

BPCC Approach for Arabic Letters Recognition

M. Albakoor
Faculty of Science
University of Aleppo
Aleppo, Syria

M. Dabsh
Faculty of Science
University of Aleppo
Aleppo, Syria

A. A. Albakkar
Scientific Research Center
Aleppo, Syria

F. Sukkar
Faculty of Informatics Engineering
University of Aleppo
Aleppo, Syria

Abstract - *This paper introduces a method of Binary Pairs and Chain Coding (BPCC) using important points for Arabic letters recognition. It is a good method paid to decrease the size of data at least 38%, reduce the number of calculations and save the time. The important points are detected. These points are coded through two counterclockwise passes; diagonal then perpendicular depending on a novel algorithm. Eight sets of binary pairs are supposed and they are used to filter the resulted code of character. Filtered code is dealt with a certain method to form an array of 85 elements. Then the array is scaled to consider as an input of NN. Two types of Arabic letter fonts of multi size are used. The results prove the recognition rate is 99.3% for Arabic Letters recognition of two fonts of multi size. The method is easy implementations, reliable to gain fast speed, high competence, hopeful results and encouraged.*

Keywords: Arabic characters, Binary pairs, coding chains, Backpropagation neural networks

1. Introduction

In spite of the high number of researches achieved recently [3, 7, 13], Arabic character recognition continues to be a difficult task. This is generally due to its cursives and the variability of Arabic characters shapes according to their position in the word. A statistical approach for recognizing cursive typewritten text [2]. Many approaches are done on recognition of Arabic cursive script as appear in [4, 6]. Efficient works are produced on recognition of cursive-character script using feature extractor [10, 11]. Works are proposed on Automatic recognition of Arabic handwritten characters using their geometrical features in [4].

This paper presents a new method of Binary Pairs and Chain Coding (BPCC) depending on important points for character coding of the Arabic letter. The image is changed into a binary image [8, 9]. The binary image is filtered depending on median filtering supposed 3-by-3

neighborhood [12, 14]. After image is thinned [10], the important points of character are defined. The important points are coded through two counterclockwise passes; diagonal then perpendicular depending on a novel algorithm. The eight basic binary pairs are assumed and used to filter the code of character. These assumed basis pairs lead to decrease the data size, to reduce the calculations and save the time. Filtered code is dealt with a certain method to form an array of 85 elements. Then the code is scaled to form an input of NN.

NN is trained using input layer of 85 neurons, one hidden layer of 500 neurons and output layer of 16 neurons. Training parameters of NN are set to the default values, momentum coefficient is 0.25, learning rate is 0.05, and the sum square error is 0.00000585. The NN uses bipolar function for training. Once the NN weights and biases are generated randomly, the range of input vector values is ranked between -1 to +1 [1, 5].

This paper is organized as the following: In the next section the feature extraction stage is clarified, section 3 describes the recognition stage based on Neural Networks and the results and discussion are summarized in section "4".

2. Feature Extraction

Input document is written in Arial and Simplified Arabic fonts of size (10,12,14). The document is scanned by scanner, then it is changed into binary image. The binary image is filtered depending on median filtering supposed 3-by-3 neighborhood.

The resulted binary image is thinned using algorithm [10]. The important points branch, start-point and end-point of character, start-point and end-point of straight piece of character are detected. The coding of character is resembled by coding these points using eight chain code technique through two passes, see Fig. 1.

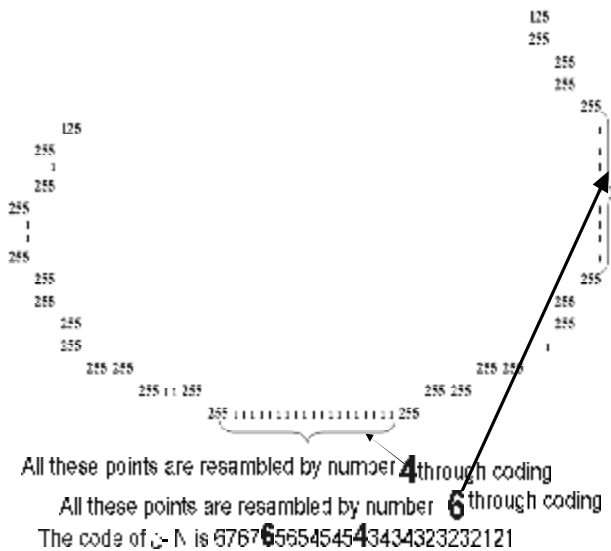


Fig.1 The coding is done on important points

The following algorithm is used for extraction of character code:

2.1 A new Character Coding Algorithm

- a. Coordinates of starting-pixel are defined considering the search as following:
 1. The image of character is divided into four parts.
 2. The search starts vertically from right to left, starting on from part 1 as clarified in table 1.

Table 1 Order search in parts of image

2	1
4	3

- b. Coding starts from starting-pixel of extracted character according to eight code technique counterclockwise, as shown in Fig. 2. considering the following states:

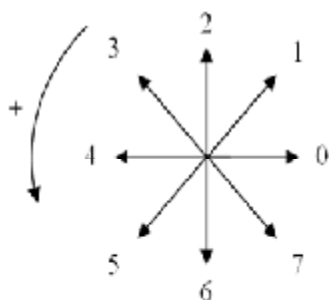


Fig. 2 Eight coding and its direction

1. Shift is made to consequent pixel depending on the following conditions:

- i. The consequent pixel has not been passed previously.
- ii. There are not any neighborhood pixels between consequent and related previous pixel of present pixel.
- iii. Coding is done according to doubled loop counterclockwise: diagonal then perpendicular, each going in ordered sequence of directions. They are 1, 3, 5, 7 for first loop, and 2, 4, 6, 0 for second loop, as shown in Fig. 5. This order of directions is organized as in table 2.

Table 2 Directions order of movement

Serial number	1	2	3	4	5	6	7	8
Directions order	1	3	5	7	2	4	6	0

2. If there is more than one direction of branch-pixel, the first direction through coding depends on the last coding number, as shown in table 3. If there is no pixel in the determined direction, the coding should be continued as in Fig. 1 considering movement start point as from the determined direction.

Table 3 Directions order through pixel coding if the pixel is a branch point.

Last coding number	1	2	3	4	5	6	7	0
First direction through coding	2	3	2	2	2	3	4	1

- c. Coding is stopped at end-pixel of extracted character .
- d. All directions of branch-pixels are considered in coding.
- e. Eight binary pairs are distinguished in character coding. The eight pairs are: $\{(1,0), (1,2), (3,2), (3,4), (5,4), (5,6), (7,6), (7,0)\}$
- f. Last coding chain is refined to numbers out of the supposed binary pairs; i.e. the numbers which can't form binary pairs of that supposed ones, see table (4-a, 4-b)

Table 4-a Overabundance numbers in coding of Arabic letter 'jeem'.

5	4	5	4	5	4	3	4	5	4	5	9	6	5	6	5	4	5	4	
5	4	5	6	5	6	5	6	5	6	7	6	7	6	7	6	7	6	7	0
7	0	7	0	7	0	7	0	7	0	1	0	1	9						

The resulted code is:

Table 4-b Coding the Arabic letter ج- 'jeem'.

5	4	5	4	5	4	3	4	5	4	9	5	6	5	4	5	4	
5	4	5	6	5	6	5	6	5	6	7	6	7	6	7	6	7	0
7	0	7	0	7	0	7	0	7	0	1	0	9					

g. Final coding matrix is calculated according to the following steps:

1. Iteration vector 'X' is supposed as zeros vector. This iteration vector consists of 9 elements as shown below:

$$X = [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0].$$

2. Iterations of binary pairs $p_{x1,x2}$ are calculated; $p_{x1,x2}$ is a binary pair of coded chain.
3. Every binary pair is resembled by a specific number, as clarified in table 6; the number '9' in table 5 refers to a branch-point.

Table 5 The eight pairs and its supposed offset.

<i>T</i> =	(1,0)	(1,2)	(3,2)	(3,4)	(5,4)	(5,6)	(7,6)	(7,0)	9
<i>Its Lookup Number</i>	1	2	3	4	5	6	7	8	9

Supposed basic matrix 'T' represents table 5 which consists of 2*9 elements. The first row of matrix refers to the binary pair and second row refers to the symbol of the binary pair, columns refer to the number of binary pairs.

4. Proposed iteration matrix 'Y' is a zeros matrix which consists of 9*9 elements. The rows of matrix represent binary pairs; and columns represent repeaters of these pairs, as shown below in table 6:

Table 6. Supposed iteration matrix

	1	2	3	4	5	6	7	8	9
Symbol of binary pairs	1	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0
	4	0	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0	0
	7	0	0	0	0	0	0	0	0
	8	0	0	0	0	0	0	0	0
	9	0	0	0	0	0	0	0	0
	Repeaters								

5. Defining 'Y' is determined by the coding chain of the extracted character regarding this procedure:

- If $p_{x1,x2} = T(1,i)$ then:

$$a. \quad j = T(2,i) \quad (1)$$

Where: $i = 1..9$

$p_{x1,x2}$ represents binary pair which is found in coding of character.

$$b. \quad X(j) = X(j) + 1 \quad (2)$$

$$Y(j, i) = Y(j, i) + X(i) \quad (3)$$

Where: $i = 1..9$

j is an index variable.
 T, X, Y are matrices proposed above

- h. Input of neural net is resembled by final iteration matrix of 81 items.
- i. Extracted character is divided into four parts, distribution percentage of the points in each part is calculated and they are added to iteration matrix. Four new items are created, at the end, matrix of 85 items is obtained and entered to the NN.

3. Neural Network

Obtained matrix of image document from feature vector as described above, is entered as an input to NN. The NN consists of one input layer with 85 neurons, one hidden layer of 500 neurons and output layer of 16 neurons due to the number of Arabic characters are reduced from 28 to 16 characters, see table 7.

Table 7. Supposed output of NN

Letter, Name	A	B	J	D	R
Pronounced	Alif	Baa	Jeem	Dal	Raa
Equivalent Arabic letter	ا	ب	ج	د	ر
Letter, Name	S	Ss	Tth	Ea	F
Pronounced	Seen	Ssad	Ttaa	Ain	Faa
Equivalent Arabic letter	س	ص	ط	ع	ف
Letter, Name	L	M	H	W	Y
Pronounced	Lam	Meem	Haa	Waw	Yaa
Equivalent Arabic letter	ل	م	ه	و	ي
Letter, Name	La				
Pronounced	Lamalif				
Equivalent Arabic letter	لآ				

These NN is trained to classify Arabic character. Training parameters of NN are set up to default values, momentum coefficient is 0.25, learning rate is 0.05 and sum square error is examined on 0.00000585. This NN uses bipolar function as an activation function. Once the NN weights and biases are generated randomly, the range of input vector values is ranked between -1 to +1, see Fig.3.

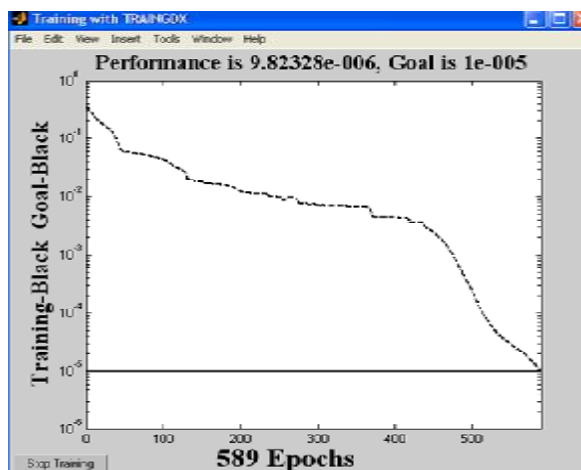


Fig. 3 The training of NN

4. Results and Discussion

In order to examine the developed algorithm, many experiments have been running, the results are illustrated below:

4.1 Considering similar letters

The output of the NN is set up to 16, taking into account Arabic letter ن - N is similar to Arabic letter ل - L, Arabic letter ك - K is similar to Arabic letter ب - B, and Arabic letter ف - F is similar to Arabic letter ق - Q.

The NN is trained on 84 Arabic characters of Arial font of multi size (10, 12, 14), and is recalled for 84 letters. The rate of recognition is 100 % at training time of 45 [seconds] with 589[cycles]. A 84 letters out of 84 letters is recognized. Also, the NN is tested on new data of Simplified Arabic font of multi size (12, 14, 16), the results are described in table 8:

Table 8: Recognition rate of different types

Size	Arial font	Simplified Arabic font
12	100 %	96 %
14	100 %	100 %
16	100 %	100 %

4.2 Joining similar letters

NN training is made taking into account joins in similar letters as follows:

4.2.1 First method “ out16-kaf_like_Ba & Noon_like_Lam & Qaf_like_Fa ”

The NN consists of input layer with 85 neurons, one hidden layer with 500 neurons and output layer with 16 neurons. It is supposed that the Arabic letter ن (‘Noon’) is similar to Arabic letter ل (‘Lam’), Arabic letter ك -K is similar to Arabic letter ب - B, and Arabic letter ف - F is similar to Arabic letter ق - Qq.

The NN is trained on 84 Arabic characters of Arial font of multi size (10, 12, 14), and is recalled for 84 letters. The rate of recognition is 100 %. Also, the NN is tested on new data of Simplified Arabic font of multi size (12, 14, 16). For Simplified Arabic font of size 12, the recognition rate is 96%. At size 14, the recognition rate is 96 %. At size 16, the recognition rate is 96%.

4.2.2 Second method “ out16-Kaf_like_Lam & Noon_like_Ba & Qaf_like_Fa”

The NN consists of input layer with 85 neurons, one hidden layer of 500 neurons and output layer of 16 neurons. It is proposed that the Arabic letter ن - N is similar to Arabic letter ب - B and Arabic letter ك - K similar to Arabic letter ل - L, also Arabic letter ف - F is similar to Arabic letter ق - Qq.

The NN is trained on 84 Arabic characters of Arial font of multi size (10, 12, 14), and is recalled for 84 letters. The rate of recognition is 100 %. Also, the NN is tested on new data of Simplified Arabic font of multi size (12, 14, 16). For Simplified Arabic font of size 12, the recognition rate is 93%. At size 14, the recognition rate is 96 %. At size 16, the recognition rate is 100%.

4.2.3 Third method “out19-Noon_like_Lam & Qaf_like_Fa”

The NN consists of input layer with 85 neurons, one hidden layer of 500 neurons and output layer of 19 neurons. It is suggested that Arabic letter ن - N similar to Arabic letter ل - L, and Arabic letter ف - F similar to Arabic letter ق - Qq.

The NN is trained on 84 Arabic characters of Arial font of multi size (10, 12, 14), and is recalled for 84 letters. The rate of recognition is 100 %. Also, the NN is tested on new data of Simplified Arabic font of multi size (12, 14, 16). For Simplified Arabic font of size 12, the recognition rate is 96%. At size 14, the recognition rate is 96 %. At size 16, the recognition rate is 100%.

5. Conclusions

In this paper a new method of Binary Pairs and Chain Coding (BPCC) is done on important points for character coding of the Arabic letter. This approach proves high results in recognition. It depends essentially on coding the important points using eight chain code through two counterclockwise passes; diagonal then perpendicular. Eight basis binary pairs are supposed and they are used to filter the chain code of character. These basis pairs abbreviate the data size paid to reduce the calculations and save the time. Filtered code is dealt with a certain method to form an array of 85 elements. Then the code is scaled to form an input to NN for recognition. The NN is trained on Arial font of multi size, the rate of recognition is 100% for trained characters. The NN is tested on a new data of Simplified Arabic font of multi size, the rate of recognition is 99.3% for data. Thus, the design and implementation of the proposed feature extractor takes important position in this research. Feature extraction is an inexpensive, treatment. The results prove the novel algorithm is a good factor to reduce the time and decrease the calculations, increase the accuracy rate and speed.

6. REFERENCES

- [1] Carl G. Looney, "Pattern Recognition Using Neural Networks", New York Oxford University Press, 1997.
- [2] El-Dabi S., Ramsis R., Kamel A., "Arabic character recognition system: A statistical approach for recognizing cursive typewritten text", *Pattern Recognition*, 23(5):485–495, 1990.
- [3] El-Sheikh T., Guindi R., "Computer recognition of Arabic cursive script", *Pattern Recognition*, 21(4):293–302, 1988.
- [4] Fahmy M., Al Ali S. , "Automatic Recognition Of Handwritten Arabic Characters Using Their Geometrical Features", Dissertation for the PhD, Egypt, 2000.
- [5] Fausett L., "Fundamental of Neural Networks", Florida, Institute of Technology, 1994.
- [6] Khorsheed M., Clocksin W. F., "Structural Features Of Cursive Arabic Script", *Proc.10th British Machine Vision Conference*, Nottingham, 1999.
- [7] Pavalidis T., A vectorizer and feature extractor for document recognition", *Computer Vision, Graphics, and Image Processing*, 35:111–127, 1986.
- [8] Pitas I., Ioannis, "Digital Image Processing Algorithms", Prentice Hall international, UK, ltd ,1993.
- [9] Rafael C. Gonzalez, Richard E. Woods, "Digital Image Processing," 2nd Edition, Prentice Hall, NJ, 2002.
- [10] Saeed K., "Image Analysis for Object Recognition," Bialystok Technical University Press, Bialystok 2004.
- [11] Saeed K., Tabêdzki M.," Intelligent Feature Extract System for Cursive- Script Recognition," 4th IEEE International Workshop on Soft Computing as Tran disciplinary Science and Technology (WSTST'05), 25-27 May, Muroran 2005.
- [12] Scott E Umbaugh, "Computer Imaging: Digital Image Analysis and Processing", CRC Press, FL 2005.
- [13] Touj S., Ben Amara S., Amiri H., "Global Feature Extraction of Off-Line Arabic ", *SMC'2002*, Hammamet, Tunisie, October 2002.
- [14] Weeks Jr, Arthur R., "Fundamentals of Electronic Image Processing", SPIE/ IEEE series on Imaging Science & Engineering, 1996.