Handwriting Arabic Words Recognition Based on Structural Features

Salim Aloud
salemali416@yahoo.com
Department of Computer Science, College of Sciences
Azzaytuna University, Libya
Received: 26 April 2018 / Accepted: 11 May 2018

Abstract

Handwriting recognition technology is the ability of a computer to recognize characters, words and other symbols that have been written by hand in natural handwriting. This study presents a method for recognition of Handwritten Arabic Words (HAW) through expanding in the way of structural features extraction by relying on geometrical information (straight lines, loops, points, and curve). The input to the system is binary images written by hand by number of people. The features are to convert the image from two dimensional into one dimensional as a victor that is to be used as a signature for the image the experiments have been conducted on a database of a thousand words representing names of a hundred Libyan cities at a rate of ten patterns for each city. Classification of the words was dependence on Artificial Neural Networks (ANN) of Multiple Layers Perceptron (MLP) type. Wherein half of the words were used to train the network and the other half to test the network. The ratio of recognition was 80.4 %.

Keywords: recognition, features extraction, structural information.

1 Introduction

Words and characters recognition methods have been improved since many years. These methods used for printed or handwritten scripts and used two different approaches of processing which are online and offline. It has been gaining more interest lately due to the increasing popularity of handheld computers, digital notebooks, and advanced cellular phones. These devices nowadays are commonly used worldwide that encouraged companies to improve their products to support multi languages. These devices can deal with many languages spoken by billions of people around the world. Arabic language is the main language of all Arabic countries with more than 280 million people are speaking this language as a first language and by 250 million as a second language. Arabic language comes as the fifth rank of most commonly used languages in the world. There are some other languages related to Arabic language. These languages have some similarities with Arabic language whence from the characters shapes or from the pronunciation [1] .the progress in Arabic language is slower than the progress in developing solutions for Latin and Asian languages [7].There are many other applications for analysis of human handwriting such as writer recognition and verification, form processing, interpreting handwritten postal addresses on envelopes and reading currency amounts on bank checks etc. The main

problem encountered when dealing with handwritten Arabic characters is that characters written by different persons representing the same character are not identical but can vary in both size and shape. unlimited variation in human handwriting styles similarities of distinct character shapes, character overlaps, and interconnections of neighboring characters. In addition, the mood of the writer and the writing situation can have an effect on writing styles [2][4][6] Handwritten recognition starts with image preparation stage by transforming it from a color image into grayscale image. Then convert it into a binary image. . The preparation stage is followed by features extraction stage during which the image is converted into a group of features in order to change them from two dimension data to one dimension data or a victor of the features. In general the features come in three main parts statistical features, structural features and global transformation. [1]. In this study the structural features are used which depend on the word's geometrical information as ratio of the length to the width, the loops, branching points, straight lines and the curve or slopes in the various directions. The process of separation of the features of each word connected with it, occurs by relying on the ANN of the MLP type. The training process takes place by allowing the network to practice on half of the number of patterns at a rate of 5 models, or patterns for each word. The training algorithm place by way of back propagation (BP) which is used to train the MLP type network [8]. The testing process was conducted on the other half of the patterns which amount to 500 words. The recognition rate was over 80%.This paper is organized as follows: Section 1 gives a brief description of Arabic Alphabet Characteristics. Section 2 explained Features extraction stage. Section3 explaining classification stage by ANN. Section 4 gives Experiments and results.

2 Arabic Language Alphabet Characteristics

The Arabic language has a lot of advantages which make it different from the other ones in terms of shape, and way of writing and direction of the writing and which are clarified as follows:

1. Arabic text (machine printed or handwritten) is written cursively and in general from right to left.
2. Arabic writing uses letters, punctuation marks, spaces, and special symbols.
3. An Arabic letter might have up to four different shapes, depending on its relative position in the word: 1: isolated, 2: connected from the left, 3: connected from the right and 4: connected from both right and left according to its place in the word like the letter (ع). Table 1.
4. Some letters exist as a combination of two letters in some certain situations, like the letter (lamelif لا) which is created by combining two letters , the letter(lam ل) and (alif ا).
5. Sixteen Arabic letters have from one to three secondary components. The type and position of the secondary components are very important features of Arabic letters. For example, Tah (ط) and Thah (ظ) differ only by the number of dots above the main body, Seen (س) and Sheen (ش), Sad (ص) and Dad (ض).

Table 1: *show four different shapes for letter (ع)*

(a) isolated	(b)connected from the left	(c)connected from the right	(d)connected from both right and left
ع	عـ	ـع	ـعـ

6. Arabic writing contains many fonts and writing styles. The letters are overlaid in some of these fonts and styles.
7. Ligatures are combinations of two and sometimes three letters into a single shape [4].

In general, the Arabic writing is written by using different writing techniques, or styles which result in letters and words having different shapes which in turn cause obscurity in any recognition system.

In general the Arabic writing may be classified into three different styles:

Typewritten: This style is generated by computer. It is the simplest one because the characters are written without overlaps or ligature.

Typeset: This style is more difficult than the typewritten because it has many ligatures and overlaps. It is used to write newspapers and books. Nowadays, this style may also be generated using computers.

Handwritten: This style is the most difficult because of the variation of writing the Arabic alphabets from one writer to another [3].

3 Features Extraction

The first step in features extraction stage is preprocessing stage. In this step the image convert from gray scale into binary image which means it has only two levels zero (0) level which represents a background, and level (1) which represents foreground. The change process occurs by using the threshold technique. And then extract connected components through convert it into labeled image [9]. Then the features extraction stage comes. The features extraction stage is considered as the most important stage of the study and the capability of any recognition system to differentiate any writing depends to a large degree on the exactness of the features extracted from the image. In this study the structural features are used.

Structural features describe the geometrical and topological characteristics of a pattern by describing its global and local properties. The structural features depend on the kind of pattern to be classified. For Arabic characters, the features consist of (ratio of the length to the width, the loops, branching points, straight lines and the curve or slops in the various directions).In this study; the structural features have been used where a word is divided as in the Figure 1. Features of each area are extracted by analyzing the connected components existing in each area.

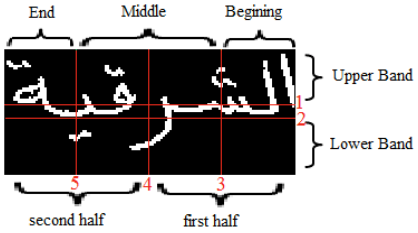

Figure 1: *show sample of the name (الشرقية)*

The letter alif (ا)usually appears in the upper part of a word and its height is twice as large as its width or more Figure 2. And the loops: each part of the binary image that has the color of the background, and whose edges have the color of the foreground, and falls within the connected components is regarded as a loop Figure 3.

Figure 2: *detect letter alif*

Figure 3: *detect loop*

The calculation of one point is found out through calculation of the area, of each component, and case where the area is less than the threshold, it is considered as a point Figure 4. And in case where the width of the component is bigger or equal to twice its length (-) then the component is considered as two points Figure 5. And in case there is a curve in the point falling above a word, the component is then regarded as three points (ٮ) Figure 6.

Figure 4: *detect one point*

Figure 5: *detect two points*

Figure 6: *detect three points*

The features that are possible to obtain in the lower part of the image are:
The letters of Arabic language that may appear underneath a word are
(ج ح خ ر ز س ش ص ض ع غ ل ن و ي) detect this features by tracing the number of crossings from the background to the foreground horizontally (h) and vertically (v) at middle connected component if h =1 and v = 1 return (curve ر) Figure 7. If h = 2 and v = 1 return (curve ن) Figure 8. if h = 1 and v = 2 return(curve ح) Figure 9.

Figure 7*: detect curve* ر

Figure8: *detect curve* ن

Figure 9: *detect curve* ح

Table 2: *shows details of features vector*

element	Description
1	Number of straight lines(ا) in beginning of word
2	Number of straight lines(ا) in middle of word
3	Number of straight lines(ا) in end of word
4	Number of loops in beginning of word
5	Number of loops in middle of word
6	Number of loops in end of word
7	Number of points(•) up the word
8	Number of points(•) down the word
9	Number of (curve ر)in first half of word
10	Number of (curve ر) in second half of word
11	Number of (curve ن)in first half of word
12	Number of (curve ن) in second half of word
13	Number of (curve ح)in first half of word
14	Number of (curve ح) in second half of word
15	Number of (letters kaf ک) in the word

The letter kaf calculated either by calculating the width of the component Figure 10.
Or by tracing the number of crossings from the background to the foreground vertically Figure 11.

Figure 10: *detect character kaf*

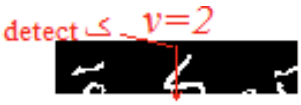

Figure 11: *detect character kaf*

In this study a features vector containing 15 elements has been set up each of which expresses a feature of the word in a way that each image is represented by a vector containing 15 elements. This vector is the one that is used in the process of training and testing of the ANN. This vector is shown in Table 2.
For example the word (توكرة) Figure 12 its features vector was as follows:

Features vector = [0 0 0 1 0 1 4 0 1 1 0 0 0 0 1]

Figure 12: *word (توكرة)*

The first, second and third elements are 0, 0, 0 and they mean that they don't contain the straight lines neither on the beginning of the word nor the middle or end of the word. The fourth component (1) means existence of a loop at the beginning of the word, the fifth nonexistence of a loop in the middle of the word, the sixth which is (1) existence of one loop at the end of the word, the seventh (4) meaning existence of four points over the word, the eight (0) nonexistence of points under the word and the ninth which is (1) meaning existence of a (curve ر) in the first half of the word and the tenth (1) which means existence of an (curveر) in the second half of the word. The eleventh and twelfth which are (0), mean nonexistence of (curve ن) in the first nor second half of the word. The thirteenth and fourteenth which are (0) mean the) curve ح) without its curve in the first half or the second one. The fifteenth (1) means existence of the letter (Kaf ك) in the word.

4 Classification Stage

The process of separation of the features of each word connected with it occurs by relying on the ANN of the MLP type which is used to separate any data even if they are not linear. The MPL network contains three layers input, hidden and output layer and in each layer there is a number of nodes Figure 13 [4]. The number of nodes in the input layer is equal to the number of elements in the features vector (15 elements). But the number of the nodes in the output layer depends on the number of the words which are to be separated (100 words). The hidden layer lies between the input and output layers. The training algorithm place by way of back propagation (BP) which is used to train the feed forward network MLP type network with supervised learning [8].

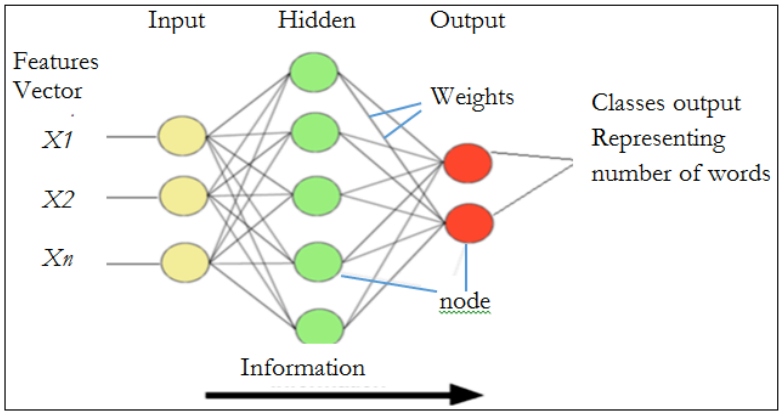

Figure 13: *ANN of the MLP type*

5 Experiments

5.1 Training Mode

In the training mode the first half of the data was selected and such a half represents 500 words at a rate of 5 forms for each word .The number of the names is 100 which requires to be 100 classes. And as the features vector includes 15 components, the number of the input layer nodes becomes 15. The number of the hidden layer's nodes

was 22. As the required number of the classes is 100, the number of the output layer's nodes was 7. And the learning rate was 0.2 the network has been trained by counting on iteration in a way that the number of the iteration was determined to be at 30000, a number that reached to $8.222 * 10^{-5}$ by the average of the errors.

5.2 Recall Mode

In this mode the values of the weights are fixed. The neural network works to determine only the input data in each class. The values of the weights are not changed nor there a calculation for the error. In the recall mode, the other half of the patterns, totaling 500 words and representing an extra 5 models for each word, have been used.

6 Results

Out of the 500 words, the result was that the ANN classified, 402 correct words while 98 words were identified incorrectly, which means the rate of recognition was 80.40 % as in the Table 3.

Table 3: *shows the results of study*

Number of words	Correct words	Incorrect words	rate of recognition
500	402	98	80.4%

7 Conclusion

In this research, the way the features are extracted has been expanded upon to involve more details on the geometrical features connected with the word written by hand, which is to discover the curve, underneath the word which appears to lean to the left or to the bottom, thereby leading to more accurate features which in turn contributed in the enhancing rate. It remains a difficult task to try to recognition of Arabic handwriting due to difference in writing styles from one person to another and to various kinds of handwriting and what accompany them in terms of overlapping and interconnection. As such, the main concern is to upgrade the capability to recognition to the maximum extent possible. And the enhancing relies in the main on the way the features are extracted, that is to say, the more exact, or accurate the features are the more better the capability to recognition becomes.

In addition, the recognition enhancing rate depends on the separation method. Through the findings of this research it was found that some errors were attributed to the network not being trained on some patterns, thus not being capable to recognition those patterns. The solution lies in increasing the patterns, or in other words, in training the network on as much patterns as possible. In the future work, the same methods connected with the findings of this research will be employed in recognizing texts written by hand.

References

[1]M.A Abuzaraida and A.M Zeki and A. M Zeki " Feature Extraction Techniques of Online Handwriting Arabic Text Recognition " 5 th International Conference on Information and Communication Technology for the Muslim Word 2013.

[2] H.EL Moubtahij, A.Halli and K.Satori " Review of Feature Extraction Techniques for Offline Handwriting Arabic Text Recognition " International Journal of Advances in Engineering & Technology, Mar.2014. Vol.7,Issue 1,pp.50-58.

[3] A. Lawgali, A. Bouridane, M. Angelova, Z. Ghassemlooy "Handwritten Arabic Character Recognition: Which Feature Extraction Method? " International Journal of Advanced Science and Technology Vol.34,September, 2011.

[4] Rafael M. O. Cruz, George D. C. Cavalcanti and Tsang Ing Ren "Handwritten Digit Recognition Using Multiple Feature Extraction Techniques and Classifier Ensemble "IWSSIP 2010 - 17th International Conference on Systems, Signals and Image Processing.

[5] Ashoka H.N. , Manjaiah D.H. , Rabin dranath Bera " Feature Extraction Technique for Neural Network Based Pattern Recognition" International Journal on Computer Science and Engineering , Vol. 4 No. 03 March 2012.

[6] Fenwa Olusayo Deborah, Omidiora Elijah Olusayo , Fakolujo Olaosebikan Alade," Development of a Feature Extraction Technique for Online Character Recognition System" Innovative Systems Design and Engineering , ISSN 2222-1727 (Paper) ISSN 2222-2871 (Online)Vol 3, No 3, 2012.

[7] G. Abandah, K. Younis, M. Khedher " HANDWRITTEN ARABIC CHARACTER RECOGNITION USING MULTIPLE CLASSIFIERS BASED ON LETTER FORM" In Proc. 5th IASTED Int'l Conf. on Signal Processing, Pattern Recognition, & Applications (SPPRA 2008), Feb 13-15, Innsbruck, Austria.

[8] A. Lawgali, A. Bouridane, M. Angelova, Z. Ghassemlooy "Handwritten Arabic Character Recognition: Which Feature Extraction Method? " International Journal of Advanced Science and Technology Vol.34,September, 2011.

[9] Rafael C. Gonzalez and Richard E. Woods "Digital Image Processing " Second Edition. Prentice Hall 2002.