Arrasikhun International Journal, Volume 4, Issue 2, Dec 2018
ISSN: 2462-2508

Propose Recognition Technique Model for Arabic Islamic Manuscripts

SHADI M. S. HILLES
Computer Science Department, Al-Madinah International University
shadihilless@gmail.com

Abstract

This paper presents a new method for Ancient Islamic Manuscripts Recognition and Arabic scrip recognition. The main contributions of this work is that the method of recognition and classification of the data used multilayer model of neural network, used as a technique Arabic Manuscript, the performance of the proposed method is assessed using samples extracted from a historical handwritten manuscript. Arabic text is cursive, and each character may have up to four different shapes based on its location in a word. Quite often old documents are subject to background damage. Examples of background damages are varying contrast, smudges, dirty background, and ink through page, outdated paper and uneven background, The hyper plane approach is employed as classifiers due to the low computation overhead during training and recall process.
Keywords: *Recognition Technique, Arabic Manuscripts, Optimal hyperplane, Support vectors, Margin, Origin.*

1. Introduction

Handwritten character recognition is one of the most difficult tasks in the pattern recognition system. There are a lot of difficult things that need many image processing techniques to solve, for examples: 1) how to separate cursive characters into an individual character, 2) how to recognize unlimited character fonts and written styles, and 3) how to distinguish characters that have the same shape but different meaning such as the character o and number 0. Many researchers try to apply many techniques for breaking through the complex problems of handwritten character recognition. There are many applications that need to take advantage of the handwritten character recognition system, namely, 1) automatic reading machine, 2) non-keyboard computer system, and 3) automatic mailing classification system. [1] The objective of this research is to try to help researchers to recognize Arabic Islamic handwritten characters by using hyperplane technique. The Arabic alphabet has 3 vowels, 25 consonant, as shown in Figure 1.

خ	ح	ج	ث	ت	ب	ا
kha	haa	jiim	thaa	taa	baa	alif
ص	ش	س	ز	ر	ذ	د
saad	shiin	siin	zaay	raa	thaal	daal
ق	ف	غ	ع	ظ	ط	ض
qaaf	faa	ghayn	ayn	thaa	taa	daad
ي	و	ه	ن	م	ل	ك
yaa	waaw	ha	nuun	miim	laam	kaaf

Figure 1.Arabic Alphabet

Arabic is a language spoken by Arabs in over 20 countries, and roughly associated with the geographic region of the Middle East and North Africa, and it is considered as a second language for several Asian countries in which Islam is the principle religion (e.g, Indonesia). In addition, non-Semitic languages such as Farsi, Urdu, Malay, and some West African languages such as Hausa have adopted the Arabic alphabet for writing [1]. Since handwritten isolated Arabic character is the domain of the proposed system, some characteristics which differs it from the other should be known, as stated by Abuhaiba in [1]. Isolated characters have the following interested features:

a. Arabic script is cursive and is written from right to left.

b. Any Arabic character has exactly on main stroke and zero or more secondary strokes as **ش، ظ، ب، س**

c. Usually, a secondary stroke does not touch the main stroke as (**ب**). If this happens, it will be in limited number of character as (**ظ**).

d. Some Arabic characters have the same shape; however, they are distinguished from each other by the addition of secondary strokes, e.g., dots, in different positions relative to the main stroke as (**ظ، ط، ت، ب**).

Sometimes, the ambiguity of the position of these secondary strokes in handwriting brings out many different readings for one word.

e. Some Arabic characters contain loops as (**ف**), but no more than two loops may be adjacent share a common link.

f. Arabic characters vary in size, particularly in width, even within the same font of type printed text.

g. Some Arabic characters use special marks to modify the character accent, such as Hamza (**ء**) and Madda (~), which are positioned at a certain distance from the character.

Table 1.Number of Samples/Character

Character	**# of samples**
ل ,م ,ه ,ب,د,ذوص,ح,ع,غ,ص,ض	14
س	28
ش	6
ت	13
ث	9

Artificial Neural Network which will be used as approach on this research is an information processing system. The Artificial Neural Network is a group of simple and interrelated cells. The cells are arranged in such way that each cell derives its input from one or more other cells and linked through weighted connections to one or more other cells. This way, input to the ANN is distributed throughout the network so that an output is in the form of one or more activated cells. Figure 2 shows the architecture of the Artificial Neural Network. It consists of 3 layers: the input layer, one hidden layer and the output layer.

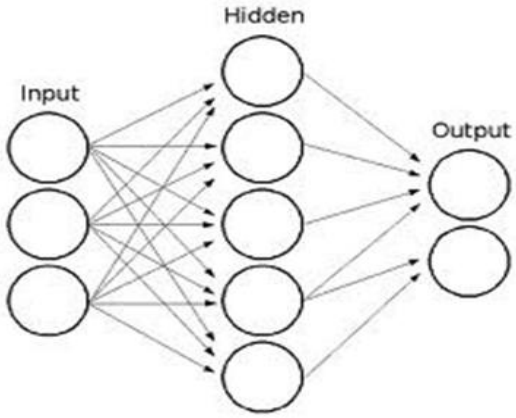

Figure 2. Neural network architecture.

2. Related Work

Character recognition has been seen as one of important Pattern recognition pillars. Arabic character Recognition has been one of the major languages to receive attention. Since high variability is expected even in printed characters, due to the large number of font styles and other reasons, Nouh in [4], suggested a standard Arabic character set, in order to facilitate computer processing of Arabic characters. Isolated characters are simulated and described by suitably chosen components (radicals). The simulated Arabic alphabet is classified utilizing a sequential tree search technique and certain correlation measurements. The disadvantage of the proposed system is the assumption that the incoming characters are generated according to specified standard rules putting strict constrains on font style design.

Al-Jawfi in [2] presents a handwriting Arabic character recognition method using LeNet NN after applying character segmentation. LeNet neural network was design to recognize a set of handwritten Arabic characters. This NN Design depend on two main stages the first to recognize character shape using pixel matrix of 16×16 an features inputs, while the second stage is to recognize the number of dots, position, and where it is a dot or zigzag using back propagation algorithms. Performance of this algorithm depends firstly of the accuracy of segmentation algorithm in addition to the noise removal. On the other hand, neural networks rely on Image Based features to recognize body shape, which may not hold all of character feature.

Al-Sheik and Al-Taweel in [5] assumed a reliable segmentation stage, which divided letters into the 4 groups of letters (initial, medial, final and isolated).

The recognition system depended on a hierarchical division by the number of strokes. One stroke letters were classified separately from two stroke letters etc., ratios between lines and position of dots in comparison to the primary stroke were defined heuristically on the data set to produce a rule-based classification. This approach had an excellent recognition rate and a good divide-and-conquer strategy by reducing the classes through hierarchical rules. However, it would be extremely sensitive to noisy data in terms of the number of strokes since the hierarchy was built on counting the exact number of strokes.

El-Wakil and Shoukry in [6] used stable features to hierarchically reduce the number of letter class considered based on template matching. The stable features were:

1. The number of dots.

2. Relative position of the dots compared with theprimary stroke.

3. Number of secondary strokes.

4. Slope of secondary stroke.

A k-nearest neighbor classifier then used primary strokes encoded as a primitive of angular directions in the stroke to determine the closest class. Recognition accuracy varied with the length of primitive strings but the optimal string length gave an accuracy of 84% by testing 7 writers on sets of 60 characters. Weighting the features manually by their relative importance gave a maximum accuracy of 93%. Like many other systems the authors showed good recognition results. Also, like many other systems, this approach's stable features were sensitive to noise and might not generalize well since the results were based on a testset of 60 characters alone.

Zafar et al. in [5] describe a simple approach involved in online handwriting recognition by avoiding lengthy pre-processing and extract only useful character information. The system evaluates the use of the Back Propagation Neural network (BPN). The recognition rates were 51% to 83% using the BPN for different sets of character samples. They tested the techniques for upper-case English alphabets for a number of different styles from different subject. We cannot generalize well since the results depend on the number of samples/characters to determine the rate of performance.

The e-government covers wide range of areas and there is proposed standard model for e-government services usage [22] which used handwritten Arabic alphabet for text recognition and morphological. Proposed method of data compression used SOFM of deep learning neural network model to compress 24 bit image with evaluated with standard method of image compression such as wavelet transform and DCT transformation and as well fractal method of image compression [23], proposed and develop method of biometric face detection and recognition by using local binary pattern and histogram of oriented gradients, the proposed model is considered automatic attendance system and evaluated used SVM [24].

3. Methodology

Consider the simplest case: linear machines trained on shared data (which we will show that the analysis for the general case - nonlinear machines trained on linearly non-separable data - a problem that is close to the quadratic programming problem). Note the training data, such as $\{ x_i, y_i\}, i = 1, \dots, l; y_i \in \{-1, +1\}, x_i \in R^n$ [8, 12]. Suppose has some of hyperplane, which separates the positive from the negative data (separate hyperplane). the x on the hyperplane satisfy the following condition: $\langle w, x\rangle + b_0 = 0$,

Where w - Normal hyperplane, than $|b_0|/\|w\|$ - perpendicular hyperplane

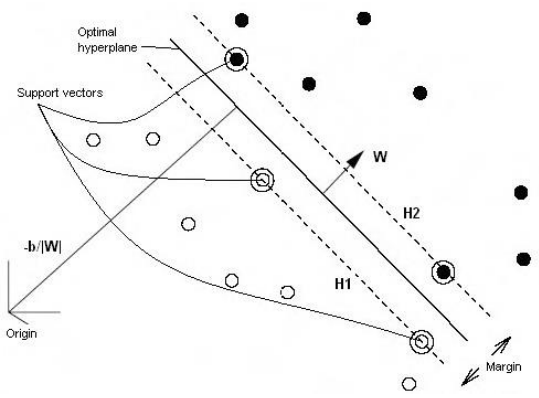

Figure 3. The optimal hyperplane for linearly separable images: Optimal hyperplane (Optimal hyperplane, Support vectors, Margin, Origin) [7]

Support vector machine - an algorithm trained to recognize two classes of objects. For multiclass problems the principle of recognition is based on the principle of "one pro-against all." Essence of support vector machines is to construct a hyperplane separating the maximum positive and negative points. Moreover, among all such hyperplane is the one for which the minimum distance (gap) from the nearest of the points possible. Support vector machine can

provide a good quality of recognition in the classification problem, without having a priori knowledge about the domain specific task, since works with an abstract vector model representation of the data. This property is unique to support vector machines [8, 9].

Support vector machine represents a category of universal neural networks of direct distribution. The technique of construction of the support vector machines for creation of the qualifier applied to identification and classification of symbols is considered.

4. Testing the classifier (Carrying out the experiment)

To verify the correct operation of the constructed classifier was selected letter alphabet series, consisting of 28 Arabic characters, from أ to ي (Alif to ya) and numerical series and Arabic numerals 0-9. Thus, the complete data set consists of 38 classes. To display symbols used matrix size of 5 × 8:

	1	2	3	4	5			1	2	3	4	5			1	2	3	4	5	
1	5	5	5	5	5	5	1	8	8	8	8		5	1						5
6	5				5	10	6					8	10	6	6				6	10
11	5					15	11		8	8	8	8	15	11	6				6	15
16	5	5	5	5		20	16	8					20	16	6				6	20
21					5	25	21	8					25	21		6	6	6		25
26					5	30	26	8					30	26						30
31	5				5	35	31	8					35	31			6			35
36		5	5	5		40	36		8	8	8	8	40	36						40

Figure 4. Examples of character representation (training data)

Each symbol is represented as an array of numbers with the symbol, i.e. Tagged, class of, the index feature of the sample and the numerical value of this pattern. For example, in such a way represented a symbol of the number "5":

35 1:1 2:1 3:1 4:1 5:1 6:1 7:0 8:0 9:0 10:1 11:1 12:0 13:0 14:0 15:0 16:1 17 : 1 18:1 19:1 20:0 21:0 22:0 23:0 24:0 25:1 26:0 27:0 28:0 29:0 30:1 31:1 32:0 33:0 34:0 35:1 36:0 37:1 38:1 39:1 40:0,

where "35" at the beginning of the line - class label, a pair of values <1:1>, <2:1>, <3:1>, ..., <40:0> - vector signs of the class and their corresponding values. If the element of the sample (at-sign) full (painted over), its numerical value is "1", otherwise "0", for example to represent numbers above "5" 1:1 2:1 ... 39:1 40:0. Numbering features left to right from top to bottom, so that the upper left corner of the matrix corresponds to the 1st sample basis, and lower right - 40.

To train a classifier (model building) was created by 9 training of samples for each class of 324 sample vector. To test the classifier (classification) was established on 12 samples for each class. The arrangement of the samples for the classification (test vector data) was chosen randomly. Total test sample 432 was created. Correctly classified (correlated to their classes) were 409 out of 432 samples, representing 94.67% accurate classification.

It should be noted that the training of the classifier was 8.3 s, and testing a classifier (classification) was conducted for 0.7 sec. This time distribution of learning is primarily due to the implementation of mathematical calculations in the code, but also with the amount of computation, which is directly proportional to the amount of data for training the classifier, the length of the unit vector of the sample data.

5. Conclusions

The described method of Arabic manuscript recognition and classification of the data at this stage of the experimental design allows to judge fairly high probability of accurate classification of data. The main advantage of this software implementation, support vector machines is the ability to use the program for the classification of data from various fields of science, the only requirement in this case is the representation of the samples the data in the format of feature vectors and their corresponding values. It should be noted the high speed of the classifier using a ready-made model.

However, the question needs to be improved to reduce the training time of the classifier, which primarily can be done by transferring the load of the mathematical-mechanical calculations using mathematical packages automated processing of information.

References

[1]. Chomtip Pornpanomchai, Dentcho N. Batanov and Nichola Dimmitt, “Recognizing Thai handwritten characters and words for humancomputerinteraction”, International Journal of Human-Computer Studies, pp. 259-279, (2001)

[2]. Al-Jawfi R., "Off Handwriting Arabic Character Recognition LeNet Using Neural Network," The International Arab Journal of Information Technology, vol. 6, no. 3, pp. 304-309, 2009.
[3]. El-Sheikh T. and El-Taweel S., "Real-time Arabic Handwritten Character Recognition," Pattern Recognition, vol. 23, no. 12, pp. 1323-1332, 1990.
[4]. Nouh A., Sultan A., and Tolba R., "On Feature Extraction and Selection for Arabic Character Recognition," Arab Gulf Journal for Scientific Research, vol. 2, no. 1, pp. 329-347, 1984.
[5]. Zafar M., Dzulkifli M., and Razid M., "Write Independent Online Handwritten Character Recognition using A Simple Approach," TheInternational Arab Journal of Information .
[6]. Burges C.J.C. A Tutorial on Support Vector Machines for Pattern Recognition., Boston, 1998.
[7]. Noble W.S., Pavlidis P. Gist: Support vector machine and kernel principal components analysis software toolkit.
[8]. Platt J., Cristianini N., Shawe-Taylor J. // Large margin DAGs for multiclass classification. In Solla S.A., Leen T.K., and Muller K.-R., editors. Advances in Neural Information Processing Systems. MIT Press, 2000. Vol. 12, P. 547–553.
[9]. Schölkopf B., Burges C.J.C., Smola A.J. Advances in Kernel Methods. Support Vector Learning, MIT Press, Cambridge, USA, 1998. Vapnik V. // NEC Journal of Advanced Technology. 2005. № 2.
[10]. Vapnik V. Statistical Learning Theory. Wiley, 1998.
[11]. Vapnik V.N. Statistical Learning Theory. Adaptive and Learning Systems for Signal Processing. Communications and Control. Wiley, New York, 1998.

[12]. Gray, R.M., 1989. Vector quantization. IEEE ASSP Mag. 1, 4–29.
[13]. Khorsheed, M.S., 2000. Automatic recognition of words in arabic manuscripts. Ph.D. thesis, University of Cambridge. Also available as University of Cambridge, Computer Laboratory Technical Report No. 495 (June 2000).
[14]. J. N. Kapur, P. K. Sahoo, and A. K. C. Wong, "A new method for graylevel picture thresholding using the entropy of the histogram," Graph. Models Image Process, vol. 29, pp. 273-285, 1985.
[15]. Y. Jui-Cheng, C. Fu-Juay, and C. Shyang, "A new criterion for automatic multilevel thresholding, " Image Processing, IEEE Transactions on, vol. 4, pp. 370-378, 1995.
[16]. W. H. Tsai, "Moment-preserving thresholding: A new approch, " Graph.
[17]. Models Image Process, vol. 19, pp. 377-379, 1985.
[18]. L. L. Sulem, and M. Sigelle, "Recognition of degraded characters using dynamic Bayesian networks", Pattern Recognition, (2008), Vol. 41 Issue 10, pp. 3092-3103.
[19]. R. Seethalakshmi, T.R. Sreeranjani, T. Balachandra, a. Singh, M. Singh, R. Ratan, and S.Kumar, "Optical character recognition for printed Tamil text", Journal of Zhejiang University SCIENCE, Vol. 6A Issue 11, pp. 1297-1305, 2005.
[20]. M. Sezgin, and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation", Journal of Electronic Imaging, Vol. 13 Issue 1, pp. 146-165,
[21]. 2004. T. M. Breuel, "Binary morphology and related operations on run-length representations", in *International Conference on Computer Vision Theory and Applications* (*VISAPP*), 2008.
[22]. Ibrahim, R., Hilles, S. M., Adam, S. M., & El-Ebiary, Y. (2016). Methodological Process for Evaluation of E-government Services base on the Federal Republic of Nigerian Citizen's E-government Services usage. Indian Journal of Science and Technology, 9(28).
[23]. Hilles, S., & Maidanuk, V. P. (2014). Self-organization feature map based on VQ components to solve image coding problem. ARPN Journal of Engineering and Applied Sciences. Vol. 9,№ 9: 1469-1475.
[24]. Mady, H. H., & Hilles, S. M. (2017). Efficient Real Time Attendance System Based on Face Detection Case Study "MEDIU Staff". International Journal on Contemporary Computer Research (IJCCR), 1(2), 21-25.