IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11705

# Recognition of Handwritten Arabic Names using Probabilistic Neural Networks

**Mohamed Nour I. Ismail[1], Mohamed Elhafiz Mustafa[2]**

[1]Department of Basic Sciences, Preparatory Year Deanship, King Faisal University, Al-Ahsa, Saudi Arabia

[2]Department of Computer Science, College of Computer and Information Sciences,

Jouf University, Skaka, Saudi Arabia

**Abstract**: Emancipation of the computers from the limited in space data entry tools (such as keyboards) and its possession of the ability of hearing and reading, remains an area of active research in computer science for more than four decades. During this period, the researchers have provided a considerable number of methods and algorithms for the computerization of hearing and reading in what is known as Pattern Recognition in computer science. One of these methods is the Holistic Approach, which has proved its efficiency in the fast recognition. The usage of neural networks with the holistic method has special importance, as it helps to determine the transition to the analytical method very easy. This paper presents successful recognition experiments of probabilistic neural networks to recognize holistically the most common Arabic names. This network was succeeded in recognizing a high proportion of Arabic names very quickly because it does not segment the words.

**Keywords**: Holistic approach, Probabilistic Neural networks, Arabic names recognition

## I. INTRODUCTION

Natural language processing systems attempt to create direct means of dealing between man and the machine through spoken or written human natural language (Arabic, English, Chinese ... etc). From here, the need arose for speech recognition and Optical Character Recognition (OCR) [1, 2, 3, 5, 6, 7]. Many studies were conducted in the area of language recognition such as Chinese, Latin, English and Japanese.

These studies yielded good results and produced technologies accessible to the user. Despite the fact that Arabic is the primary language for millions of people around the world [3] the amount of research on the computerization of Arabic, particularly in the area of recognition is very small when compared to other languages.

There are two methods for the recognition of writing. The first one is the Analytical Approach, which depends on segmenting the word into characters and then recognizing the parts (characters, or parts of characters). The other method is the Holistic Approach [4]. The most important characteristic of the Holistic Approach is that it deals with the word as a single unit for recognition without the need for segmentation. This makes the recognition process very fast. The main shortcoming of this method is that it deals with a limited vocabulary. The large increase in the speed of computational processing and in storage capacity and progress in parallel processing have drawn attention to this method.

Arabic names used today have so much repetition, such as names of the prophets (Mohammad " محمد " , Ibrahim "ابراهيم" , Jesus " عيسى" , Moses " موسى".. etc.) . Figure 2 contains sample images for the name Mohamed. Names of the Caliphs ( Abubaker " ابوبكر" , Omer "عمر" , Osman "عثمان" , Ali "علي") and compound names whose first element is Abd (slave of God) (Abdullah" عبد الله" , Abdul Rahman " عبد الرحمن " ... etc.), and there are many examples of repetitive names (such as Adil " عادل" , Awad " عوض" Saud "سعود" Fahd " فهد"... etc.), together with a few common names. Therefore, the idea of designing a system that uses the Holistic Approach to quickly recognize the common names and resort to the use of the Analytical Approach to recognize the names that are not common, which could be done by new methods like deep learning [10, 11, 12]. This paper examines the effectiveness of the first part of this system, which is the use of probabilistic neural networks to recognize the most common Arabic names in one shot.

The Holistic Approach is one of the Statistical Pattern Recognition methods [5] and it depends on the statistical properties of patterns, and at the same time, it falls under the Neural Networks because its structure is quite similar to neural networks. Neural networks are of the most popular methods used in the field of pattern recognition.

## II. SUST-ARG NAMES DATASET

The Data set used in this paper SUST-ARG is a data set designed and collected by the Pattern Recognition Research Group, College of Computer Science and Information Technology, Sudan University of Science and Technology [9]. The data collected from a local application certificate request form. The form contains the necessary data needed to obtain

 This work is licensed under a Creative Commons Attribution 4.0 International License 

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11705

the graduation certificate, and the data include the quadruple name of the applicant in Arabic and English. Figure 1, shows the upper part of the form.

The formation of data set passed through several stages, starting from the process of scanning the forms. The forms were kept in the form of image files of type (Bit map image). The next stage was for cropping names from the forms and storing each name in a separate file and Figure (2) shows samples for the name "Mohamed" after the cropping. After that the names were sorted out and all images related to a specific name were stored in a separate folder

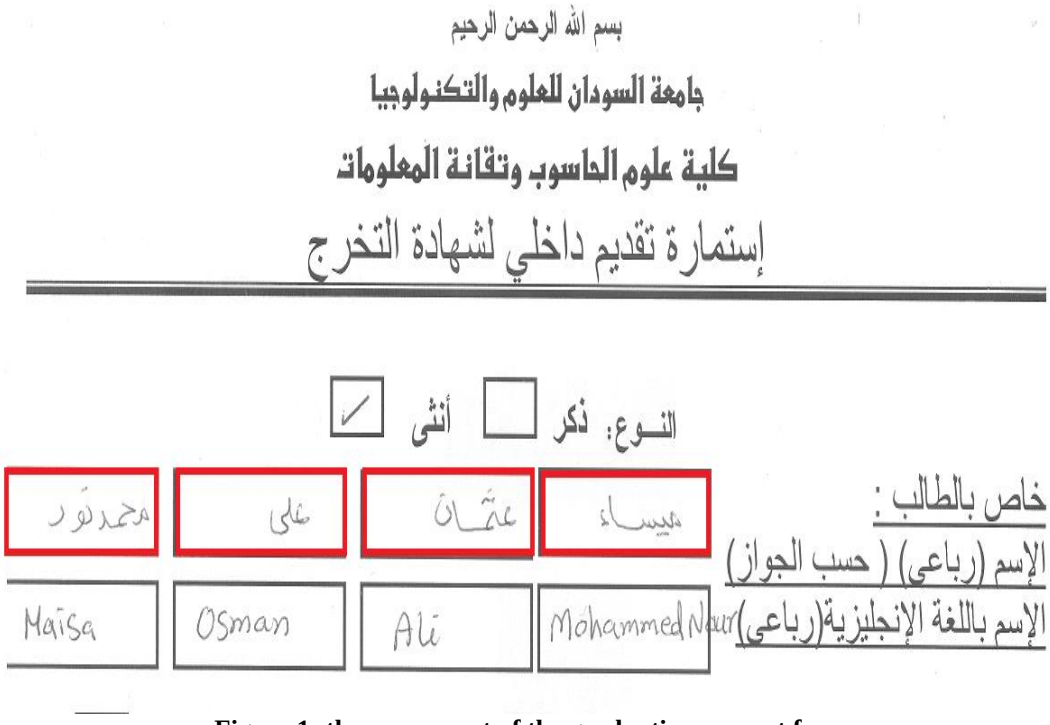


**Figure 1: the upper part of the graduation request form**

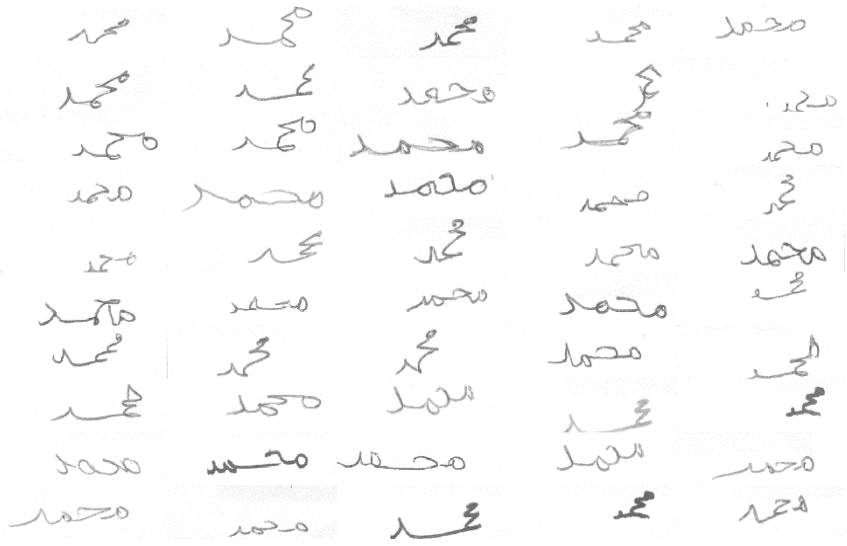

Figure 2: examples from SUST ALG data set for the name Mohamed written in ARABIC

## A. Details of the Data Set and Statistics

The data consist of 8028 names written by 2007 students at the University of Sudan for Science and Technology. Table (1) shows the twenty names that are more repetitive and the frequency of their repetition in the data set.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11705

## III. THE RECOGNITION SYSTEM

The proposed system has two main stages illustrated bellow.

Pre-processing and feature extraction

Before feature extraction, the following image processing steps were performed to enhance the images:

- Setting the intensity of the image colors
- Noise removal
- Removal of surplus white areas of the image
- Image sizes unification

Feature extraction

The features used are the pixels representing the images. Since the number of distinctive features of the image is very large (270 * 90 = 24300), we have used PCA to extract thirty features [6].

The recognizer

Figure 3. depicts the topology of neural networks designed for classification stage for the proposed system. The network consists of four layers:

- The First layer: This is input layer. It has 30 neurons each of which represents a feature.
- The second layer: This is the Patterns layer in which there are 50 neurons for each class. As we have 20 classes this layer consists 1000 neurons.
- The third layer: This is the Class layer; in this layer there are 20 neurons, one neuron for each class. All the outcome of the second layer representing one class pattern accumulate in the neuron that represents the class in the third layer, for example, the output of the pattern of the name " Ahmad" accumulate in the neuron representing the name "Ahmed", and the weights of these neurons are considered equal and thus every class is multiplied by 1/n where n is the number of patterns per class.
- The Fourth layer: It has one neuron whose mission is to select the neuron with the largest output.

Table1: Statistics of the most common names

| Percentage | Sample | Name | # |
|---|---|---|---|
| 2.57 | 177 | إبراهيم | 1 |
| 7.33 | 504 | احمد | 2 |
| 0.90 | 62 | ادم | 3 |
| 0.94 | 65 | بابكر | 4 |
| 2.43 | 167 | حسن | 5 |
| 1.10 | 76 | حسين | 6 |
| 0.87 | 60 | سليمان | 7 |
| 2.95 | 203 | عثمان | 8 |
| 3.58 | 246 | علي | 9 |
| 1.45 | 100 | عمر | 10 |
| 13.40 | 922 | محمد | 11 |
| 1.00 | 69 | محمود | 12 |
| 1.22 | 84 | مصطفى | 13 |
| 0.87 | 60 | موسى | 14 |
| 1.29 | 89 | يوسف | 15 |
| 1.46 | 101 | حامد | 16 |
| 1.31 | 91 | صالح | 17 |
| 3.52 | 242 | ابوبكر | 18 |
| 1.49 | 104 | طارق | 19 |
| 1.32 | 96 | الامين | 20 |

 This work is licensed under a Creative Commons Attribution 4.0 International License 

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11705

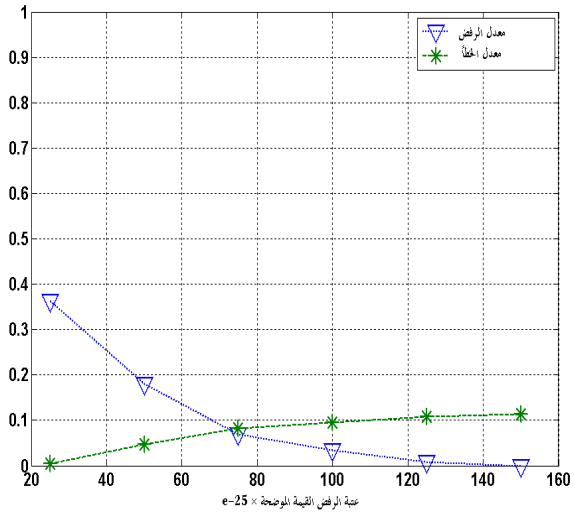


Figure 3: this figure shows the relationship of error rate to the threshold value.

The Probabilistic Neural Network is based on statistical concepts which are both Bayes theory for classification and the Probability Density Function [7, 8]. The classification based on comparing density function of the classes and then the class with the highest density function will be adopted. The following equation represents the processes of the third and fourth layers.

$$w\_c = \max_{j=1..c} \left\{ \sum_{i=1}^{k} \exp[\,(x^T w_{ji} - 1)/\sigma^2\,] \right\}$$

Where:

| | |
|---|---|
| w_c | : The winner class. |
| c | : The number of classes (15 in the network used) |
| k | :The number of patterns of the specified class (50 in the network used) |
| **X** | : Features (30 in the network used) |
| **W** | weights of the second layers, which are the features of training patterns (matrix dimensions 30× 15 × 50) |
| σ | coefficient of smoothing |

## IV. EXPERIMENTS AND RESULTS

Table (1) represents the most frequent Twenty Arabic names according to our data set; these names represent 41.9% of the data set. Each name represented by at least 60 samples in the data set. Therefore, we have used 50 samples from each name in the network, and the 10 samples from each name for testing. This process was repeated 10 times, and the average error was 0.11±0.03. It is known that in such systems all classes can give very small probability values, and consequently, if the chosen class has low probability, the validity of whether the sample belongs to it is questionable. Thus, in such a case it is preferable to choose a rejection threshold whereby the system rejects the classification if the probability of the winner class is less than this threshold and therefore the results of the classification will be more valid and reliable. To determine this threshold value, we repeated the experiments six times each time the threshold value is increased. The summary of this result is depicted in Table 2.

We noticed that when the rejection amounts is set to 1e-25 the network rejects one third, but at the same time the error rate becomes less than 1%. Figure 3, shows the relationship of error rate to the threshold value.

Other experiments were carried out by selecting 50 samples from SUST-ARG names data set outside the 20 names. The correct decision for these samples is to reject them altogether. Using very small rejection threshold of (1e-125), the system classified all the names and when the level of rejection was raised gradually upto 1e-25; the network rejected all the 50 names except only one name.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022
DOI: 10.17148/IJARCCE.2022.11705

## V. CONCLUSION AND FUTURE WORK

This paper used the Holistic Approach for recognizing hand-written Arabic names. Where the Probabilistic Neural Network was used for the recognition. The paper also presented a new data set related to handwritten Arabic Names (SUST-ARG-names). The goal of this paper is to study the idea of building a system which recognizes very quickly the common Arabic names. The system resorts to the Analytical Approach if it fails to recognize the name as a common name. A Probabilistic Neural Network was used in system to holistically recognize 15 names whose frequency in our data set is 60 and above and represents 41% of the whole data set. By choosing an appropriate rejection threshold, the network was able to recognize 64% of the 20 names where the error was less than 1%.

The results of the experiment are an encouragement for moving ahead in the design of a comprehensive system which quickly recognizes the common Arabic words and leaves the process of identifying less common ones to another system. The other system segments the word into letters to recognize each letter alone. The Holistic Approach can also help in the compound names whose first element is preceded by Abd (Slave of God) and nicknames. All these problems are issues for further research in this topic.

Table 3: Results of the experiments with different threshold.

| Rejection threshold | Error | Rejection |
|---|---|---|
| 1 e-150 | 0.02 0.114±0.03 | 0.98 0 |
| 1 e-125 | 0.44 0.108 ±0.04 | 0.56 0.008 |
| 1 e-100 | 0.84 0.096 ±0.04 | 0.1600 0.034 |
| 1 e-75 | 0.96 0.082 ±0.03 | 0.0400 0.069 |
| 1 e-50 | 1 0.048 ±0.02 | 0 0.181 |
| 1 e-25 | 1 0.004 ±0.00 | 0 0.364 |

## REFERENCES

[1] R. Plamondon and SN Srihari " On-Line and Off-Line Handwriting Recognition :A Comprehensive Survey ," IEEE Trans.Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 68-89, Jan.. 2000.

[2] Tal Steinherz, Ehud Rivlin, Nathan Intrator, "Offline cursive script word recognition-a survey," International Journal on Documents Analysis and Recognition (IJDAR), September 1999.

[3] A. Amin , "Off-line Arabic character recognition: the state of the art" , Pattern Recognition 31 (5) (1998) 517) 530. Madhvanath S., Govindaraju V., The Role of Holistic Paradigms in Handwritten Word Recognition, IEEE Trans. On Pattern Analysis and Machine Intelligence, On Pattern Analysis and Machine Intelligence, vol. vol. 23no.2, February 2001. 23 no.2, February 2001.

[4] A. K. Jain, R. P. Duin, and M. MapJianchang "Statistical Pattern Recognition: A Review," IEEE Trans.Pattern Analysis and Machine Intelligence, Jan.y

[5] C. Bishop, Pattern Recognition and Machine Learning. Springer, 2006

[6] R. Callan, The Essence of Neural Networks ,Prentice Hall Europe 1999

[7] P. Picton, Neural Networks, Palgrave Publisher Ltd, 2nd, 2000.

[8] Mohamed EM Musa. Towards building standard datasets for arabic recognition. International Journal of Engineering and Advanced Research Technology (IJEART), 2(2), 2016.

[9] Murtada Khalafallah Elbashir and Mohamed Elhafiz Mustafa. Convolutional neural network model for arabic handwritten characters recognition. International Journal of Advanced Research in Computer and Communication Engineering, 7(11), 2018

[10] Mohamed E. Mustafa, Murtada K. Elbashir " A Deep Learning Approach for Handwritten Arabic Names Recognition", International Journal of Advanced Computer Science and Applications,Vol. 11, No. 1, 2020

[11] Mohamed Elleuch, Najiba Tagougui, and Monji Kherallah. Arabic handwritten characters recognition using deep belief neural networks. In 2015 IEEE 12th International Multi-Conference on Systems, Signals & Devices (SSD15), pages 1–5. IEEE, 2015