World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:3, 2007

Arabic Character Recognition using Artificial Neural Networks and Statistical Analysis

Ahmad M. Sarhan, and Omar I. Al Helalat

***Abstract*—**In this paper, an Arabic letter recognition system based on Artificial Neural Networks (ANNs) and statistical analysis for feature extraction is presented. The ANN is trained using the Least Mean Squares (LMS) algorithm. In the proposed system, each typed Arabic letter is represented by a matrix of binary numbers that are used as input to a simple feature extraction system whose output, in addition to the input matrix, are fed to an ANN. Simulation results are provided and show that the proposed system always produces a lower Mean Squared Error (MSE) and higher success rates than the current ANN solutions.

***Keywords*—**ANN, Backpropagation, Gaussian, LMS, MSE, Neuron, standard deviation, Widrow-Hoff rule.

I. Introduction

THE goal of a letter recognition system is to transform a text document typed on paper into a digital format that can be manipulated by word processing software. The system is required to identify a given input data/letter form by mapping it to a single letter in a given letter set. This process can be quite involved since there are several valid forms that a letter may take. This is largely due to the many fonts and styles (bold type, italic type, etc.…) that can be used. The motivation behind developing letter recognition systems is inspired by their wide range of applications including archiving documents, automatic reading of checks, and number plate reading. In this paper, a new scheme for the recognition of typed Arabic letters is developed. The proposed method is compared with the existing current solution to the problem. Simulation results prove that the proposed system always produces a lower MSE than the standard ANN-based approach.

A. Arabic Letter Recognition

Arabic belongs to the group of Semitic alphabetical scripts in which mainly the consonants are represented in writing, while the markings of vowels (using diacritics) is optional and is rarely used. Arabic is spoken by more than 300 million people and is the official language of many countries.

Enormous amount of research has been undertaken in the field of recognizing typed and handwritten Latin, Chinese, and Indian letters. Little progress, however, has been made in the recognition of Arabic letters, mainly due to their cursive nature. Unlike most of the other languages, both typed and hand-written Arabic letters are cursive. Furthermore, Arabic letters can take more shapes than Latin letters.

Other problems facing Arabic letter recognition systems include:

a) The unevenness of Arabic fonts; i.e., a certain letter in a specific font can be misinterpreted as a different letter in another font. In Arabic, some letter pairs may be combined together to form another letter, that is often referred to as a ligature. The only mandatory ligature is the (Lam Alef). Other ligatures are optional. Ligatures greatly complicate the segmentation task of an Optical Character Recognition (OCR) system.
b) Arabic has 28 letters, each of which can be linked in three different ways or separated depending on the case. Therefore, each letter can have up to four different forms depending on its position.
c) Arabic letters have different heights, which puts an extra burden on the noise detection task of the OCR system.
d) Line mingling, a phenomenon exhibited by improperly spaced documents.

II. Artificial Neural Networks

ANNs were introduced by McCulloch and Pitts in 1943. ANNs are trainable algorithms that can "learn" to solve complex problems from training data that consists of a set of pairs of inputs and desired outputs (targets). They can be trained to perform a specific task such as prediction, and classification. ANNs have been applied successfully in many fields including speech recognition, image processing, and adaptive control.

An ANN consists of interconnected processing elements called neurons that work together to produce an output.

A. Single Neuron and the Least-Mean-Square (LMS) Algorithm

The neuron (Fig. 1) is the basic building block of an ANN.

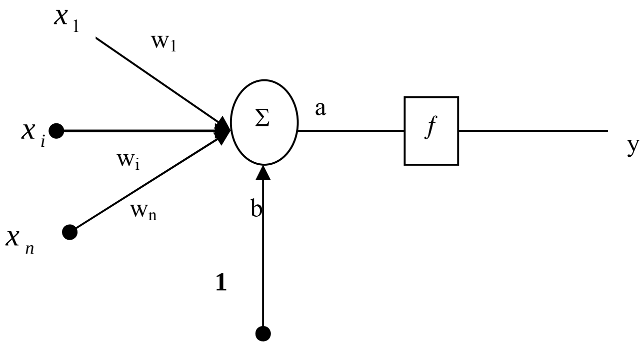


Fig. 1 The structure of a single neuron

The output a, of the neuron is a weighted linear combination of its inputs. The function *f* in Fig. 1 is called the transfer or scaling function. Some commonly used transfer functions are shown in Fig. 2.

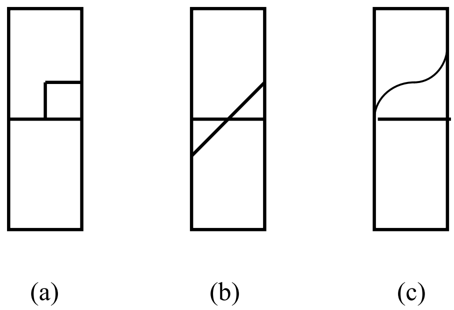


Fig. 2 Transfer functions
(a) hard line (b) pure line, and (c) logsigmoid

Using training data (input—target pairs), the weights of the neuron can be iteratively adjusted to give local or global optima. Optimum weights in the sense of Least Squared Errors were derived by Widrow and Hoff [9] and the algorithm was called the LMS algorithm and is commonly known as the *Widrow-Hoff rule* and has become a widely accepted algorithm .In the LMS algorithm, the network weights are moved along the negative of the gradient of the performance function. Specifically, after each iteration or epoch (new set of input—target pairs) the weights are adjusted according to the following rule

$$\mathbf{W} \leftarrow \mathbf{W} + \mu\, e\, \mathbf{X}, \qquad (1)$$

where μ is the learning /adaptation speed, and the input vector $X \varepsilon R^n$ is given by

$$\mathbf{X} = [x_1, x_2, x_3, x_4, x_5,\ldots, x_n]^T, \qquad (2)$$

and $W \varepsilon R^n$ is the vector of weights and is given by

$$\mathbf{W} = [w_1, w_2, w_3, w_4, \ldots, w_n]^T. \qquad (3)$$

The output a of the neuron is defined by the following expression

$$a = w_1 x_1 + w_2 x_2 + w_3 x_3 + \ldots + w_n x_n + b, \qquad (4)$$

where *b* is a bias value that is not part of the input. The neuron's scaled output *y* is given by *y = f (a)*, where *f* is the transfer function (Fig. 2). The error e is the difference between the neuron's output and the desired output.

The initial values of the weights can be set explicitly if apriori information is available. Alternatively and in most practical cases, the weights are initially set to zeros or some random values.

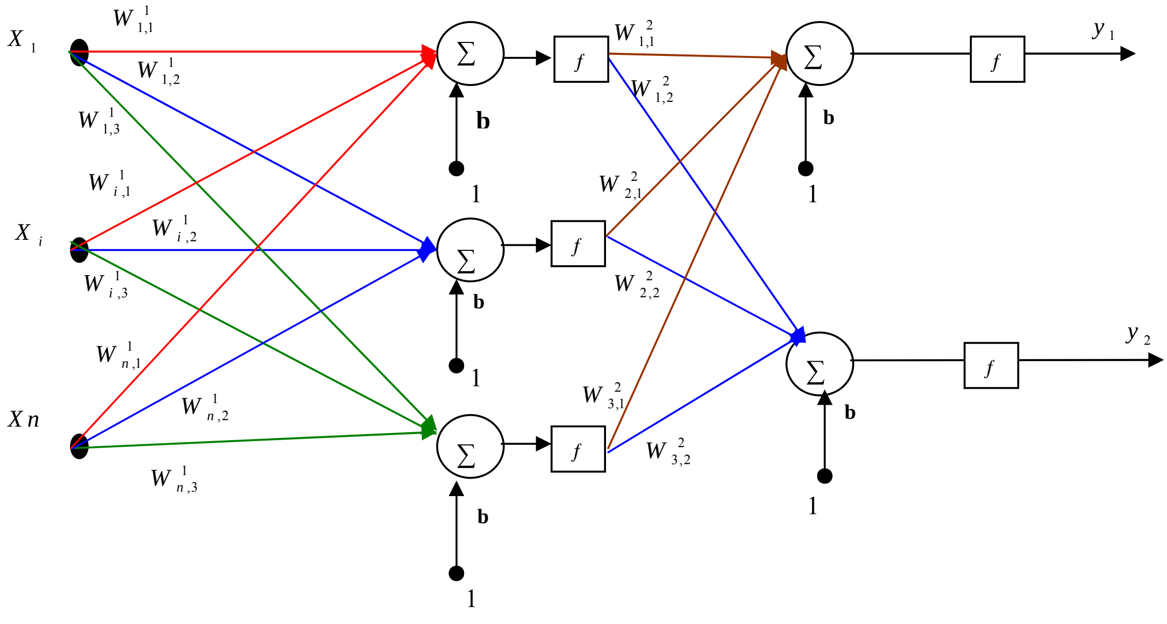


Fig. 3 A two-layer ANN

B. Multilayer ANN

In general, a multilayer ANN has the architecture depicted in Fig. 3, which shows a two-layer network.

ANN is designed so that the application of a set of inputs produces a desired set of outputs. One of the main letteristics of an ANN is its set of weights.

Note that $w^{k}_{i,j}$ in Fig. 3 denotes the weight for the i^{th} input in the j^{th} neuron of the k^{th} layer.

C. Backpropagation Algorithm

When a multilayer ANN uses the *Widrow-Hoff* learning rule and nonlinear differentiable transfer functions, the resulting ANN, known as Backpropagation, can approximate almost any function that has a finite number of discontinuities. Properly trained backpropagation networks have proven to give reasonable answers when presented with inputs that they have never seen. In many cases, it is possible to train a network on a representative set of (input—target) pairs and get good results without training the network on all possible (input—target) pairs.

III. Methods

As shown in Fig. 5, the proposed system, referred to henceforth as *netStd,* consists of two stages, a feature extraction stage followed by an ANN which constitutes the second stage. Fig. 6 shows the current approach to the Arabic letter recognition problem, called here *netOld*. As seen from Fig. 6, netOld bypasses the feature extraction stage and depends entirely on the ANN to achieve the recognition task. In the simulation section, a performance comparison between the two systems is presented.

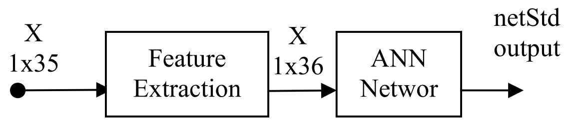


Fig. 5 Flow diagram of the proposed network, netStd

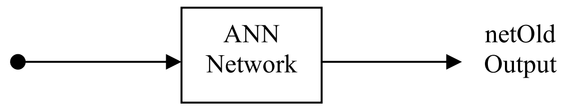
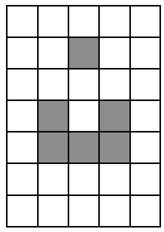


Fig. 6 Flow diagram of the standard network, netOld

A. Letter Modeling

In the proposed system, each letter is represented by a matrix of 7 x 5 binary pixels, producing a 35-element input vector. As an example, Fig. 7 shows the binary image of the Arabic letter "Noon" and its corresponding matrix representation. The number of Arabic letters is 28, producing 28 different classes or system outputs. A careful examination of the matrix representations of the Arabic letters shows that the pixel values are highly uncorrelated. A graph showing the standard deviation of the matrices of the Arabic letters is shown in Fig. 8, which reveals that there is a considerable variance in the standard deviations of the Arabic letters. This fact instigated the use of the standard deviation as a valid feature that can significantly help the ANN in its classification job.

The flow of the system is as follows. The input to the feature extraction stage consists of a vector of 35 elements representing an Arabic letter. The output of the feature extraction stage is a 36-element vector, with the additional element being the standard deviation of the original 35 elements. The 36-element vector is the fed as input to the ANN stage.



(a)

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

(b)

Fig. 7 The Arabic letter "Noon" (a) Bit map image, and (b) matrix representation

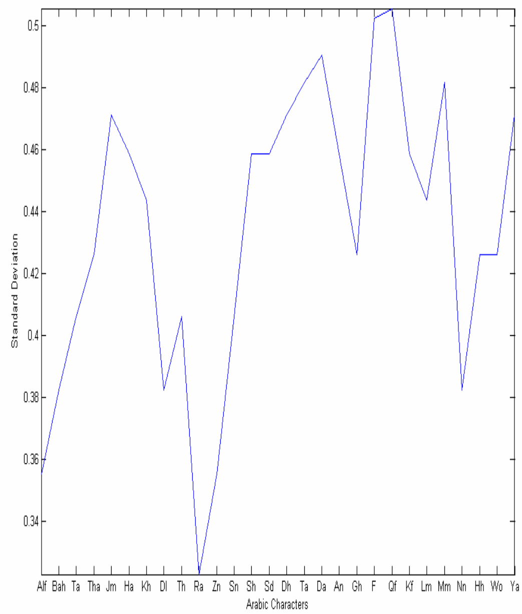


Fig. 8 Standard deviation of the Arabic letters

B. ANN Design

It was shown in [1] and [17] that the best ANN solution for the problem at hand is a two-layer network (1 hidden and 1 output) with 10 neurons in the hidden layer. The same network is also used here for both netOld and netStd systems.

Furthermore, log-sigmoid functions, Eq. (5), were used as the transfer functions of the output layer since they can approximate binary values.

$$\log sig(n) = \frac{1}{1+e^{-n}} \tag{5}$$

The network receives the 36 Boolean values as a 36-element input vector. It has 28 outputs, where each output corresponds to an Arabic letter.

IV. Simulations

netOld and netStd systems were trained with sets of noisy and clean inputs. The contaminating noise was the Gaussian noise which has a probability density function (pdf) given by

$$f(x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}} \tag{6}$$

where μ and σ are the mean and standard deviation of the noise, respectively.

Both networks were then presented with new sets of noisy inputs, where the noise was additive Gaussian noise, Eq. (6). Fig. 7 shows the performance of both networks when presented with Gaussian noise of zero mean and varying standard deviation (noise level). As can be seen from Fig. 9, netStd always outperforms netOld and produces a lower MSE.

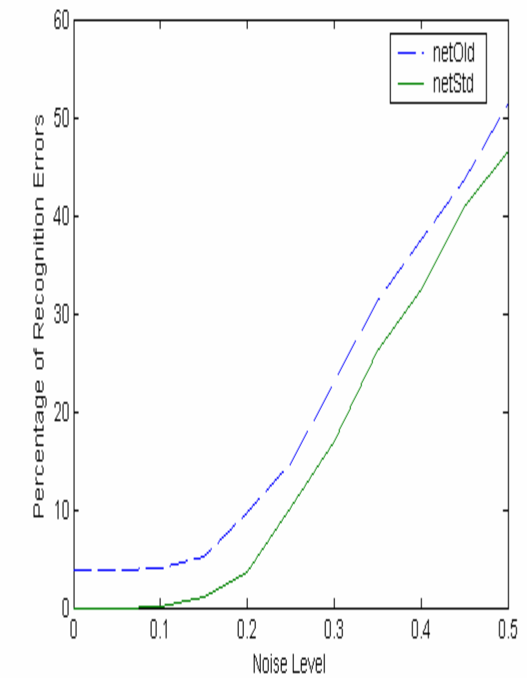


Fig. 9 MSE for netOld and netStd

V. Conclusion

In this paper, a new system for the recognition of typed Arabic letters is presented. The system is composed of two stages, a feature extraction stage, followed by an ANN classification stage. The performance of the system was compared to the traditional solution that bypasses the feature extraction stage and consists only of an ANN stage.

A statistical analysis on the Arabic letters was performed and showed that the pixel values of the Arabic letters are highly uncorrelated. That fact provoked the investigation of the use of the standard deviation as a distinctive feature of the letters. It was shown that the standard deviation of the pixel values provides distinctive information about the letters.

The system inputs consist of the 28 typed Arabic letters. Each letter is represented by a matrix of 7 x 5 binary pixels, which is presented to a feature extraction stage. The original matrix, in addition to the one element generated by the feature extraction stage, composes a 29 – element vector which is then fed to an ANN stage.

The ANN stage is based upon the traditional ANN used in character recognition. Specifically, the ANN is composed of two layers. The first layer consists of 10 neurons and the second layer consists of 28 neurons, the number of Arabic letters.

In comparison with the traditional ANN solution, the additional standard deviation element has proven to significantly help the ANN reach its classification goal. Simulation results indicate clearly that the proposed system always produces a lower MSE and higher success rates than the current ANN solutions.

References

[1] Amin and G. Masini, "Machine recognition of multifont printed Arabic texts", in Proc.8th Int. Conf. Patt. Recogn. (Paris, France), pp. 392-395, 1986.

[2] A.M. Sarhan and R. C. Hardie, "Partition-based filters", In *Proceedings of the 1995 IEEE National Aerospace and Electronic Conference (NAECON),* volume 2, Dayton, Ohio, May 1995.

[3] A.M. Sarhan, R. C. Hardie, and K. E. Barner, "Partition-based adaptive estimation of single-response evoked potentials", In *Proceedings of the 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* May 1995.

[4] A.M. Sarhan. N*onlinear partition-based filters for signal restoration.* Ph.D. Thesis University of Dayton, July 1996.

[5] D. J. Hand. *Discrimination and Classification*. New York: Wiley, 1981.

[6] E. W. Brown, "Letter Recognition by Feature Point Extraction", Northeastern University internal paper, 1992.

[7] F. Hussain and J. Cowell, “Character Recognition of Arabic and Latin Scripts”, *Proceedings, IEEE International Conference on Information Visualization*, pp. 51–56, 2000.

[8] H. Al-Yousefi and S. S. Udpa, "Recognition of handwritten Arabic characters," in Proc. SPIE 32nd Ann. Int. Tech. Symp. Opt. Optoelectric Applied Sci. Eng. (San Diego, CA), Aug. 1988.

[9] Haykin S.,Adaptive Filter Theory, Englewood Cli s.N.J: Prentice Hall(3ed), 1996.

[10] J.F Canny. “A Computational Approach to Edge Detection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-6, pp. 679-698, 1986.

[11] K. E. Barner, A. M. Sarhan, and R. C. Hardie. "Partition-based weighted sum filters for image restoration". *IEEE Transactions on Image Processing,* Vol. 8, No. 5, May 1999.

[12] K. Khatatneh ,"Probabilistic Artificial Neural Network for Recognizing the Arabic. Hand Written Characters", Journal of Computer Science 3 (12), 881-886, 2006.

[13] K. Badi and M. Shimura, "Machine recognition of Arabic Cursive Script" Trans. Inst. Electron. Commun. Eng., Vol. E65, no. 2, pp. 107-114, Feb. 1982.

[14] K. Badi and M. Shimura, "Machine recognition of Arabic cursive scripts" in Pattern Recognition in Practice. Amsterdam: North Holland, 1980.

[15] L. Hammami and D. Berkani ,"Recognition system for printed multi-font and multi-size Arabic characters", The Arabian Journal for Science and Engineering, Volume 27, Number 1B, pp:57-72, April, 2002.
[16] M. Altuwaijri, M. A. Bayoumi, "Arabic Text Recognition Using Neural Network" ISCAS 94. IEEE International Symposium on Circuits and systems, Volume 6, 30 May-2 June 1994.
[17] N. Ben Amor, N. Essoukri Ben Amara: "A hybrid approach for Multifont Arabic Characters Recognition", 5th WSEAS Int. Conf. On Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED'06) Madrid, Spain, February 15-17, 2006.
[18] R. A. Dosari, R. C. Hardie, and A. M. Sarhan. "Multi-channel nonlinear filters for signal restoration". In *Proceedings of the 1997 IEEE National Aerospace and Electronic Conference (NAECON),* Dayton, Ohio, May 1997.

Ahmad M. Sarhan was born in Jordan in 1970. He received a BSc. in biomedical engineering form Syracuse University, NY, USA in 1991 and MS. in electrical engineering from Syracuse University in 1992, and a Ph.D. in electrical engineering from University of Dayton, Ohio, USA. .He is currently an associate professor of Engineering and Computing Sciences at New York Institute of Technology, Abu Dhabi. He has worked at several universities including Yamouk University, Jordan, Amman University, Jordan, Ajman University, UAE, Higher Colleges of Technology, UAE. He also worked in the industry as an executive manager of a medical company, UAE. He received several rewards including, The Best Student Paper Award for the paper “Partition-Based Filters,” the Most Outstanding Paper Published in the NAECON Conference, Dayton, Ohio, May 1995. Dr. Sarhan received a Tuition scholarship from Syracuse University and a Teaching and Research Assistantship from the University of Dayton. His main research interest includes adaptive signal processing and pattern recognition. is a member of several organizations.

Omar Al Helalat has been working for the Ministry of Education, U.A.E since 2004, he received a BSc in Mathematics from U.A.E.University, and MSc in Computer Science from New York Institute of Technology in 2007.His research interests are: Pattern Recognition and Neural Networks.