International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:5, 2008

A New Recognition Scheme for Machine-Printed Arabic Texts based on Neural Networks

Z. Shaaban

***Abstract*—** This paper presents a new approach to tackle the problem of recognizing machine-printed Arabic texts. Because of the difficulty of recognizing cursive Arabic words, the text has to be normalized and segmented to be ready for the recognition stage. The new scheme for recognizing Arabic characters depends on multiple parallel neural networks classifier. The classifier has two phases. The first phase categories the input character into one of eight groups. The second phase classifies the character into one of the Arabic character classes in the group. The system achieved high recognition rate.

***Keywords*—**Neural Networks, character recognition, feature extraction, multiple networks, Arabic text.

I. Introduction

OPTICAL character recognition(OCR) is one of the most successful applications of image processing and pattern recognition field. A lot of work has been done on English texts but a little work on Arabic texts because of the difficulty of segmenting cursive Arabic words into isolated characters. There are some differences between Arabic language and others. First, Arabic language is written from right to left. Second, An Arabic word consists of one or more cursive sub-words, each consists of one or more characters. Third, some Arabic characters have exactly the same shape.

Many Arabic researchers have done some works on online and offline Arabic cursive texts. They found that the segmentation process is the most important step to achieve good results in recognition. In reference [3], a review for the recognition of Arabic texts is given. A combination of several features (multi- features) with multi-stage classifiers is a very reliable approach for a character classification [4]. Many researchers have been interested in building multiple classifiers to achieve better performance [5].

Kimura and others [4,6] have successfully applied a combination of several classifiers for recognizing totally unconstrained hand-written numerals. The experimental results showed that die performance of the classifiers has improved significantly.

Kimura and others [6] achieved good results by a combination of a statistical classifier using a modified quadratic discriminant function with features derived from chain codes of the character contour and a tree classifier based on structural features. Suen [12] suggested a multiple expert system that is a combination of four classifiers based on the understanding of individual opinions observed in human interaction. They achieved a higher recognition rate. Srihari [11] achieved a good performance on a combination of template matching, mixed statistical-structural classifier based on character contours and structural classifier using features such as strokes and holes.

Also, Srihari [11] combined several neural networks that use five different features. The result of using different features to feed different networks is better than using single features. Cao and others [4] proposed a multistage expert system for hand-written numerals using neural networks. It is based on a modified directional histogram feature extraction method. The proposed system has produced a high performance by using multiple sub-classifiers.

This paper suggests a new scheme for the recognition of machine-printed Arabic characters [14]. There are three main advantages of this scheme. First, the proposed scheme utilizes two stages of classifiers. Second, it is also uses combined features and multiple classifiers to increase the recognition rate of the system. Third, the Arabic characters are divided into eight groups to reduce the learning time of the network and to speed up the recognition process. This paper describes the multiple parallel neural networks applied in recognition of cursive Arabic characters.

The rest of the paper is organized as follows : Section II gives the preprocessing stage of the system. Section III shows the proposed multi-font Arabic character recognition scheme. In section IV, Results and conclusions are presented.

II. Preprocessing

The preprocessing stage concentrates on the followings:

1. scanning the text using 300dpi and digitizing it into a digital image.
2. converting the gray-scale image into binary image.
3. normalizing the text
4. Segmenting the text into lines and the lines into words as shown in Fig. 1.

Fig. 1 Segmented Arabic text

5. Separating the words into characters.

The whole preprocessing procedure is shown in Fig. 2.

III. Recognition System

At the beginning of the work using the back-propagation algorithm, the network is performed with 32 by 32 inputs(binary image(0,l)) in the input layer and 50 hidden nodes and 28 nodes in the output layer. The number of training data was one sample per class. The network did not converge, even though it was trained for a long time. Based on method of divide and conquer, the problem is divided into sub-problems. The proposed approach, which has two stages of classification, is shown in Fig. 3. The two-stage classifier separates the classification problem into different individuals. To gain more efficient performance, the characters are divided into several groups using some knowledge about the similarity between the characters. The design of the classifier is made based on (i) a similarity of characters, (ii) a multistage classification and (iii) a combination of different types of features and classifiers.

This type of neural network approach, where the mapping from inputs to outputs is achieved with different smaller networks, needs less learning time than approaches that use a single network.

The Arabic characters are divided into several groups based on the observation of similar characters. The 28 characters are separated into the following eight groups: G1 = { أ، ك، ل}, G2 = { ب،ت،ث،ن،ي،ى }, G3 = { ح،ج،خ }, G4 = { د،ذ،ر،ز،و },G5 = { س،ش }, G6 = { ص،ض،ط }, G7 = { ع،غ،ف،ق }, G8 = { م، هـ }.

A new design of machine-printed Arabic character classifier is proposed as shown in Fig. 3. The classifier is concerned with combined features(i.e., Hu moments, visual features and SVD eigen values) proposed in [7,8,9,10]. It is a multistage classifier that has two stages. The first stage that is based on combined features extracted from the normalized character is a neural network classifier. It performs as a clustering technique to split the input character images into one of eight character groups based on the character confusion problem. The second stage has eight sub-classifiers. Each sub-classifier is a multiple parallel neural network (PNN) classifier. Each of the PNN has two networks work in parallel. The inputs for the first and second networks are the moments invariants and svd features extracted from the normalized original binary image (i.e., the character here is normalized in terms of scaling, translation and slanting, respectively). The final decision is made by picking up the largest value of the averaged output values of the two networks.

IV. Results and Conclusion

The proposed system has been tested on more than 100 Arabic text images. The recognition rate was very high (98%).

This paper described a new approach for the recognition of multi-font Arabic texts. It depends on multiple parallel neural networks. The classifier achieved very high recognition rate. I would suggest that this approach could be extended for hand-written Arabic text.

Acknowledgment

This paper received financial support towards the cost of its publication from the Deanship of Research and Graduate Studies at Applied Science University, Amman, Jordan.

References

[1] I.S. Abuhaiba and P. Ahmed, "Restoration of Temporal Information in Off-Line Arabic Handwriting," Pattern Recognition(26), No. 7, July 1993, pp. 1009-1017.

[2] M. Altuwaijri and A. Bayoumi, "A new recognition system for multi-font Arabic Cursive words, " Proceedings of ICECS'95, Amman-Jordan, pp.298-303,1995.

[3] A. Amin, H.B. Al-Sadoun and S. Fischer, "Hand-Printed Arabic Character-Recognition System Using an Artificial Network," Pattern Recognition(29), No. 4, April 1996, pp. 663-675.

[4] J. Cao, M. Ahmadi and M. Shridhar, "Recognition of handwritten numerals with multiple feature and multistage classifier," Pattern Recognition,1995, vol.28, no.2, pp. 153-160.

[5] S. Günter and H. Bunke, "Multiple classifier systems in off-line handwritten word recognition - on the influence of training set and vocabulary size," Int. Journal of Pattern Recognition and Art. Intelligence, 2004, vol. 18, no. 7, pages 1303 - 1320.

[6] F. Kimura and M. Shridhar, "Handwritten numerical recognition based on multiple algorithms," Pattern Recognition,1991, vol.24, no.lO.pp.969-983.

[7] Z. Shaaban and G. Sulong, "Uppercase hand-printed character recognition using parallel neural architecture," Proc. of the 1ASTED International conference modeling and simulation Pittsburgh-USA, 1995 pp.307-309

[8] Z. Shaaban and Z. Sulong, "Recognition of connected handwritten characters based on moments invariants using neural networks," Proceedings of ACCV'95 Second Asian Conference on Computer Vision, 1995, Singapore pp.l- 335-339

[9] Z. Shaaban, G. Sulong and B. Duin, "Recognition of handprinted characters using distance transform and moment invariants via parallel neural networks," Proceedings of ICECS'95 International conference on electronics, circuits and systems '95, 1995, Amman-Jordan, pp.393-398.

[10] Z. Shaaban, G. Sulong and B. Duin, "Symbol recognition based on distance transform," Proc. of the IASTED International conference on signal and image processing and applications, 1996 Annecy-France, pp.225-230.

[11] S. N. Srihari, "Recognition of handwritten and machine printed text of postal address interpretation," In ([4]).

[12] C. Y. Suen, "Computer recognition of unconstrained handwritten numerals," Proceedings of the IEEE vol.80, no.7, pp.1162-1180,1992.

[13] L. Xu, A. Krzyzak and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," IEEE transactions on systems, man, and Cybernetics vol.22, no.3, pp.418-435,1992.

[14] Z. Shaaban, Algorithms for off-line upper-case hand-written text recognition and its associated processes, Ph.D. Thesis, University of Technology Malaysia, 1996.

Zyad M. Shaaban was born in Jordan in 1969. He received a BSc. in Computer Science from Yarmouk University, Irbid, Jordan in 1992 and a Ph.D. in Computer science from University of Technology, Johor Bahru, Malaysia in 1996. He is currently an assistant professor of computer Science at Faculty of Information Technology, Applied Science University, Jordan. Dr. Zyad received a Fellowship from University of Technology, Malaysia and he was working on handwritten text recognition project. His research interests are: Handwritten Character Recognition, Moments Invariants, Neural Networks, Face Recognition, Arabic Text Recognition.

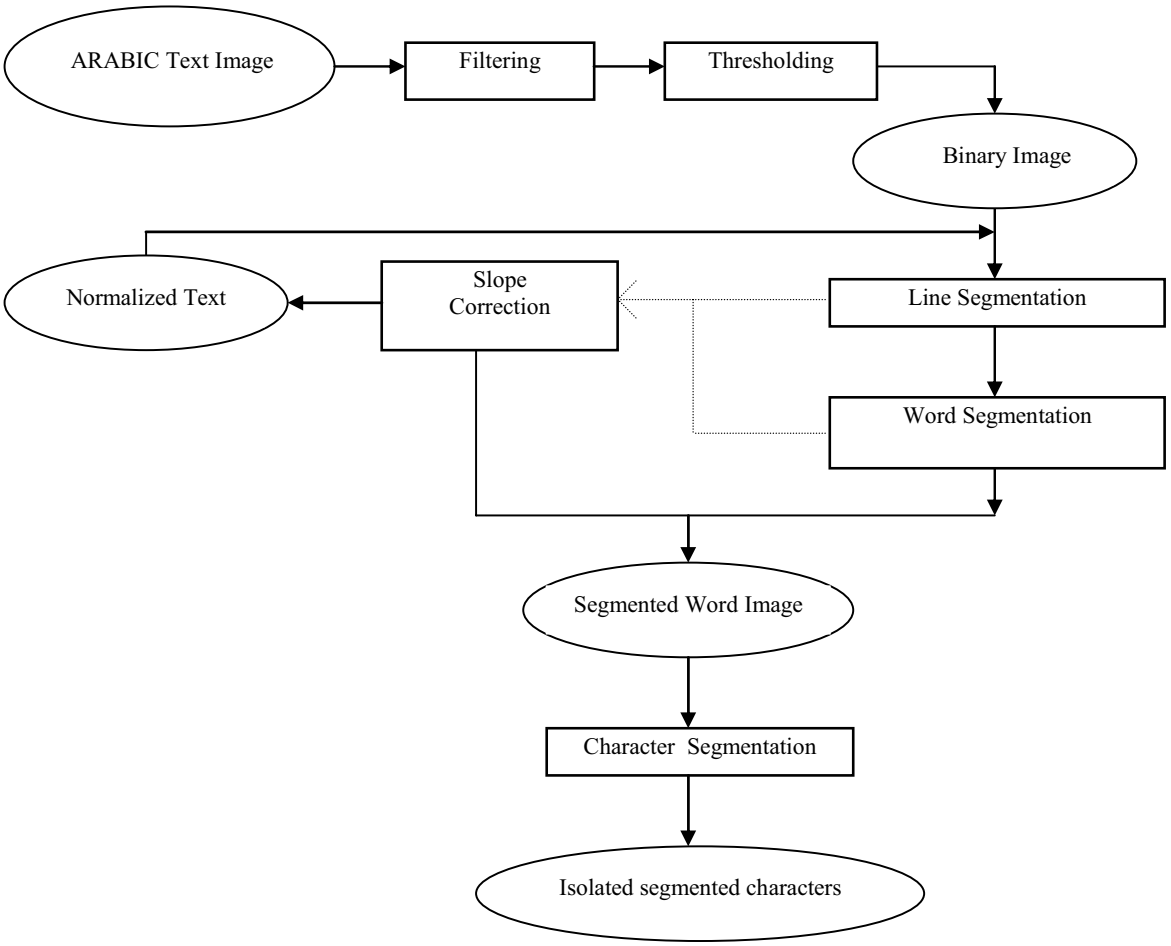

Fig. 2 Overview of the proposed text image preprocessing

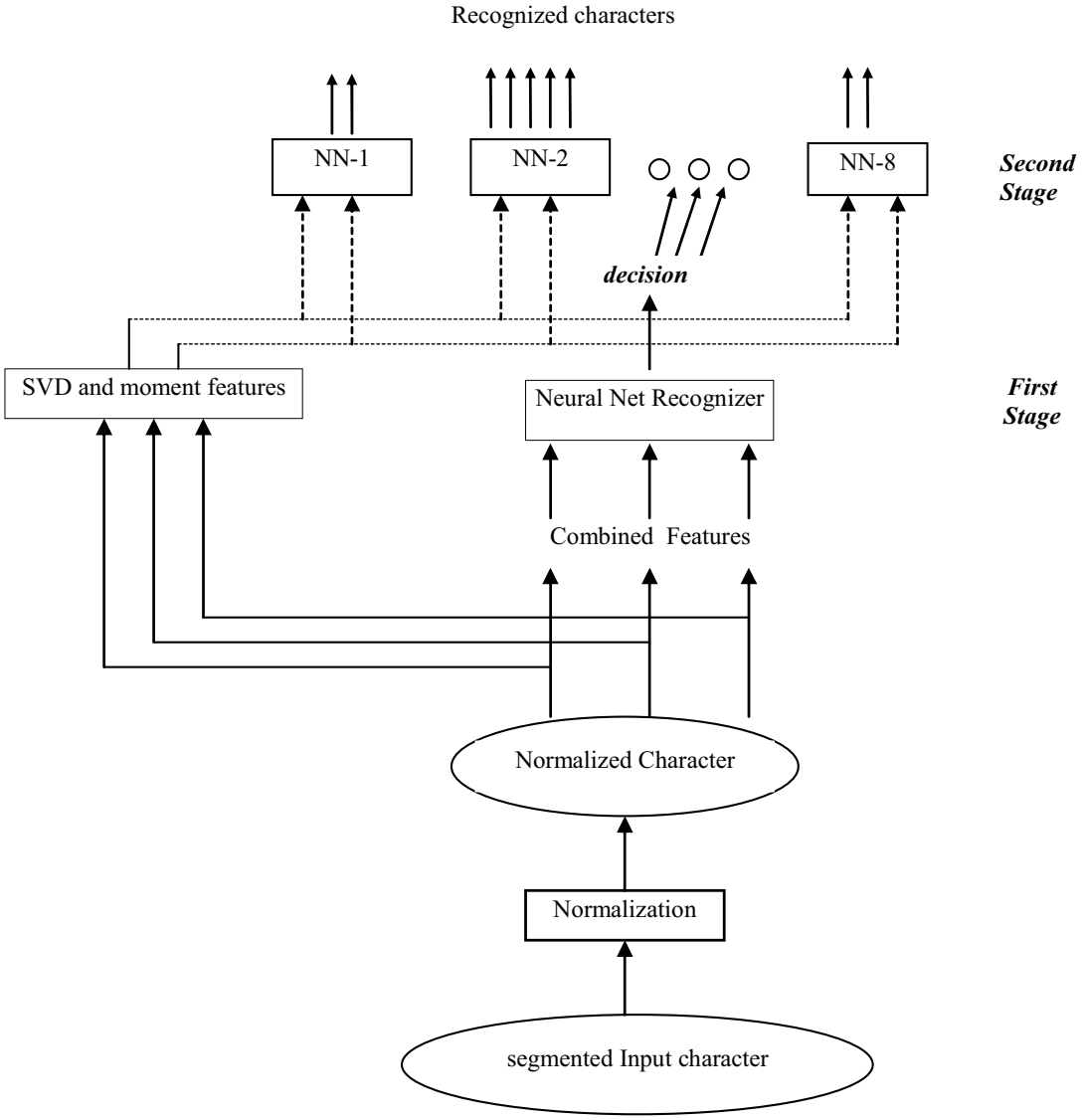

Fig. 3 A new scheme of Arabic character classifier