International Journal of Soft Computing 14 (1): 1-5, 2019
ISSN: 1816-9503

Recognition of Similar Shaped Handwritten Arabic Characters Using Neural Network

Rashad A. Al-Jawfi
Department of Mathematics, Faculty of Science and Arts, Najran University, Najran, Kingdom of Saudia Arabia (KSA)

Abstract: Recognition of Arabic handwriting characters is a difficult task due to similar appearance of some different characters. However, the selection of the method for feature extraction remains the most important step for achieving high recognition accuracy. In this study, a novel method is provided to recognize handwritten Arabic characters based on their features extraction and adaptive smoothing technique. In this study, combination of two approaches will be introduce, one of them is feature selections methods and the other is adaptive smoothing technique from smooth shape of character. Combination of both these approaches leads to the better results.

Key words: Character recognition, image segmentation, pattern matching, smoothing, handwriting, extraction, approaches

INTRODUCTION

In recent years, handwritten Arabic character recognition (as more of other languages) has grabbed a lot of attention as Arabic being primary official language in more than 20 countries and has wide applications in areas like passport, railways, postal address reading, etc. More than 200 million people speak this language as their native speaking and over 1 billion people use it in several religion-related activities.

In general, the character recognition procedure consists of two steps feature extraction where each character is represented as a feature vector and classification of the vectors into a number of classes (Kavallieratou *et al*., 2002).

Earlier, traditional classifiers such as Nearest Neighbor (NN) (Parez-Cortes *et al*., 2000; Zhang and Srihari, 2002) were adopted for character recognition, however, they exhibit certain limitations. Machine Learning (ML) algorithms (Liu *et al*., 2002) provide a promising alternative in character recognition based on the feature set given to them. A variety of features can be extracted such as primitives, profiles etc.

Literature review: Beside the main goal of any Optical Character Recognition (OCR) system which is simulating human's reading capability, the accuracy and time consuming are very important issues in this aspect. Template matching works effectively for recognition of standard fonts but gives poor performance with handwritten characters and when the size of dataset grows. It is not an effective technique if there is font discrepancy (Prasad *et al*., 2009). Based on the latest survey which is published by Lorigo and Govindaraaju (2006) all covered papers presented their proposals seeking high accuracy and less time. We can classify their research into three main categories preprocessing problems, features extraction problems and recognition problems. Many researchers have used skeletonization in their proposed preprocessing stages Mozaffari *et al*., 2005; Alma'adeed *et al*., 2002). Few used techniques such as wavelet or fractal like (Mowlaei *et al*. 2002; Mozaffari *et al*., 2004). Later, discriminative classifiers such as Artificial Neural Network (ANN) and Support Vector Machine (SVM) grabbed a lot of attention. Vamvakas *et al*. (2007) compared the performance of three classifiers: Naive Bayes, K-NN and SVM and attained best performance with SVM. However SVM suffers from limitation of selection of kernel. ANNs can adapt to changes in the data and learn the characteristics of input signal (Kahraman *et al*., 2004). Also, ANNs consume less storage and computation than SVMs (Verma, 1995; Jane and Pund, 2012) presented a system for HCR using MLP and RBFN networks in the task of handwritten Hindi character recognition where the mostly used classifiers based on ANN are MLP and RBFN. Dabra *et al*. (2011) presented a novel feature set using machine learning for recognition of similar shaped handwritten Hindi characters. Archana Jane and others (Sutha and Ramaraj, 2007) presented a novel feature set for recognition of similar shaped handwritten Marathi characters using artificial neural network.

Table 1: Arabic characters and their forms as different positions in the word

Letter	Single	Beginning	Middle	Ending
Alef	ا	ا	ا	ا
Baa	ب	بـ	ـبـ	ـب
Taa	ت	تـ	ـتـ	ـت
Thaa	ث	ثـ	ـثـ	ـث
Jeem	ج	جـ	ـجـ	ـج
Haa	ح	حـ	ـحـ	ـح
Khaa	خ	خـ	ـخـ	ـخ
Dal	د	د	ـد	ـد
Thal	ذ	ذ	ـذ	ـذ
Raa	ر	ر	ـر	ـر
Zai	ز	ز	ـز	ـز
Seen	س	سـ	ـسـ	ـس
Sheen	ش	شـ	ـشـ	ـش
Sadd	ص	صـ	ـصـ	ـص
Dadd	ض	ضـ	ـضـ	ـض
Tah	ط	طـ	ـطـ	ـط
Thah	ظ	ظـ	ـظـ	ـظ
Ayn	ع	عـ	ـعـ	ـع
Ghyn	غ	غـ	ـغـ	ـغ
Faa	ف	فـ	ـفـ	ـف
Qaf	ق	قـ	ـقـ	ـق
Kaf	ك	كـ	ـكـ	ـك
Lam	ل	لـ	ـلـ	ـل
Meem	م	مـ	ـمـ	ـم
Noon	ن	نـ	ـنـ	ـن
Ha	ه	هـ	ـهـ	ـه
Waw	و	و	ـو	ـو
Yaa	ي	يـ	ـيـ	ـي

Wakabayashi *et al.* (2009) proposed an F-ratio to improve results of similar shaped characters. Yang *et al.* (2010) proposed a method for similar handwritten Chinese character recognition. Many variations have been used in order to overcome the main disadvantage of ANNs which is time consuming. Fathy and Syiam (1996), presented a parallel design for ackpropagation neural networks approach in order to accelerate the computation process.

Arabic character characteristics: Arabic writing like English in term of using letters, numbers and punctuation but there are number of characteristics which make Arabic cursive writing is unique compared to other languages. These characteristics can be summarized as follow (Ahmed and Zakaria, 1996). Arabic is written from right to left in both printed and handwritten forms. The shape of the character varies according to its position in the word (Table 1). Each character has either two or four different forms. Off course this will increase the number of classes to be recognized from 28-100.

Arabic is always written cursively. Words are separated by spaces. There are 6 characters can be connected only from the right, these are (ا، د، ذ، ر، ز، و) and these six characters if appeared in a word will cause the word to be divided into blocks of connected components called sub words thus, a word can have one or more sub words. Sub words are also separated by spaces but usually shorter than the one between words.

Character width and character height differ from one character to another in addition to that, the width and height vary across the different shapes of the same character in different position in the word. The 15 characters have dots associated with the character, they can be above or below the primary part and some characters share the same primary part and distinguished from each other by the secondary part (the dots) (Table 1).

Alif-Maqsora (ى) shares the same primary part of character ي but without dots. This character appears only at the end of the word. Hamza (ء) is not really a letter, it is a complementary shape appears in the following cases. Always: with character ك in the separated or final forms. Here, it is used to distinguish it from letter ل. Separated: May appear at the beginning in the middle or at the end of a word. This is the only case in which the character can't be connected from both sides.

Similar character in Arabic language: In this study, combinations on pixel rating an image, feature extractions and image pattern matching are proposed. This method consists of two phases, training and testing. Training on images consists of listing all handwritten images with respect to its standard Arabic character images. We have 106 Arabic characters for training and trained 15 handwritten characters with respect to each corresponding standard Arabic character images (Table 2).

In Arabic language there are many similar characters, in this research by similarity we mean the characters with the same shape and additional point different between them (Table 2). Jane and Pund (2012) show that the no of images in training affects turnaround time of entire process execution:

$$T_i \propto N$$

Where:
T_i = Turnaround time
N = No. of training images

An accuracy of result depends on the no of trained handwritten images per standard Marathi character image:

$$A_c \propto \frac{Him}{Sim}$$

Where:
A_c = Accuracy of character recognition
Him = No. of Handwritten patterns
Sim = Standard Arabic image = 1

Testing is a phase where No. of handwritten character image is tested against training set handwritten images. Testing consists of the following phases.

Table 2: Arabic similar characters

Letter	Single	Beginning	Middle	Ending
G1				
Baa	ب	بـ	ـبـ	ـب
Taa	ت	تـ	ـتـ	ـت
Thaa	ث	ثـ	ـثـ	ـث
G2				
Jeem	ج	جـ	ـجـ	ـج
Haa	ح	حـ	ـحـ	ـح
Khaa	خ	خـ	ـخـ	ـخ
G3				
Dal	د	د	ـد	ـد
Thal	ذ	ذ	ـذ	ـذ
G4				
Raa	ر	ر	ـر	ـر
Zai	ز	ز	ـز	ـز
G5				
Seen	س	سـ	ـسـ	ـس
Sheen	ش	شـ	ـشـ	ـش
G6				
Sadd	ص	صـ	ـصـ	ـص
Dadd	ض	ضـ	ـضـ	ـض
G7				
Tah	ط	طـ	ـطـ	ـط
Thah	ظ	ظـ	ـظـ	ـظ
G8				
Ayn	ع	عـ	ـعـ	ـع
Ghyn	غ	غـ	ـغـ	ـغ

RGB-to-binary image conversion: As we maintained training images as a binary images, there is need to convert testing image into binary image. Binary image avoids unnecessary image segmentation and features extraction.

Pixel rate an image: Pixel rate an image used to identify an image pixel value either on or off. We set pixels height and width equal to 10 of which we got result shown in Table 1 height, width of an pixel rate image should be proportionate to size of an binary image. From the Table 1, it is observed that having pixels size 15×32 leads to loss of pixels which are represented with an equation:

$$P_l = \int_1^n P_n (\log_2 n)$$

Where :

P_l = Number of pixels loss

P_n = Pixels new

n = Number of off pixels in binary image

Image edge smoothing: From literature review, it is observed that patterns in handwritten characters have large deviation factor with respect to its standard image pattern. Deviation factor is a mod difference value between no of pixels patterns in training and testing images and represented with (Fig. 1 and 2):

$$D_f = | T_r - T_s |$$

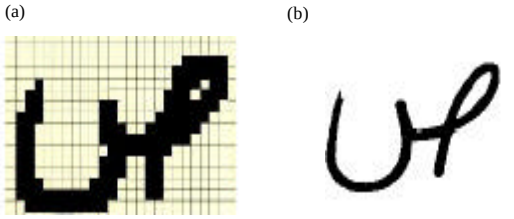

Fig. 1: Handwritten and binary; a) Input binary image and b) Input handwriteen image

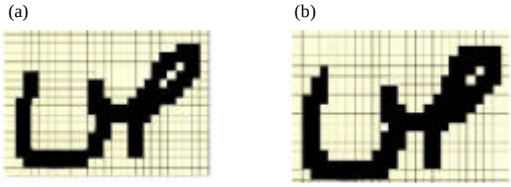

Fig. 2: Binary and pixel; a) Pixel rate image and b) Input binary image

Where

T_r = No. of segments of training image

T_s = No. of segments of testing image

Image segmentation:

- To recognize character, segmentation is done based on their patterns of size 2×2
- Pattern matching

MATERIALS AND METHODS

Proposed method: Different from the previous approaches in this study, we propose a novel algorithm based on critical regions to classify similar pairs.

Noting the fact that similar pairs usually share common radicals and are just different in some regions, we try to detect those regions which are critical for discriminating two similar characters. Take the similar pair of characters “ت” and “ث” as example in Fig. 3. “ت ” and “ث” have the same bottom radical but are different in the top. Hence, we can easily distinguish “ت” from “ث” only by its top radical “ ˙”or recognize “ث” from “ت” by the top radical “ ˆ”. This motivates us to distinguish similar pairs by appropriately locating and exploiting the critical region information. The algorithm steeps can be conclusion as:

Algorithm 1; image steps

1 = Input image

2 = Normalization of image

3 = Extraction of the future of image

Reduction the image dimension:

A. = If the pair is similar, extraction critical region feature into two class, go to 5

B. = If the pair is not similar, go to 5

5. = End

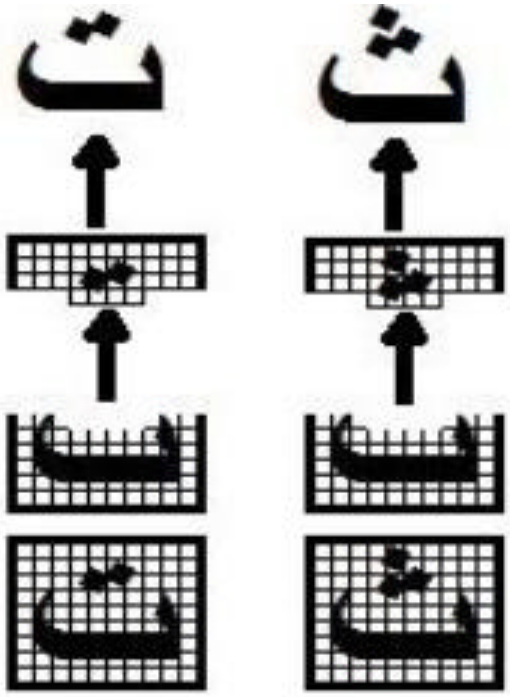

Fig. 3: Common radicals and different regions

RESULTS AND DISCUSSION

The data set that is being used in order to test and measure the proposed system performance of this study are consists of that 70 different separated characters. The data set were recorded by 7 different persons.

CONCLUSION

From the study of literature survey and proposed method, we conclude that proposed method gives considerable and expected accuracy than previous character recognitions techniques like HMM, ML, NBP, etc. Experiment results shows that, proposed method achieved an accuracy nearer to 98% provided no. of training samples per standard Arabic images should be maximum as possible as.

RECOMMENDATIONS

In the process of recognizing handwritten character, human brains may fails that's why to keep an expectations to achieve 100% accuracy is not expectable. A future resaerch is needed to correctly analyze segments patterns and fuzzy rules mentioned to achieve better accuracy which should be independent of no of training set images.

ACKNOWLEDGEMENT

This research is supported by the Scientific Research Deanship at Najran University, Kingdom of Saudi Arabia (KSA), research project number ESCI/13/31.

REFERENCES

Ahmed, M.Z. and M.S. Zakaria, 1996. Challenges in recognizing Arabic characters. International Islamic University Malaysia, Malaysia.

Alma'adeed, S., C. Higgens and D. Elliman, 2002. Recognition of off-line handwritten Arabic words using hidden Markov model approach. Proceedings of the 16th International Conference on Pattern Recognition Vol. 3, August 11-15, 2002, IEEE, Quebec City, Quebec, Canada, ISBN:0-7695-1695-X, pp: 481-484.

Dabra, S., S. Agrawal and R.K. Challa, 2011. A novel feature set for recognition of similar shaped handwritten Hindi characters using machine learning. Proceedings of the 1st International Conference on Computer Science, Engineering and Applications (CCSEA) Vol. 2, July 15-17, 2011, CS&IT-CSCP, Chennai, India, pp: 25-35.

Fathy, S.K. and M.M. Syiam, 1996. A parallel design and implementation for backpropagation neural network using MIMD architecture. Proceedings of the 8th International Conference on Industrial Applications in Power Systems, Computer Science and Telecommunications (MELECON'96) Vol. 3, May 16, 1996, IEEE, Bari, Italy, pp: 1472-1475.

Jane, A.P. and M.A. Pund, 2012. Recognition of similar shaped handwritten Marathi characters using artificial neural network. Intl. J. Eng. Res. Appl., 3: 63-67.

Kahraman, F., A. Capar, A. Ayvaci, H. Demirel and M. Gokmen, 2004. Comparison of SVM and ANN performance for handwritten character classification. Proceedings of the IEEE 12th International Conference on Signal Processing and Communications Applications, April 30-30, 2004, IEEE, Kusadasi, Turkey, Turkey, ISBN:0-7803-8318-4, pp: 615-618.

Kavallieratou, E., N. Fakotakis and G. Kokkinakis, 2002. Handwritten character recognition based on structural characteristics. Proceedings of the 16th International Conference on Pattern Recognition (ICPR'02) Vol. 3, August 11-15, 2002, IEEE, Computer Society, Washington, DC, USA., ISBN:0-7695-1695-X, pp: 139-142.

Liu, C.L., H. Sako and H. Fujisawa, 2002. Performance evaluation of pattern classifiers for handwritten character recognition. Intl. J. Doc. Anal. Recognit., 4: 191-204.

Lorigo, L.M. and V. Govindaraju, 2006. Offline Arabic handwriting recognition: A survey. IEEE Trans. Pattern Anal. Mach. Intell., 28: 712-724.

Mowlaei, A., K. Faez and A.T. Haghighat, 2002. Feature extraction with wavelet transform for recognition of isolated handwritten farsi/arabic characters and numerals. Proceedings of the 14th International Conference on Digital Signal Processing, July 1-3, 2002, USA., pp: 923-926.

Mozaffari, S., K. Faez and H.R. Kanan, 2004. Feature comparison between fractal codes and wavelet transform in handwritten alphanumeric recognition using SVM classifier. Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004) Vol. 2, August 26-26, 2004, IEEE, Cambridge, UK., ISBN:0-7695-2128-2, pp: 331-334.

Mozaffari, S., K. Faez and M. Ziaratban, 2005. Structural decomposition and statistical description of Farsi/Arabic handwritten numeric characters. Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR'05), August 31-Setempber 1, 2005, IEEE, Seoul, South Korea, ISBN:0-7695-2420-6, pp: 237-241.

Perez-Cortes, J.C., R. Llobet and J. Arlandis, 2000. Fast and accurate handwritten character recognition using approximate nearest neighbours search on large databases. Proceedings of the Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), August 30-September 1, 2000, Springer, Berlin, Germany, ISBN:978-3-540-67946-2, pp: 767-776.

Prasad, J.R., U.V. Kulkarni and R.S. Prasad, 2009. Offline handwritten character recognition of Gujrati script using pattern matching. Proceedings of the 3rd International Conference on Anti-counterfeiting, Security and Identification in Communication (ASID 2009), August 20-22, 2009, IEEE, Hong Kong, China, ISBN:978-1-4244-3883-9, pp: 611-615.

Sutha, J. and N. Ramaraj, 2007. Neural network based offline Tamil handwritten character recognition system. Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007) Vol. 2, December 13-15, 2007, IEEE, Sivakasi, Tamil Nadu, India, ISBN:0-7695-3050-8, pp: 446-450.

Vamvakas, G., B. Gatos, S. Petridis and N. Stamatopoulos, 2007. An efficient feature extraction and dimensionality reduction scheme for isolated Greek handwritten character recognition. Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007) Vol. 2, September 23-26, 2007, IEEE, Parana, Brazil, ISBN:978-0-7695-2822-9, pp: 1073-1077.

Verma, B.K., 1995. Handwritten Hindi character recognition using multilayer Perceptron and radial basis function neural networks. Proceedings of the IEEE International Conference on Neural Networks Vol. 4, November 27-December 1, 1995, IEEE, Perth, Australia, ISBN:0-7803-2768-3, pp: 2111-2115.

Wakabayashi, T., U. Pal, F. Kimura and Y. Miyake, 2009. F-ratio based weighted feature extraction for similar shape character recognition. Proceedings of the10th International Conference on Document Analysis and Recognition (ICDAR'09), July 26-29, 2009, IEEE, Barcelona, Spain, ISBN:978-1-4244-4500-4, pp: 196-200.

Yang, F., X.D. Tian, X. Zhang and X.B. Jia, 2010. An improved method for similar handwritten Chinese character recognition. Proceedings of the 2010 3rd International Symposium on Intelligent Information Technology and Security Informatics (IITSI), April 2-4, 2010, IEEE, Jinggangshan, China, ISBN:978-1-4244-6730-3, pp: 419-422.

Zhang, B. and S.N. Srihari, 2002. A fast algorithm for finding K-nearest neighbors with non-metric dissimilarity. Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognitio, August 6-8, 2002, IEEE, Ontario, Canada, ISBN:0-7695-1692-0, pp: 1-13.