WJCMS

Wasit Journal of Computer and Mathematics Science

Journal Homepage: https://wjcm.uowasit.edu.iq/index.php/WJCM

e-ISSN: 2788-5879 p-ISSN: 2788-5879

A Novel Arabic Words Recognition System Using Hyperplane Classifier

Dr.Mahmoud Abdegadir Khalifa[1,*], Dr.Ammar Mohammed Ali[2], Dr.Saif Ali Abd Alradha Alsaidi[3], Dr.liying Zheng[4], Dr.Nahla Fadel Alwan[2] and Dr.Gadiaa Saeed Mahdi[2]

[1]College of Computing and Information Technology University of Bisha, Bisha, Saudi Arabia
[2]Chemical Engineering Department, University of Technology, Iraq
[3]College of Education For Pure Science , Wasit University, Iraq
[4]Harbin Engineering University, China

*Corresponding Author: Dr.Mahmoud Abdegadir Khalifa

DOI: https://doi.org/10.31185/wjcm.Vol1.Iss2. 45
Received: February 2022; Accepted: April 2022; Available online: June 2022

ABSTRACT: Topic of exhaustive study for about past decades has been carried out in machine imitation of human reading. a small number of investigates have been accepted on the detection of cursive font writing like Arabic texts for its individual challenge and difficulty .In this work, a novel technique for automatic Arabic font recognition is proposed to demonstrate an suitable recognition rate for multi fonts styles and multi sizes of Arabic word images.The scheme can be classified into a number of steps. First, segmenting Arabic line into words depending on the vertical projection and dynamic threshold then we implicated each Arabic word as a class by ignoring segmenting the word into characters .Second ,normalizing step, the size of Arabic word images varies from each other .The system converts the images that contribution into a new size that is divisible by "N" without remainder, to decrease the difficulty of feature extraction and recognition of the system that may allow images from different resources, Third, feature extraction step which is based on apply the ratio of vertical sliding strips as a features. Finally, multi class support vector machine (one versus one technique)is used as a classifier .This method was estimated on off line printed fonts, five Arabic fonts, (Andalus, Arial, Simplified Arabic, Tahoma and Traditional Arabic) were used and the average recognition rate of all fonts was 95.744%.

Keywords: Support vector machine, one-against-one SVMs technique, vertical projection profile, horizontal projection profile

1. INTRODUCTION

Arabic text recognition is an important and demanding task not only for those who speak Arabic but also for non-Arabic national such as Persian and Urdu that use Arabic characters in his language. The most important characteristics of Arabic language can be summarized (Fig.1).

a number of associated work focus on both machine-print and handwriting, with much more discussion of machine-print (B. Al-Badr and S. Mahmoud, [1]; M.S. Khorsheed, [2]) and hand written (Liana M. Lorigo, And Venu Govindaraju [3]), many methods had been created to establish good and useful Arabic OCR systems. Some of these schemes can be briefly established as follows.

M.S, Khorsheed, [4] developed a system was based on HMM Toolkit to recognize multi-font Arabic text. (Mehmmood Abdulla Abd, [5]) introduced system to recognize printed Arabic character recognition and utilizing support vectors

FIGURE 1. Characteristics of Arabic text. (1) Direction of Arabic writing from "right" to "left". (2) Some characters are not connectable from the left side with the succeeding character. (3) A word consisting of six characters. (4) A word consisting of two separate characters. (5) A word consisting of the same number of characters but the word have different size (6) The same character with different shapes depends on its position in the word.(7) Different characters with different sizes. (8) Different characters with a same numbers of dots but different position upper and lower the base line. (9) Different characters withwith a different number of dots. (10) Basie line.

machine SVMs using one-against-all technique in the classification phase. (Menasri et al., [6]) extracted seventy-four baseline-dependant feature vectors from the graphemes and hybrid HMM/NN to recognize handwritten Arabic words.

Husni A. et al., [7] described system for recognition printed Arabic text by applying hierarchical sliding window.

Jakob Sternbya, [8] explored the application of a template matching scheme to the recognition of Arabic script. (Sami Ben Moussa et al., [9]) proposed method to recognize Arabic font, using global texture analysis based on fractal geometry. (Morteza Zahedi, [10]) has proposed a new method for Farsi/Arabic automatic font recognition. (khalifa and yang bing Ru, [11]) have computed the Euclidean distance between pairs of objects in n-by-m data matrix X based on the point 's operator of extrema and classify printed and handwritten Arabic words using one against one class SVM.

The intention of this paper aimed to examination problem of automatic recognition Arabic script. Scheming and applying a recognition method of printed Arabic text to answer these issues.

2. ARABIC TEXT SEGMENTATION

Most of the Eliminating unnecessary area: generally there are various useless regions which have no contribution to the recognition. Here, in this section talks about the eliminating unnecessary margin empty region by clipping the word image to get the regains which just consist of the text, line and word.

The image transformed into binary format (0,1) throughout white text with black background, then the row and column scan statistical analysis was done to count the numbers of white pixels showed in horizontal and vertical directions, in that order. To remove the top, bottom, left and right boundary in the same way. Here we use the actual boundaries of the word to extract the crop image (Fig. 2).

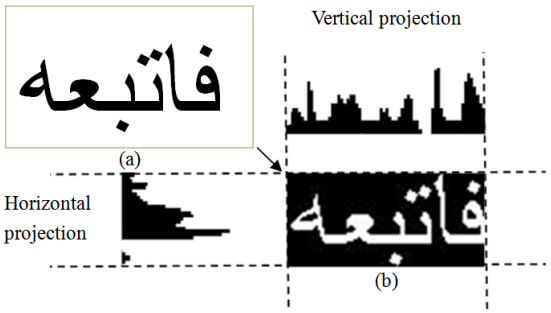


FIGURE 2. Crop Arabic word (a) to get new image (b) by removing unwanted region using vertical and horizontal projection profile

Word Segmentation: Because 1)Arabic characters in a word/sub-word are associated at one side (right side) and 2)Arabic words might be consisted of one or more sub-words, there is a distance between two sub-words, in addition to between two words. on the other hand, the space between two words is typically larger than the space between two sub-words.

Thus, segmenting line into words depends on the technique of vertical projection shape that computes the number of the black pixels using the Eq. of the vector Vproj given by:

$$vproj(i] = \sum_{i=0}^{m-1} I[i, j]$$

where I is a given image text, m is number of rows in a given text.

If the summation of Vprojis equal to zero, that means the gap has started, and we need to calculate the distance of each gap to find the value of the dynamic threshold (T) depend- ing on the mean equation. If the number of zeros is > T that means the line does not seg- ment, or else segments it. The example of product of this method shows in (Fig. 3).

FIGURE 3. Example of applying Vertical Projection to segment Arabic line to the words

3. FEATURE EXTRACTION

Printed word has a geometrical shape and this shape is regular whatever the size of the word is modify, in this case, we employed this feature to extract a set of statistical features to get a unique representation to the printed Arabic words images in different sizes. The direction of the Arabic word from right to left is believed as the feature extraction axis. As shown in these steps.

Step1: Pre-processing. A word image is binarazed into white with black background to get a binarazation image and then clip the word image by remove unnecessary area.

Step2: Size of Arabic word images are differ from each other. In order to decrease the difficulty of feature extraction and recognition procedure of this system by accepting image file from different sources, the system implements the action of normalization through transforms different sized image file into a new size by making the width of the image file is divisible by "N" without remainder, where N in this research = 72 then the size of Arabic word image is (LW x WW), where LW is the length of the word image and the WW is the width of the word image.

Step 3: Divided the new image into 72 vertical strips. The length of each slide strip window is the same as the word length(LW) and the width of the slide strip window is (WW/72).

Step 4: Then extract feature from each strip by using the summation of all pixels in each strip as element in one vector named B.

Depend on technique of vertical projection profile, by calculating the number of the white pixels in each strips using the equation of the vector Vproj as shown in this equation.

$$vproj(i] = \sum_{i=0}^{m-1} I[i, j]$$

Where m= number of the rows in the strip

Step 5: From vector B we generate a new vector named C that has 36 elements by compute the ratio between the contiguous elements of vector B after added 2 to the numerator and 1 to the denominator to avoid divide by zero as shown in these equations:

$$C(1) = ((B(1) + 2)/(B(2) + 1)) * 10, C(2) = ((B(3) + 2)/(B(4) + 1)) * 10$$
$$C(3) = ((B(5) + 2)/(B(6) + 1)) * 10 \ldots C(36) = ((B(71) + 2)/(B(72) + 1)) * 10$$

4. SUPPORT VECTOR MACHINE

The SVMs consider as a type of hyperplane classifier, it developed based on the statistical learning theory of (V. Vapnik., [12]) , This classifier associated to the error bound of generalization, since it focus on maximizing a geometric margin of hyperplane. The study of SVMs has reported from the mid-1990s,

It wins popularity due to many good features and definitely performance in the fields of nonlinear and high dimensional pattern recognition, but the application of SVMs related to pattern recognition has still continues.SVM classifier in general is a binary (two-class) linear classifier and use kernel function to represents the inner product of two vectors in linear/nonlinear feature space. And it is not straightforward to turn them into multi-class (N-class) recognition systems. There have been many references for describing the details of SVMs, like N. (Cristianini and J. Shawe-Taylor, [13]; B. Sch¨olkopf and A. J. Smola, [14]; Christopher J.C., [15]).

The implementation of SVM in the proposed system: The features of off-line Arabic word has been extracted from a given word images as discussed previously. The feature vector has feed as a row in one matrix to create one model for all words, that have all feature vectors together and each one of these vectors have been labeled to distinguish one class from the others . In general, classification task usually involves training and testing sets (Fig.4) which consist of data (class labels and a feature vectors).The task of recognition allocates each word (class) within predefined matrix that has all feature vectors. In this work an optimized support vector machines (SVMs) to classify the input Arabic words by applying one-against-one (1-v-1) SVMs technique.

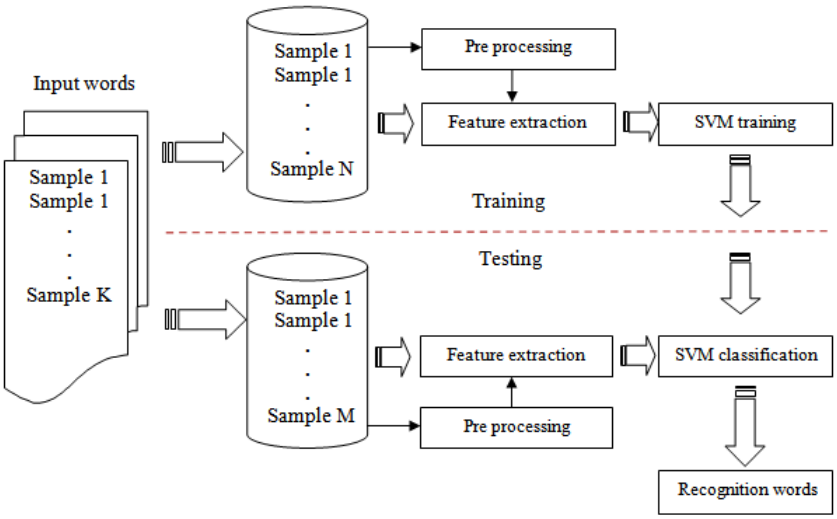


FIGURE 4. The main architecture of implementation SVM in the proposed systems

5. EXPERIMENT RESULTS AND ANALYSIS

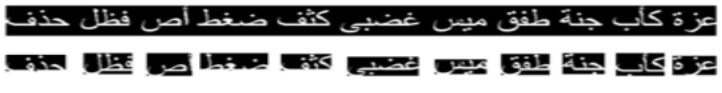
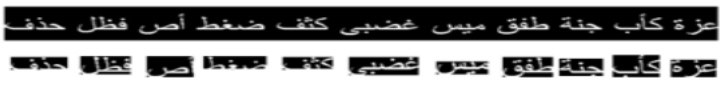
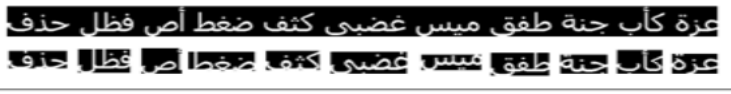
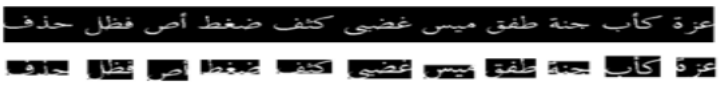
Experimental datasets: The proposed system has been tested on the printed Arabic text using public dataset (PATS-A01) . We prepare five experimental datasets for analysis based on this database. This dataset named printed Arabic Text Set A01 (PATS-A01) created by(Husni A. et al.,2008). It consists 2766 text line images the line images are available in eight fonts: (Arial, Tahoma, Akhbar, Thuluth, Naskh, Simplified Arabic, Andalus, and Traditional Arabic).

Results and analysis of word segmentation: This section evaluates the methods for segmenting lines into words as previously described. The lines that are used in this experiment have been selected randomly from PATS-A01 dataset, each line is typed in five Arabic fonts(Andalus, Arial, Simplified Arabic, Tahoma and Traditional Arabic). It works well on all fonts with a segmentation rate ofabout 100%. (Table 1) shows segmentation rates and (Table 2) shows some examples of successful segmentation results.

Table 1. Segmentation rates for each fonts

Segmentation rate for five fonts from PATS-A01 dataset	Font name
100 %	**Andalus**
100 %	**Arial**
100 %	**Simplified Arabic**
98.485 %	**Tahoma**
96.97 %	**Traditional Arabic**
99.091 %	**Average**

Table 2. Shows some examples of successful segmentation results

Font name	Arabic line samples	Success/ Failure
Andalus		yes
Arial		yes
Simplified		yes
Tahoma		yes
Traditional		yes

Results and analysis of word recognition: The proposed recognition method has been evaluated on printed Arabic words, five different Arabic fonts were used (Andalus, Arial, Simplified Arabic, Tahoma and Traditional Arabic).

6. TEST THE PERFORMANCE OF ALL FEATURES ON EACH FONT SEPARATELY WITH DIFFERENT SIZE:

This experiment has been applied on (dataset 1, dataset 2 , dataset 3, dataset 4, dataset 5) respectively, Totally each one of the dataset has 396 Arabic words and these words are divided into 66 classes of Arabic words. And the samples in each dataset are typed in one of five different fonts (Andalus, Arial, Simplified Arabic, Tahoma and Traditional Arabic) correspondingly. Therefore to distinguish between these words, the size of each Arabic word has been changed. We will simply resize the model size in each class according to original size as follows:

Model1 = original image size

Model2 = 1.2(original image size)*

Model3 = 1.4(original image size)*

Model4 = 1.6(original image size)*

Table 3. The recognition rate (%) of each fonts

Approximate unrecognized samples	Approximate recognized samples	Recognition rate %	The font name that used in the dataset	Dataset name
17	**379**	**95.799**	**Andalus**	**Dataset1**
16	**380**	**95.845**	**Arial**	**Dataset 2**
15	**381**	**96.183**	**Simplified Arabic**	**Dataset 3**
18	**378**	**95.356**	**Tahoma**	**Dataset 4**
18	**378**	**95.539**	**Traditional Arabic**	**Dataset 5**
17	**379**	**95.7444**	**The Average**	

Model5 = 1.8*(original image size)

Model6 = 2*(original image size)

When the words carried to the system, 36 features extract from each word. It was able to be recognized from each dataset(379 ,380 , 381 ,378 ,378 and 379) approximate models and produced recognition rate (95.799%,95.845%,96.183%,95.356% and 95.539%) correspondingly. See (Table 3).

7. CONCLUSION

Due to specific characteristics of Arabic scripts, such as cursiveness, recognition of Arabic words is considerable more complex than the recognition of English words.

We have proposed a novel scheme for automatic Arabic font recognition which is based on concern vertical sliding strips summation of the words images, then the ratio between two adjacent slides is considered as a feature, this feature feed to the words models which consist of word class labels and word class features. at last, the words models fetched to the classification stage using multi class support vector machine as a classifier by using 1-v-1 technique.

When the words input to the system, 36 features extracted from each word and the average recognition rate of all fonts was 95.744%.

The results obtained are extremely hopeful and have shown that the system work well and fast on off printed Arabic words and can be employ this system in the future to handwritten Arabic words. Also can be test the system on large scale date set.

FUNDING

None

ACKNOWLEDGEMENT

None

CONFLICTS OF INTEREST

The author declares no conflict of interest.

REFERENCES

[1] B. Al-Badr and S. Mahmoud *Survey and bibliography of Arabic optical text recognition," signal proceeding*, vol. 1, pp. 14–16, 1995.
[2] M. S. Khorsheed, "Off-Line Arabic Character Recognition A Review," *Pattern Analysis and Applications*, vol. 5, pp. 31–45, 2002.
[3] L. M. Lorigo, , and V. Govindaraju, "Offline Arabic Handwriting Recognition: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, 2006.
[4] M. S. Khorsheed, "Offline recognition of omnifont Arabic text using the HMM ToolKit (HTK)," *Pattern Recognition Letters*, vol. 28, pp. 1563–1571, 2007.
[5] M. Abdulla and Abd, "Effective Arabic Character Recognition using Support Vector Machines," *Innovations and Advanced Techniques in Computer and Information Sciences and Engineering*, pp. 7–11, 2007.
[6] F. Menasri, N. Vincent, E. Augustin, and M. Cheriet, "Shape-based alphabet for off-line Arabic handwriting recognition," *9th International Conference on Document Analysis and Recognition (ICDAR)*, vol. 2, pp. 969–973, 2007.
[7] A. A. Husni, A. M. Sabri, and S. Q. Rami, "Recognition of off-line printed Arabic text using hidden Markov models," *Signal Processing*, vol. 88, pp. 2902–2912, 2008.
[8] J. Sternbya, "On-line Arabic hand writing recognition with templates," *Pattern Recognition*, vol. 42, pp. 3278–3286, 2009.
[9] A. S. B. Moussa, A. Zahour, A. M. Benabdelhafid, and Alimi, "New features using fractal multi-dimensions for generalized Arabic font recognition," *Pattern Recognition Letters*, vol. 31, pp. 361–371, 2010.
[10] M. Zahedia and S. Eslamia, "Farsi/Arabic Optical Font Recognition Using SIFT Features," *Procedia Computer Science*, vol. 3, pp. 1055–1059, 2011.
[11] M. Khalifa and Y. B. Ru, "A Novel Word Based Arabic handwritten recognition system using SVM classifier," *Advanced Research on Electronic Commerce, Web Application, and Communications in Computer and Information Science*, vol. 143, pp. 163–171, 2011.
[12] V. Vapnik, "The nature of statistical learning theory," Springer, 1995.
[13] N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines and there Kernel-Based Learning Methods, Cambridge University Press," 2000.
[14] B. S. olkopf and A. J. Smola, *Learning with Kernels*. Cambridge,: MIT Press.
[15] J. C. Christopher, "A Tutorial on Support Vector Machines for Pattern Recognition," 2005. http://aya.technion.ac.il/karniel/CMCC/SVM-tutorial.pdf.
[16] A. H. M. Alaidi, S. A. A. A. Alsaidi, and O. H. Yahya, "Plate detection and recognition of Iraqi license plate using KNN algorithm," *Journal of Education College Wasit University*, vol. 1, no. 26, pp. 449–460, 2017.
[17] I. A. Aljazaery, J. Sadiq, and R. M. Al_Airaji, "Face Patterns Analysis and recognition System based on Quantum Neural Network QNN," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 16, no. 9, pp. 2022–2022.
[18] A. H. M. Alaidi, R. M. Al_Airaji, H. T. Alrikabi, I. A. Aljazaery, and S. H. Abbood, "Dark Web Illegal Activities Crawling and Classifying Using Data Mining Techniques," *International Journal of Interactive Mobile Technologies*, vol. 16, no. 10, pp. 2022–2022.