Journal of Computer Science 6 (6): 619-624, 2010
ISSN 1549-3636

A Tool to Develop Arabic Handwriting Recognition System Using Genetic Approach

[1]Hanan Aljuaid, [1]Zulkifli Muhammad and [2]Muhammad Sarfraz
[1]Department of Computer Graphics and Multimedia, University Technology Malaysia,
Johor Bharu, Skudai 81310, Malaysia
[2]Department of Information Science, Kuwait University, Kuwait, Safat 13060

Abstract: Problem statement: Significant movement has been made in handwriting recognition technology over the last few years. Up until now, Arabic handwriting recognition systems have been limited to small and medium vocabulary applications, since most of them often rely on a database during the recognition process. The facility of dealing with large database, however, opens up many more applications. **Approach:** This study presented a complete system to recognize off-line Arabic handwriting image and Arabic handwriting and printed text database AHPD-UTM that used to implement and test the system. That system start from preprocessing and segmentation phases that deepened on thinning the image and found the V and H projection profile until recognition phase by genetic algorithm. **Results:** The genetic algorithm stand on feature extraction algorithm that defined six feature for each segment beak. The system can be recognized Arabic handwriting with 87% accuracy. The confusion and rejection rates are 8.4, those causes for several problems like characters with broken loops and character segmentation problem. **Conclusion:** Peak connection solved some of the segmentation problems and helped to provide better accuracy.

Key words: Arabic characters recognition, genetic algorithm, feature extraction, Arabic characters pattern, OCR, AOCR, off-line characters recognition

INTRODUCTION

Character Recognition (CR) mechanization occupies an intensive research region of the pattern recognition research area. CR automation means translating images of characters into an editable text, in other words, it represents an attempt to simulate the human reading process. In other said handwriting recognition is a very challenging task due to the existence of many difficulties such as the high variability of the handwritten styles and shapes, uncertainty of human writing, writing skew or slant, segmentation of the words into characters and the size of the lexicon (Amin and Kavianifar, 1997).

The problem of handwriting recognition can be classified into two main groups, off-line and on-line recognition, according to the format of handwriting inputs (Mener *et al*., 1994; Kherallah *et al*., 2008). In offline recognition, only the image of the handwriting is available, while in the on-line case temporal information such as pen tip coordinates, as a function of time, is also available. Many applications require off-line HWR capabilities such as bank processing, mail sorting, document archiving, commercial form-reading and office automation. So far, off-line HWR remains an open problem, in spite of a dramatic boost of research (Koerich *et al*., 2003; Plamondon and Srihari, 2000; Vinciarelli, 2002) in this field and the latest improvement in recognition methodologies (El-Yacoubi *et al*., 1999; Vinciarelli *et al*., 2004; Lorigo and Govindaraju, 2006).

At present, there are many different methods to recognize the Arabic character. The use of genetic algorithm to recognize a character has been a new algorithm used in this problem. Genetic algorithms offer a particularly attractive approach for this kind of problems since they are generally quite effective for rapid global search. Moreover, genetic algorithms are very effective in solving large-scale problems, but what is the Gas?

Genetic Algorithm (GAS) is a search technique used in computer science to find approximate solutions to optimization and search problems and is inspired by evolutionary biology such as inheritance, mutation, natural selection and recombination. Genetic algorithms are typically implemented as a computer simulation in which a population of abstract representations of candidate solutions to an optimization problem evolves toward better solutions. Traditionally, solutions are

Corresponding Author: Hanan Aljuaid, Department of Computer Graphics and Multimedia, University Technology Malaysia, Johor Bharu, Skudai 81310, Malaysia

represented in binary as strings of 0s and 1s, but different encodings are also possible. The evolution starts from a population of completely random individuals and happens in generations. In each generation, the fitness of the whole population is evaluated, multiple individuals are stochastically selected from the current population (based on their fitness), modified (mutated or recombined) to form a new population, which becomes current in the next iteration of the algorithm as show in Fig. 1. Evolutionary algorithms work on populations, instead of single solutions. In this way the search is performed in a parallel manner.

This study describes an extended version of an off-line handwritten Arabic word recognition system based on Feature extraction approach and Genetic Algorithms.

Most recognition systems in use today are developed for applications with a restricted lexicon of words. These systems are focused on certain applications such as the reading of cheque amounts or postal addresses, which are proven to be realistic and profitable. The development of recognition systems, however, needs a large amount of data to train and test the system. The implementation of a system requires real world data but data from the bank or the postal system are often confidential and inaccessible for non-commercial research. As the amount of data is crucial for a reliable training of recognition systems, artificial data collected on special forms instead of scarce real world data can be used. Despite the disadvantage of using artificial data, the data labeling process is made simpler due to the fact that the forms can be adapted to the automatic labeling process.

Database AHPD-UTM: Collecting data: The data collected in database called Arabic handwriting and printed text database AHPD-UTM. It is have two types of data printed text and handwriting. Printed text is an image of Arabic words written as text types.

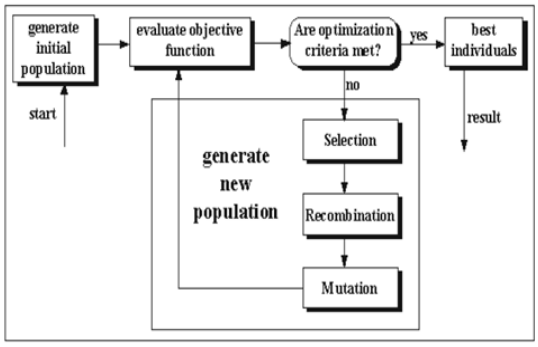

Fig. 1: Structure of a single population evolutionary algorithm

Where, handwriting is an image of Arabic words written by human.

However, Printed text contented 50 Paragraphs of different type of categories in Literatures, Sciences, Arts, religions Geographies and History collected from Arabic Wikipedia, stored as paragraphs, printed and scanned as BMP image files format.

On the other hand, handwriting images collected by filling Form filled by 100 writers. The writers were classified in to 6 groups according to their ages and gender, (Table 1). The writers are aged from 5 years to about 60 years. Their occupations are undergraduate students, administration staff, faculty members, housewives and schoolboys. The writers who belonged to group 1 are from King Abdul Aziz School boys and girls in Alkhobar in Saudi Arabia. While the faculty members and the undergraduate students were from King Saud University in Riyadh in Saudi Arabia. Filling the Form by writers in different age and gender groups given the words an authentic readability level. Accordingly, the readability level started from a low level in group 1 and went up in the highest groups.

The content of the filling form has about 150 words from Al-Quran, the holy book of Muslims and some other general words in Arabic language. The collected forms are depended to represent each character shape one time at least. The Form has 5 studies the first page is for personal information. The second and third study contain a Table of 5 columns by 30 row sorted alphabetic (from Alf (أ)-Ya'a (ي)) each row has the different shapes of the characters which had the row. The last row present the special case of the Zigzag shapes (Hamzah). The last two study has empty Tables like that in study two and three are filling by the writers in the. It was taken in to consideration that the writing of the words must be unconstrained and the separating of the scanned words must be easy and not time consuming.

Analyzing the data: The data of AHPD-UTM is an image; it might be in one of the following:

- An image containing one and only one sub word
- An image containing a word of more than one sub word
- A document of one or multi text lines like the paragraph

Table 1: The writers' groups

Group	Writers' gender	Writers' age
1	Male	Between 5-15 years
2	Female	Between 5-15 years
3	Male	Between 15-30 years
4	Female	Between 15-30 years
5	Male	Above 30 years
6	Female	Above 30 years


في شهر واحد إستضاف "هلنان شيرد" إثنين من المؤتمرات الدولية الهامة، التى حضرها عدد من الوزراء والمعنيين فى مصر والخارج ..ففى الفترة من ١٣ إلى ١٥ يناير الماضى عقد الإتحاد الدولى لتنظيم الأسرة - إقليم العالم العربى - مؤتمرة الذى ناقش فيه وسائل تفعيل توصيات المؤتمر الدولى للسكان والتنمية الذى إنعقد بالقاهرة خلال سبتمبر ٩٣ . هذا وقد إشترك وزير الصحة ، ود. ماهر مهران وزير السكان السابق.
كما إستضاف الفندق المؤتمر العلمى الثالث لجمعية جراحة القلب والصدر المصرية لمناقشة العيوب الخلقية المركبة للاطفال . كما ناقش المؤتمر إستعمال الصمامات الآدمية لإستبدال الصمامات التالفة، وكذلك الطرق المثلى لعلاج إصابات حوادث القلب والصدر. هذا وقد حضر المؤتمر وفود من جراحى إيطاليا وإنجلترا والولايات المتحدة، فضلاً عن فريق من الجراحين المصريين والعرب .إشترك أيضاً فى جلسات المؤتمر د. حسين بهاء الدين وزير التعليم، ود. إسماعيل وزير الصحة ، ورأس المؤتمر د. أنور بليغ،

Fig. 2: Printed text paragraph

Table 2: Number of words, sub-words and letters in one form

Form	Word	Sub-word	Character
	150	270	528

Table 3: Number of writer's, words, sub-words and letters in each group

	Writers		Database		
Categories	M	F	W	SW	L
5-15	12	16	4200	7560	14784
15-30	13	27	6000	10800	21120
Above 30	15	17	4800	8640	16896
Total	40	60	15000	27000	52800

The image that has a paragraph is printed text type, but the handwriting image has one word of one sub word or more. Accordingly, there are two samples are analyzed in our data: Analyzing documents that has been collected from Arabic Wikipedia and analyzing handwritten words that has been collected from our filling forms.

In the first sample, paragraphs from six to ten lines were selected randomly from Arabic Wikipedia. Then, each paragraph was coordinated with Arial font type and different text size around from 12-16. After that, it was printed and scanned as an image of BMP format, an example in Fig. 2. Special tagging scheme was used to facilities the access to the BMPs. The scheme to tagging the files of the paragraph images was:

Example: P15

Letter represents the paragraph "P"	The serial number of the paragraph

In the other sample, the forms were scanned as 1-pixel color BMP image files for study 4 and 5 (handwritten study). The BMP format was selected because it is easily manipulated. When the scanning was completed, a process was stated to separate each word. The scheme to tagging the files of the words images was:

Example: 2F55-54

The number of the group	Letter to represent The Gender F, M	The serial number of the sample plus "-"	The serial number of the word

Storing resulted samples into database: The image resulted from data analysis storing into database. The database has two main directories. One has 50 BMP images of printed text paragraph and the other has six sub directories, each one has been saved the words for each groups.

Thus, the resulted database has 15000 handwritten words written by 100 writers as shown in Table 2 and 3.

As mentioned, the database has samples of different writers from different ages and genders which give the word an authentic readability level as shown in Table 3.

MATERIALS AND METHODS

Recognition algorithm: To recognize Arabic characters the preprocessing and the segmentation stages must be done before the feature extraction and the recognition stages. In this study the preprocessing stage depended upon the thinning the image of Arabic word then find the vertical and horizontal projection profile.

The horizontal projection is defined as:

$$h(i) = \Sigma p(i, j))$$

And the vertical projection as:

$$v(j) = \sum p(i, j)$$

Where:

I = the row number
J = the column number
P = the pixel value. It is 0 for white pixel (or background), or 1 for black pixel (or for ground)

The segmentation of the word to characters depends in the preprocessing stage.

However, the Arabic language has 28 characters but there is more than one shape for each character depending on the position of it in the word. The different shapes of the characters were collected to attaché more than 200 shapes of the Arabic characters. The feature extraction of each shape described by the

feature extraction algorithm that to define one chromosome of the genetic algorithm chromosomes to recognize the character shape that his feature described in that chromosome.

Feature extraction: After the segmentation of a word has been done, the feature of each character must be detected to recognize the shape of the character. The recognition algorithm stands in six features of each characters shape that are the length of the character, the width of it, if it has a loop or not, if there is a right character to connect to it, if there is a left character to connect to and if there is a complement of the character like the zigzag shape (Hamza), one point, two point or three point.

The Arabic characters are written in a cursive way and all the characters in the word are stand in the baseline. The feature extraction algorithm depend in this fact, so in this algorithm the Arabic characters image divided to three areas the first one is the baseline area that was detected by the horizontal projection profile where The longest spike represents the baseline as show in Fig. 3. The second area is the area upper the baseline this area usually content the upper complementary and the upper length of the character like ك, ـلـ. The third area is the area below the base line that usually content the down complementary the character and the down length of the character like و، ر،ج (Fig. 3).

The vertical and horizontal projection profile calculated for each area individually, to help in detected the feature of the character. Each character shape has six features that are unique for it only no other shape has that features compatibility.

After defined the area of image and it is projection profile, the feature extraction algorithm start by tracing the boundary of the image and traveling around the eight neighborhoods from right to left. In one vector array has six cells that's call feature vector where each cell represent one feature of the characters shape by integers value as show in Table 4 that describe the feature vector of character shape ـه that is short, has small width, has loop in the baseline, does not has right connection, has left connection and does not have complementary. This vector is unique for the character shape ـه.

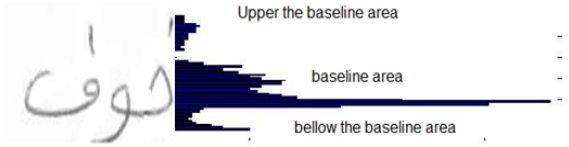

Fig. 3: Detect the baseline by H projection profile

Each feature detected according to the following algorithm:

1. Length: It is denoted by 0 if the character shape is short or 1 if the character shape is one. The length of the character shape detected by the baseline area of it is in the baseline area only it is short otherwise it is high
2. Width: It is denoted by 0 if the character shape has small width and 1 if the character shape has large width. The width of the character shape depends in the vertical projection profile of the image and the thinning algorithm where all the images have one pixel width
3. Loop: It is denoted by 0 if the character shape has no loop, 1 if the character shape has one loop in the baseline area above the baseline like ـعـ, -1 if the character shape has one loop bellow the baseline ـهـ, 2 if the character shape has two loop above the baseline like هـ, 3 of the character shape has open loop in the left said and it is stand in the baseline like ـحـ or -3 if the character shape has open loop in the right said and it is stand in the baseline like ـعـ. The type of loop detected according to the following algorithm:
   a. If the subtraction of the entire nearest pixels row in the same column is >1and the subtraction of the entire nearest pixels column in the same row is >1. In this case the loop type is 1 if the last row is the baseline otherwise it is -1
   b. If the subtraction of the entire nearest pixels row in the same column is >1 except the first column which has one pixel only and the second one which has two pixel behind other the loop type 3 otherwise the loop type -3
   c. If there is more than two pixels in the same column the subtraction of it >1 and the subtraction of all the rows in the same column and all columns in the same row >1. The loop type is 2
4. Right connection: denoted by 0 if there is no character in the right said otherwise it is 1
   It is detected by the vertical projection profile of the baseline area if there are ones in the right said of the character width, it denoted by 1 otherwise 0

Table 4: Feature vector that represent character ـه

Complementary	Connection				
Character	Left	Right	Loop	Width	Length
0	1	0	-1	0	0

5. Left connection: denoted by 0 if there is no character in the left said otherwise it is 1. It is detected by the vertical projection profile of the baseline area if there are ones in the left said of the character width, it is denoted by 1 otherwise 0
6. Complementary character: It is denoted by 0 if there are no complementary, 1 if there is one dote above the baseline area, -1 if there is one note below the baseline area, 2 if there are two dots above the baseline area, -2 if there are two dots below the baseline area, 3 if there are three dots above the baseline area and 4 if there is a zigzag shape (Hamza) above the baseline area. It is detected by using the vertical projection profiles of the areas

Genetic algorithm:

Coding the population: Each chromosome is a structure has the Arabic character Unicode and 6 double numbers that describes the feature of it.

The fitness function: The fitness of a chromosome represents the degree of match between the feature of this chromosome and the feature extracted from the feature extraction function.

The selection-reproduction operator: This operator reproduces a competition between different chromosomes. Geometrically, this operator makes sure that the best chromosome will not be lost while the worse chromosomes are discarded. In order to introduce new chromosome in the population, two other special operators are used: The crossover operator and the mutation operator.

The crossover operator: Chromosomes are mated at random with a higher probability for the chromosomes that undergoes crossing. A sub chromosome is selected in each parent and the resulting offspring is build by concatenation of the sub chromosome.

The mutation operator: In order to extend the explored solution space and to avoid the recombination of the same Chromosome, the mutation operator is used to alternate some chromosomes and with combination of the crossover operator, this will create new and possibly better solutions as in the Fig. 4.

Fig. 4: Example of mutation operator

System implementation: The implementation of the system starting from the preprocessing stage where the image thinned and the vertical and horizontal projection profile found. Then the image entered to the segmentation stage which segment the image to small peaks depending in the vertical projection of the thinning image. After that, the feature vector of each peak is found by the feature extraction algorithm. Next each feature vector sends to genetic algorithm to search for the chromosome that has the best fitness according to that the character are recognize in the other part of the chromosome which have the Unicode of it. If there no chromosome fitness to the feature vector, the feature vector returns to the feature extraction algorithm to concatenate the peak of it with the nearest neighbor peak and detect the feature vector of it.

RESULTS AND DISCUSSION

Experimental and result: In order to check the accuracy of the Arabic handwriting recognition system using genetic algorithm, handwriting samples in AHPD-UTM has been used. The collected characters were 52800 characters.

In the training phase, 11450 characters were used to perform the algorithm and 4000 characters was ignored for the poor of writing and the difficult to be recognized it be human's reader 37350.

The system was tested by 37350 characters. These experiments show that the system can recognize Arabic handwriting with 87% accuracy. The confusion and rejection rates are 8.4, those causes for several problems like characters with broken loops and character segmentation problem. Peak connection solves some of the segmentation problems and help to provide better accuracy.

The recognition problem solved by genetic algorithm which may yield a different solution for the same word in each time that depends in the number of population and iteration. But, there is a big similarity between the original solution and the proposed one. In instance, we considered the population size effect. The final result is composed of one character in which the fitness function has a correct value. We remark that the tested letters are better recognized if the population size is less or equal to 100. However, the executed time increases if the population size is about 100 as seen in the last column of Table 5.

In order to test the influence of the population on the recognition rate, we redid the execution with population of varied size (10, 20, 30, 40, 50, 60, 70, 80, 90, 100, and 1000). The recognition rate reaches with 10 population size as show in Fig. 5 if the population size decreased then the recognition rate increase.

Table5: Example of the population result and the execution time

Tested character	Proposed character	Fitness value	Population size	Generation number	Execution time (sec)
خ	خ	6	10	10	3
	ت	5	100	12	6
	خ	6			
و	و	6	10	11	7
	و	6	100	15	9
ف	ن	5	10	10	11
ف	ف	6			
	ق	5	10	20	13
	ن	5			
	ف	6			

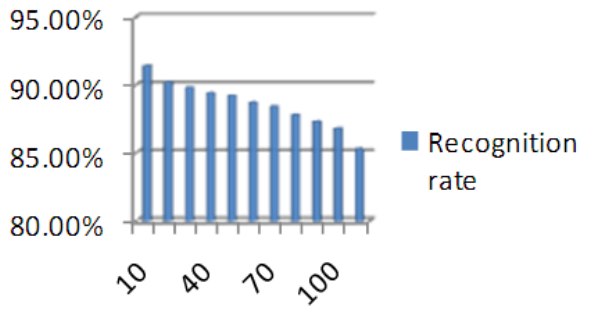

Fig. 5: Relation between population size and recognition rate

The system implemented in Mat lab 2007 language on Intel(R) Core (TM) 2Due CPU, 2.20 GHz.

CONCLUSION

This study shows that the Arabic database AHPD-UTM is possible to obtain interesting off-line Arabic handwriting recognition rate. The system was used to recognize Arabic characters in all their form. First of all, the projection and thinning method used to solve segmentation and feature extraction method. The conjunction method solve the over segmentation problem. The recognition problem solved by genetic algorithm which may yield a different solution for the same word in each time that depends in the number of population and iteration, but there are a big similarity between the original solution and the proposed. To solve that problem we keep the three first solutions in the early iteration.

This study shows also the benefits of AHPD-UTM. Where this data base cloud be used by other researchers to test their systems. It is written by 100 writers with 50 samples of printed text.

REFERENCES

Amin, A. and M. Kavianifar, 1997. Automatic recognition of printed arabic text using neural network classifier. Proceeding of the International Conference on Image Analysis and Processing, Aug. 18-20, IEEE Xplore Press, Ulm, pp: 616-623. DOI: 10.1109/ICDAR.1997.620576

El-Yacoubi, A., M. Gilloux, R. Sabourin and C. Suen, 1999. An HMM-based approach for off-line unconstrained handwritten word modeling and recognition. IEEE Trans. Patt. Anal. Mach. Intell., 21: 752-760. DOI: 10.1109/34.784288

Kherallah, M., F. Bouri and A.M. Alimi, 2008. On-line Arabic handwriting recognition system based on visual encoding and genetic algorithm. Eng. Appli. Artif. Intell., 22: 153-170. DOI: 10.1016/j.engappai.2008.05.010

Koerich, A., R. Sabourin and C. Suen, 2003. Large vocabulary off-line handwriting recognition: A survey. Patt. Anal. Appli., 6: 97-121. DOI: 10.1007/s10044-002-0169-3

Lorigo, L.M. and V. Govindaraju, 2006. Offline Arabic handwriting recognition: a survey. IEEE Trans. Patt. Anal. Mach. Intell., 28: 712-724. DOI: 10.1109/TPAMI.2006.102

Mener, G., G. Lorette and P. Gentric, 1994. A genetic algorithm for on-line cursive handwriting recognition. Proceedings of the International Conference od Pattern Recognition, IEEE Xplore Press, Jerusalem, pp: 522-525. http://direct.bl.uk/bld/PlaceOrder.do?UIN=024352213&ETOC=EN&from=searchengine

Plamondon, R. and S.N. Srihari, 2000. On-line and off-line handwriting recognition: a comprehensive survey. IEEE Trans. Patt. Anal. Mach. Intell., 22: 63-84. DOI: 10.1109/34.824821

Vinciarelli, A., 2002. A survey on off-line cursive word recognition. Patt. Recog., 35: 1433-1446. DOI: 10.1016/S0031-3203(01)00129-7

Vinciarelli, A., S. Bengio and H. Bunke, 2004. Offline recognition of unconstrained handwritten texts using HMMs and statistical language models. IEEE IEEE Trans. Patt. Anal. Mach. Intell., 26: 709-720. DOI: 10.1109/TPAMI.2004.14