Printed Arabic Words Recognition Using Genetic Algorithm

Asst.Teach.Rasha H.Ali
University of Baghdad /Computer Department - Education College for women-

Abstract:

The automatic recognition of text on scanned images has enabled many application such as searching for words in large volumes of documents, automatic sorting of postal mail, and convenient editing of previously printed documents. In this paper, a Printed Arabic word recognition system using Wavelet Transform, and Genetic Algorithms (GA) is proposed. The system consist three steps. In first step pre-processing are applied on the input image. Secondly features are extracted, which will be taken as the input to the genetic algorithm. In feature extraction stage the feature vector extracted by using 1-level linear wavelet decomposition technique and only the approximation are taken as the input to the (GA). While the third step is a classification which is carried out by GA. The proposed approaches are tested on a ten Arabic word. The recognition rate is 90%.

Keywords:- **Image Recognition, Word Recognition, Wavelet Transform, Genetic Algorithm, Arabic Recognition.**

1- Introduction

An optical character recognition system typically consists of the following processing steps: digitization, preprocessing, segmentation, feature extraction, recognition using one or more classifiers and contextual verification or post-processing.

Recognition of Arabic characters represents an important goal, not only for Arabic speaking countries, but also for Curds, Persians and Aurdo speaking Indians. However, in spite of the progress of machine character recognition techniques (both printed and handwritten) of Latin, Chinese and Japanese characters, research of Arabic characters has been slowly gaining momentum since the early 1980s. The main reason for such delay is the different characteristics of Arabic writing from other writings. This also results from the fact that techniques developed for other writings cannot be easily applied to Arabic writing [1].

Arabic writing is unlike English, it is written from right to left, and it consists of 28 characters. Characters do not contain upper and lower case and their shape depend on its position in the word (isolated, beginning, median and end) [2]. The Arabic letter may be connected from one side only or both sides, which depend on the word itself, also each word, may be composed of one unit (connected characters) or more. Some letter contains ascenders and descenders [3]. In Arabic word recognition system, printed typed document is scanned and used as input to the system. Then the document is prepared for processing: this stage is called preprocessing stage which includes noise removal, binarization and baseline estimation. The document is then segmented into lines and then lines segmented into

words, sub-words, or characters by using horizontal and vertical projections [4].

2-Genetic Algorithm

Genetic algorithms are a stochastic search algorithm, which uses probability to guide the search. It was first suggested by John Halland in the seventies. Over the last twenty years, it has been used to solve a wide range of search, optimization, and machine learning. Genetic algorithms are a class of parallel adaptive search algorithms based on the mechanics of natural selection and natural genetic system. It can find the near global optimal solution in a large solution space quickly. It has been used extensively in many application areas, such as image processing, pattern Recognition, feature selection, and machine learning [5].It is a powerful search technique that mimics natural selection and genetic operators. Its power comes from its ability to combine good pieces from different solutions and assemble them into a single super solution [6]. Genetic algorithms are initial population of solution called individuals is (randomly) generated, the solutions are evaluated. The algorithm creates new generations of population by genetic operations, such as reproduction, crossover and mutation.

The next generation consists of the possible survivors (i.e. the best individuals of the previous generation) and of the new individuals obtained from the previous population by the genetic operations. The best source of information about Gas is Holland's adaptation in natural and artificial systems; each position in the string is called a gene. The possible values of each gene are called alleles. A particular string is called a genotype. The population of strings also called the gene pool. The organism or behavior pattern specified by a genotype is called a phenotype. If the organism represented is a function with one or more inputs, these inputs are called detectors [7].

3-Wavelet Analysis

Wavelet transform is a representation of a signal in terms of set of basic functions, which is obtained by dilation and translation of a basic wavelet. Since wavelets are short time oscillatory functions having finite support length (limited duration both in time and frequency), they are localized in both time (special) and frequency domains. The joint spatial-frequency resolution obtained by wavelet transform makes it a good candidate for the extraction of details as well as approximations of images [8].

4- The Proposed System

In this section, a description for the recognition system has been demonstrated as follows.

4-1 The Data set

The system has been applied on ten Arabic words printed with font

size 16, with font type Simplified Arabic, because the nature of connected Arabic letter each letter have four pattern, in this paper we saved only two pattern for each letter. The number of patterns for characters is 66 pattern. The Arabic words are

(جامعة- بغداد- كلية- التربية- للبنات- قسم- علوم- الحاسبات- مكتبة- تعليم)

4-2 Architecture of the proposed system

The proposed system is used for recognition printed Arabic word using Wavelet in feature extraction and genetic algorithm as classification. This system contain three main stages: preprocessing stage, feature extraction stage, classification stage. Figure (1) shows the architecture of the proposed system.

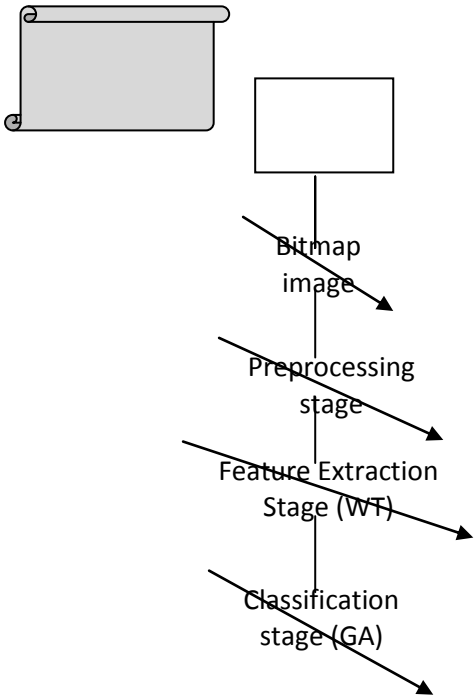

Fig. (1) The Architecture of proposed recognition system

4.3. Preprocessing stage:-

The preprocessing stage includes some steps after convert the document in to image:

- Binarization :- in this step the bitmap image convert the scanned image to binary image contain 0 and 1[9].
- Segmentation: - this step is very important in recognition system because segmenting the words into characters performed a higher recognition rates, but this segmentation suffers from segmentation problems such as over segmentation, under segmentation or misplaced segmentation which affects the recognition performance in a negative way. Therefore, in this paper we separate the paper into text lines then each line segmented into segments this segments may be contain one or more words by using the vertical projection at

last segment the word or sub word to letter using baseline, and threshold to detect the begin and the end of each letter[10].

- Fixed the size of each letter fixed frame because the connected of Arabic letter comes in different size form segmentation.

4.4. Feature Extraction:-

Feature extraction is the process of getting useful information from binarized files to be used for classification purposes in this paper, we used the discrete wavelet transform (Haar) to extract the information and produced the feature vector as input to classification using genetic algorithm. The Haar Wavelet is a simplest and the fastest wavelet transformation, which operates on data by calculating the sums and the differences of the adjacent elements.

4.5. Classification:-

In this stage the classification is perform using genetic algorithm to compare between the test word and the patterns of characters that are stored in file by the following main steps to performed this stage. Figure (2) shows the classification stage using GA.

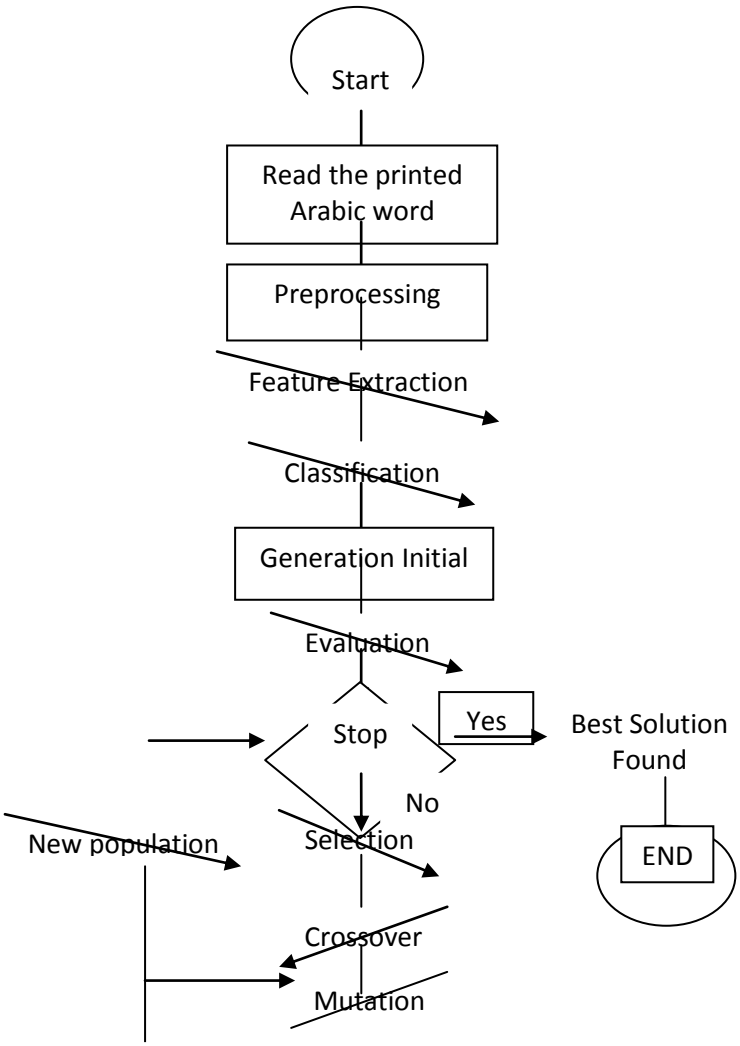

Fig. (2) Classification Using GA

a. Initial of population

The initial population contain 50 individual managed by Genetic Algorithms, the initial population was generated depend on the word wanted be to recognized.

b. Encoding

Binary encoding is used to encode the feature vector of an image for each word wanted to be recognize as shown in table(1).

Table (1) the binary code of word

The Word	The Word Binary encoding
كلية	1111010101010111111001111111010101111110111

c. Evaluation

To evaluate the population, the fitness of each individual or chromosome are be calculated, which is the differences between the characters of word to be recognized and the patterns of characters of words that are stored in database, whenever the distance value is less or near to zero then the word is recognized. It's formula is derived from the Euclidean distance using Mean Square Error (MSE) as shown in Eq. (1) [11].

$$\text{Distance (A,B)= MSE} = \frac{1}{N}\sum_{i=1}^{n}\sqrt{(A-B)^2} \quad \text{.....} \quad (1)$$

Where:-

N: is number of individual.

A: is test file or unknown word.

B: is reference file or patterns are stored in data set.

d. Selecting

After evaluating individuals of the population, the elitist selection method will be used; this method allows the Genetic Algorithm to retain a number of best individuals for the next generation. These individuals may be lost if they are not selected to reproduce.

e. Crossover

Single point crossover was used feature vector by replace the value between the characters of word to be recognized and the patterns of characters that are stored in database. With regard do not make maximum change in value of unknown word as shown in following table (2)

Table (2) Single Point Crossover

Unknown word	10101110011110110101110 **11**
Stored Pattern	11101100111010101110101 **01**
Unknown word after Crossover	10101110011110110101110 **01**
Stored Pattern after Crossover	11101100111010101110101 **11**

f. Mutation

A mutation operation in this paper is simple change because keeping the feature extraction is very important in this work, so if the word did not recognized, the mutation operation will be done by replacing the individual. The probability of mutation is 0.001.

g. The Stop Criterion

The stop criterion of the proposed approach is either the word is recognized or all populations have been covered.

5- Experimented Work

The experiments are conducted on 10 Arabic words. For example recognize the word (جامعة), this word printed with font 16 and type simplified Arabic font. At the beginning, load the word image shown in figure (3).

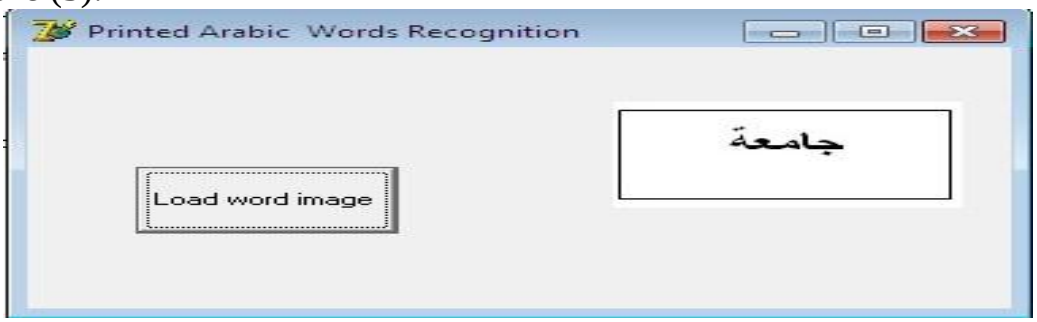

Fig. (3)The loading of word image

Then a preprocessing and feature extraction stage with the encoding of genetic algorithm has been merging in one step as shown in figure (4).

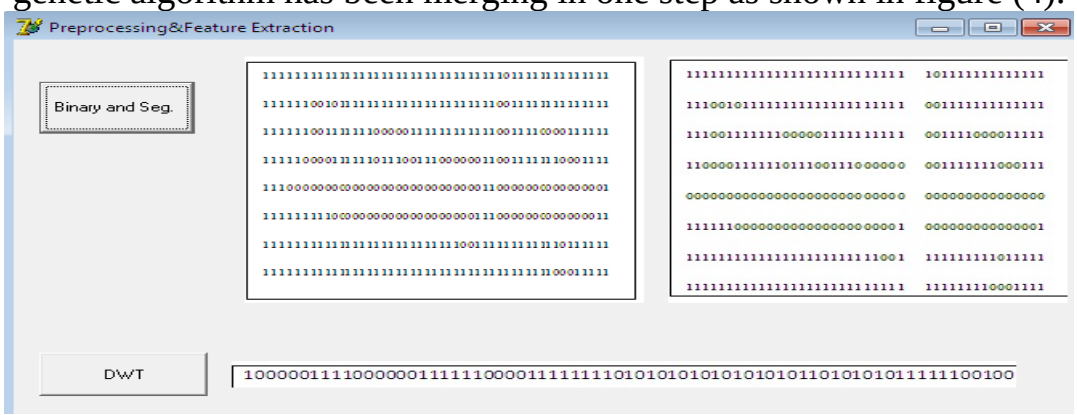

Fig. (4) The Preprocessing and Feature and encoding of GA

Finally the genetic algorithm operations have been applied on the word for recognizing. As shown in figure (5).while table (3) shows the recognition rate for each word in the data set.

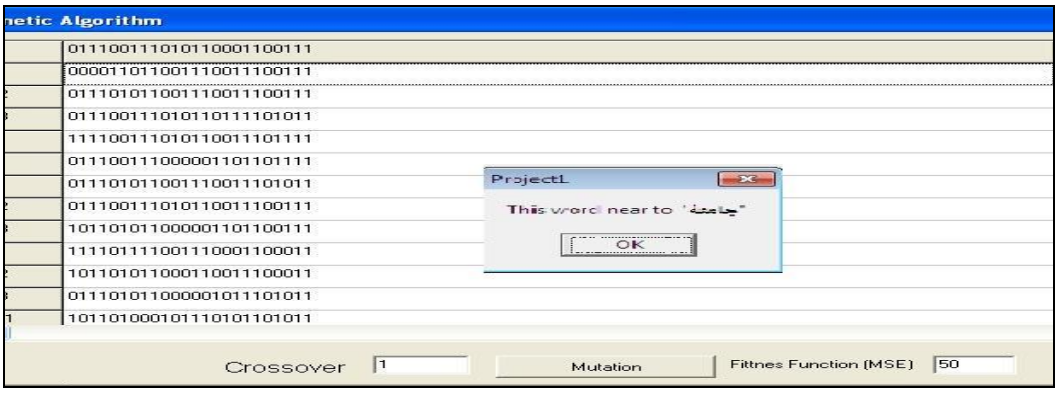

Fig. (5) Genetic Algorithm

Table (3) Recognition rate for each word

جامعة	90%
بغداد	80%
كلية	95%
التربية	85%
للبنات	100%
قسم	85%
علوم	90%
الحاسبات	80%
مكتبة	100%
تعليم	95%

6- Conclusion

This paper proposed a method for recognizing Arabic words which is based on discrete wavelet transform (DWT) with simplest type (Haar) and genetic algorithm. The proposed method is applied on ten Printed Arabic words. The proposed method included three stages: first preprocessing stage (binarization, segmentation), second feature extraction stage (DWT), third classification stage (genetic algorithm). The recognition rate is 90%. The system was implemented by Delphi programming language (version 7).

Reference

[1]- Amin, A., A. Kaced, J.P. Haton and Moher, 1980.Handwritten Arabic character recognition by I.R.A.C. system. Proc. Fifth. Intl. Conf. Pattern Recognition, pp: 721-731.

[2]-Al-Hajj Mohamad, R.; Likforman-Sulem, L.; Mokbel, C.; Combining Slanted-Frame Classifiers for Improved HMM-based Arabic Handwriting Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, No. 7, pp. 1165-1177, 2009.

[3]- Abdulaziz, E.Alsaif, K.I., "Radon Transformation for Arabic Character Recognition", International conference on Computer and Communication Engineering, Kuala Lumpur, Malaysia, 13-15 May,2008, pages 433-438, 2008.
[4]- El rube. Ibrahim, El Sonni. Mohamed, and Soha S. Saleh," Printed Arabic Sub-Word Recognition Using Moments", World Academy of Science, Engineering and Technology, Egypt, 2010.
[5]- Jasradj U. Dange, "Introduction to Genetic Algorithms" 2001.

[6]- Phili Kohn, "Combing Genetic Algorithm and Neural Networks" M.Sc. Thesis,University of Tennessee,1994.

[7]- Rasheed, Sh. A., "Genetic Algorithms Application in Pattern Recognition", Master's thesis, National Computer Center Higher Education Institute, 2000.
[8]- C.FAN, H.WANG, F.ZHANG," Improved Wavelet-based Illumination Normalization Algorithm for Face Recognition" The 1st International Conference on Information Science and Engineering (ICISE2009) IEEE.2009, p: 583-586.

[9]-J. Sauvola, M. Pietikainen, Adaptive document image binarization, Pattern Recognition 33 (2) (2000) 225–236.
[10]- Jawad H AlKhateeb, Jianmin Jiang, Jinchang Ren, and Stan S Ipson, "Component-based Segmentation of Words from Handwritten Arabic Text", World Academy of Science, Engineering and Technology, vol.41, pp.344-348, 2008.
[11]- Mathematic distance, an article from Wikipédia, the free encyclopedia. http://fr.wikipedia.org/wiki/Distance_%28math%C3%A9matiques%29.

تمييز الكلمات العربية المطبوعة باستخدام الخوارزمية الجينية

م.م. رشا حسين علي

جامعة بغداد/ كلية التربية للبنات

الملخص:

إنَّ تمييز التلقائي للنصوص المصورة باستخدام الماسح الضوئي تستعمل في تطبيقات عديدة منها البحث التلقائي للنصوص الكبيرة الحجم، الترتيب التلقائي للرسائل في البريد الالكتروني، وتحرير النصوص المطبوعة سابقا. في هذا البحث نقترح طريقة لتمييز الكلمات العربية المطبوعة باستخدام تحويل المويجة والخورازمية الجينية. النظام المقترح يتكون من ثلاث مراحل: تتم في المرحلة الاولى معالجة أولية للصورة وفي الثانية تتم عملية استخلاص الصفات وفي المرحلة الاخيرة تتم عملية التصنيف وتمييز الكلمة. تم الاعتماد على الصفات الرئيسية التي يقدمها تحويل المويجة بالاعتماد على حزمة الترددات الواطئة والتي تعتبر كمدخل للخوارزمية الجينية والتي تتولى عملية تصنيف الكلمة وتمييزها. طبق النظام على عشر كلمات عربية مطبوعة. بلغت نسبة التمييز الكلي لجميع الكلمات ٩٠%.

الكلمات المفتاحية:– تمييز الصورة، تمييز الكلمات، التحويل المويجي، الخوارزمية الجينية،التمييز العربي