Diyala Journal of Engineering Sciences

ISSN 1999-8716

Printed in Iraq

Vol. 05, No. 01, pp.79-87, June 2012

ON-LINE HANDWRITTEN ARABIC CHARACTER RECOGNITION BASED ON GENETIC ALGORITHM

Haithem Abd Al-RaheemTaha
Assistant Lecturer
Electrical Department, Engineering College, Al-Mustansirya University.
(Received:2/2/2011 ; Accepted:27/10/2011)

ABSTRACT:- On-line Arabic handwritten character recognition is one of the most challenging problems in pattern recognition field. By now, printed Arabic character recognition and on-line Arabic handwritten recognition has been gradually practical, while offline Arabic handwritten character recognition is still considered as "The hardest problem to conquer" in this field due to its own complexity. Recently, it becomes a hot topic with the release of database, which is the first text-level database and is concerned about the area of realistic Arabic handwritten character recognition.

At the realistic Arabic handwritten text recognition and explore two aspects of the problem. Firstly, a system based on segmentation-recognition integrated framework was developed for Arabic handwriting recognition. Secondly, the parameters of embedded classifier initialed at character-level training were discriminatively re-trained at string level.

The segmentation-recognition integrated framework runs as follows: the written character is first over-segmented into primitive segments, and then the consecutive segments are combined into candidate patterns. The embedded classifier is used to classify all the candidate patterns in segmentation lattice. According to Genetic Algorithm (Crossover, mutation, and population), the system outputs the optimal path in segmentation-recognition lattice, which is the final recognition result. The embedded classifier is first trained at character level on isolated character and then the parameters are updated at string level on string samples.

Keywords: Arabic Character, handwritten recognition, Genetic Algorithm.

1. INTRODUCTION

Handwritten Arabic character recognition has been the focus of intelligent computer interface research, a typical handwritten Arabic character recognition system generally consists of two components: the front word recognizer and back-end language decoder (post processing system). The complex structure of Arabic characters changeable, and front word recognizer accuracy is limited, and also the decoder post-processing language often used in handwritten Arabic character recognizer to improve the performance. We will use front word after-treatment system with integrated to improve handwritten Arabic character recognition system for the overall performance and get a greater impact[1]. General font word output a character matrix of candidate H_i (C_{ji},d_{ij}) C_{ji} is the first images of handwritten Arabic characters and candidate, d_{ji} is to identify the distance, it means $I_{i\ j}$ class candidate and the characters of the pattern similarity between the I_i. Which, d_{ii} represents front word is very important information, in the establishment of handwritten Arabic character recognition post-processing system, to have it integrated with the terms of the statistical information together, constitute a post-processing language of the state space decoder value of the node.

However,both together,the search for the best in the state space ofcandidate have a direct impact statement[2].

Genetic algorithms (GA) is a global, robust and relevant to the data search technology, which is likely to be automatically and efficiently search the system parameters to obtain better performance of a good way. In this paper, genetic algorithm optimization outstanding ability to propose a genetic algorithm parameters based on automatic optimization.

2. HANDWRITTEN ARABIC CHARACTER RECOGNITION MODEL

A typical schematic of handwritten Arabic character recognition as shown in Figure (1).

Front word input terminal receiving a handwritten Arabic character image sequence I = I_1 I_2 ⋯, I_n, generate a candidate vector H_i (C_{ji}, d_{ji}), of which, C_{ji} that the first i- handwritten Arabic character images I_i and the first j is candidate, d_{ji} is to identify the distance, it means I_{ij}-class candidate and the characters of the pattern similarity between the I_i. candidate statement S= C_j^1 1
C_j^2 2 ⋯ C j_nn, where C_{ji}i ∈ H_i, and 0 <i≤n decoder post-processing language of the candidate will use the language model matrix H = H_1H_2⋯ H_n decode, and select the best candidate statement S as output.

$$\hat{S} = \arg_s \max P(S|I) =$$

$$\arg_s \max \frac{P(S|I)\,P(s)}{P(I)}$$

$$\arg_s max P(I|S)\,P(S) \ldots\ldots (1)$$

Equation (1), front word provides P(I|S) estimates, and P(S) provided by the decoder post-processing language.

3. SOLVING *p (I|S)*

P(*I*|*S*) can be expressed as

$$P(I|S) = \prod_{i=1}^{n} P\left(\frac{I_i}{C_{i,j}}\right) \ldots\ldots\ldots (2)$$

By the Bayes formula, can be

$$P(I_i|S_{i,j}) = \frac{P\left(\frac{C_{ji}}{I_i}\right) * P(I_i)}{P(C_{ji})} \ldots\ldots (3)$$

Equation (3), for H_i for all the candidate Arabic characters, I_i is the same, so the p (I_i) can be omitted. P(C_{ji}i) is a candidate character C_{ji}i priori pattern class corresponding probability, and the front word is generally believed that p (C_{ji}i) is evenly distributed, so p (Cjii) can be omitted. Therefore, equation (1) can be rewritten as

$$\hat{S} = \arg_s max P(I|S) =$$

$$P(I|S) = \prod_{i=1}^{n} P\left(\frac{I_i}{C_{i,j}}\right) * P(S) \ldots\ldots\ldots (4)$$

p(C_{ji}i|I_i) candidate characters C_{ji}i by the model class and I_i calculated the distance between the [4]:

$$P\left(\frac{C_{i,j}}{I_i}\right) = \frac{1/d_{ij}}{\sum_{i=1}^{m_i} 1/d_{ki}} = D_{ji} \ldots\ldots (5)$$

I_i where m_i is the number of candidate characters, D_{ji} said that after identifying the probability of the transformed distance

$$\sum_{k=1}^{m} D_{kj} = 1$$

4. SOLVING *p (S)*

Intuitively, a candidate for the adjacent columns to form an Arabic character word can help us to choose an optimal candidate statement. In addition, since the word-based model of Arabic N-gram entropy is far more character-based Arabic language model for N-gram low. Therefore, the language decoder post-processing, but should use the word-based model. According to Arabic word segmentation algorithm, the candidate statement, S = $C_{j11}C_{j22} \cdots C_{jnn}$ can be divided into a word sequence.

$$S = w_1^m = w\,1\;w\,2 \ldots w_m$$

So that $S = W_1^m =^{def} W\,1\;W\,2 \ldots W_m$, is the Arabic word in real time. Therefore, in the post-processing language decoder be based on Arabic words model probability p(S) as in formula

$$P(S) = P(w\,1)\prod_{i=2}^{m} P(w\,i\,|w\,i-1) \ldots (6)$$

Of equation (6) applications Bayes formula,

$$P(S) = P(w\,1)\prod_{i=2}^{m} \frac{P(w\,i\,|w\,i-1)}{P(w\,i-1)} \ldots\ldots (7)$$

Meanwhile, the corresponding word recognition will also be converted from the D ji candidate statement in the identification of *S* in the word from the DW_j:

$$DW_j = \prod_{j=k}^{1} D_{ji} \ldots\ldots\ldots (8)$$

One, DW j wj word recognition that distance, the word length is 1-k +1.

As for training the model can not be for infinite text, data sparseness problem be smoothing to solve this problem[3], the smoothing formula mentioned as follows:

$$P(w\,i\,|w\,i-1) = \lambda_b P(w\,i\,|w\,i-1) + \lambda_b\, P(w\,i) \ldots (9)$$

Where

$$\hat{S} = \arg_s max P(S|I) =$$

$$\arg_s max \prod_{i=1}^{m} DW_i * P(w\,1)$$

$$\prod_{i=2}^{m} \frac{P(w\,i\,|w\,i-1)}{P(w\,i-1)} \ldots\ldots (10)$$

$$\hat{S} = \arg_s max \lambda_1 \sum_{i=1}^{m} 1nP(w\,i + \lambda_2 p(w\,1) \sum_{i=2}^{m} (1nP(w\,i\,w\,i-1) - 1nP(w\,i-1)) \ldots\ldots (11)$$

Where λ_1 and λ_2 are fitting parameters of the system, to optimize two parameters on the final recognition rate. A flowchart of process of the proposed work be shown in figure (2).

5. BASIC BINARY CODED GENETIC ALGORITHM

Consider the following objective function of the problem:

$$\min_{X \in \Omega} f(X) \dots\dots (12)$$

Where X = $(x_1, x_2, \cdots, x_n) \in R^n$ as the independent variables, Ω is X, the domain of f is the solution space $X \in \Omega$ to the real number field $F \in R$ as a mapping, that is the objective function[4].

Genetic algorithms (GA) can be described as formal GA = (N, X_{init}, SΩ, $\otimes\Omega$, $\nabla\Omega$, f, τ), where, N is population size, $X_{init} = (X_1, X_2, \cdots, X_N)$ for the initial population, X_i ($1 \le i \le N$) for the population of individuals, as encoded (encoding) of the binary string, SΩ, $\otimes\Omega$, $\nabla\Omega$ genetic algorithm were three basic operators: select (selection), cross-operating (crossover), mutation (mutation), by the three operators to the initial population X_{init} from starting to produce the desired improvement of population, f is objective function, τ for the algorithm termination conditions, usually expressed as the maximum number of iterations algorithm execution form.

6. DETERMINATION OF THE OBJECTIVE FUNCTION

In any use of genetic algorithms, we must first solve the most critical issue is which determine the merits of the individual to determine the objective function[5].

In equation (11), $\tilde{S}$ can be the objective function, but the arithmetic is more complex and less intuitive. From the known characteristics of handwriting recognition.

7. RESULTS

a. Handwritten Arabic Characters

In the form we will starting writing in image field any character belong to Arabic language, which later will be recognized, in same time the buffer data will check if this character is found in database in front word as shown in figure (3).

b. Recognizing

After the character been recognized, by get the data from database, will found the similarity of data which in matrix of buffer data, and will show the result in data image, as seen in figure (4).

c. Learning

In case of the Arabic character not recognized well, that mean its not found in database (not recognized), in this case, we will have to learn the machine about this new Arabic character, in our case in figure (5) is "ق" which the result showed in data image is similar to "ث".

d. Optimization and Detection

After we learned the machine regards the word in study "ق", cleared the image of handwritten, and re write the character "ق", as seen in figure (6), the result in data image and percentage of recognized of character.

Example:
a. Recognition Stage: as shown in figure (7)
b. Learning Stage: as shown in figure (8)
c. Optimization and Detection Stage: as shown in figure (9)

7. CONCLUSIONS

This paper derived line handwritten Arabic character recognition system, front-end word recognizer and the effective integration of back-end language decoder mathematical model, then describe the use of genetic algorithms (GA) optimization of the system parameters.

As seen in the results, the accuracy is high, and its depends on the way to draw the character. Even if was a word, will search each character and find the best accuracy (highest).

Also, the practical application of online handwritten character recognition in our study is demand robust and highly accurate recognition along with low memory requirements. By combines the advantages of generative classifiers to address the similarity of between-class samples, while taking into account the variability of writing Arabic styles within the same character class. To model the significant writing Arabic styles in a memory-efficient manner. However, by created software application with a large number of training samples is needed up front, which is not desirable or available in many practical applications, we allow a recognition to begin with a small number of training samples, and adapts the classifier to the new samples presented to the system during recognition.

In addition to that, by depending on propagation algorithm for system parameter optimization method is a very simple and effective way, in other areas of artificial intelligence will also have a very broad application prospects.

8. REFERENCES

1. Jane wightwick, Mohmoud Caafar; 2005, "Mastering Arabic Script a Guide to Handwritten", plagrave (macmillan); p.88-p90.
2. Marcus Liwicki, Horst Bunke; RECOGNITION OF WHITEBOARD NOTES Online, Offline and Comb Dominik ´ Sle¸zak William I. Grosky Niki Pissinou Timothy K. Shih, Tai-hoon Kim Byeong-Ho Kang; Communications in Computer and Information Science; 2009.
3. MOHAMED CHERIET; NAWWAF KHARMA; CHENG-LIN LIU; CHING Y. SUEN; CHARACTER RECOGNITION SYSTEMS A Guide for Students and Practioners; 2007.
4. Jeng-Shyang Pan, Shyi-Ming Chen, Ngoc Thanh Nguyen; Computational Collective Intelligence; 2010.
5. Jeng-Shyang Pan, Shyi-Ming Chen, Ngoc Thanh Nguyen; Computational Collective Intelligence; 2010.

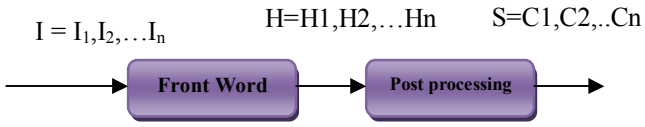

Fig.(1): Handwritten Arabic Character Process.

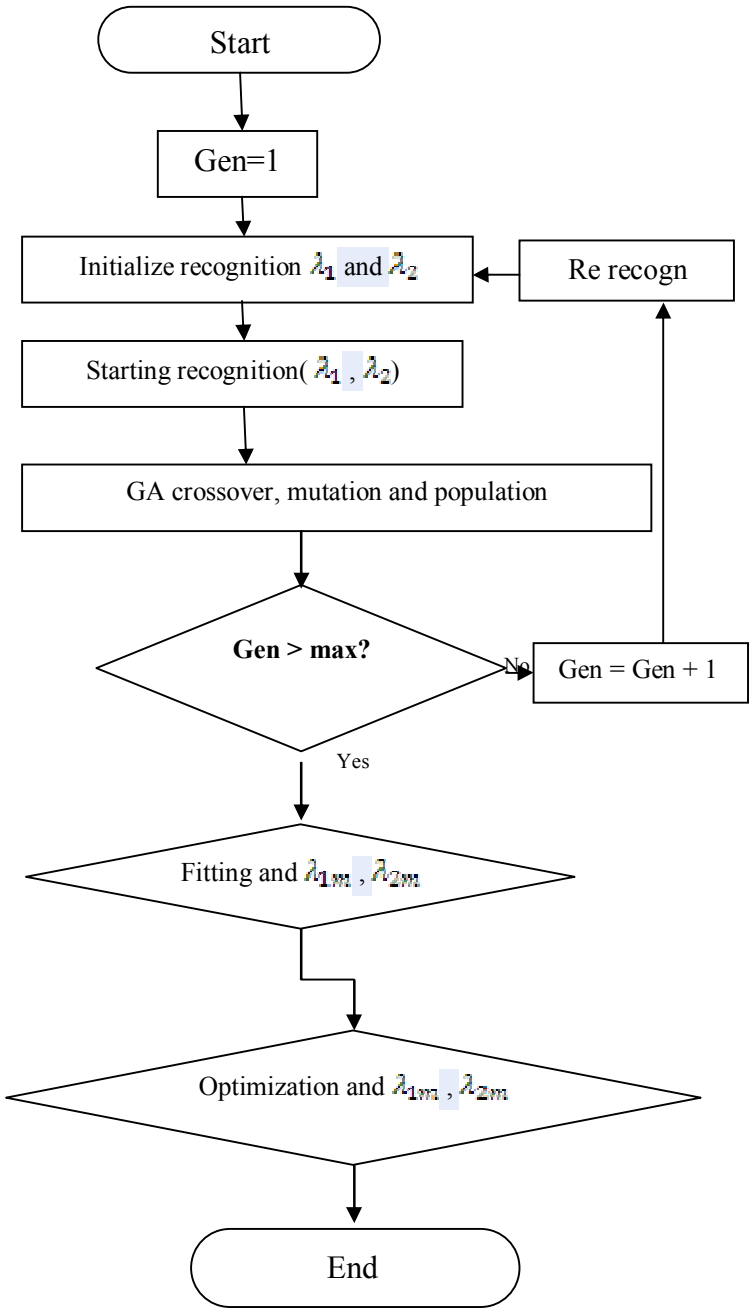

Fig.(2): online using genetic algorithm flowchart.

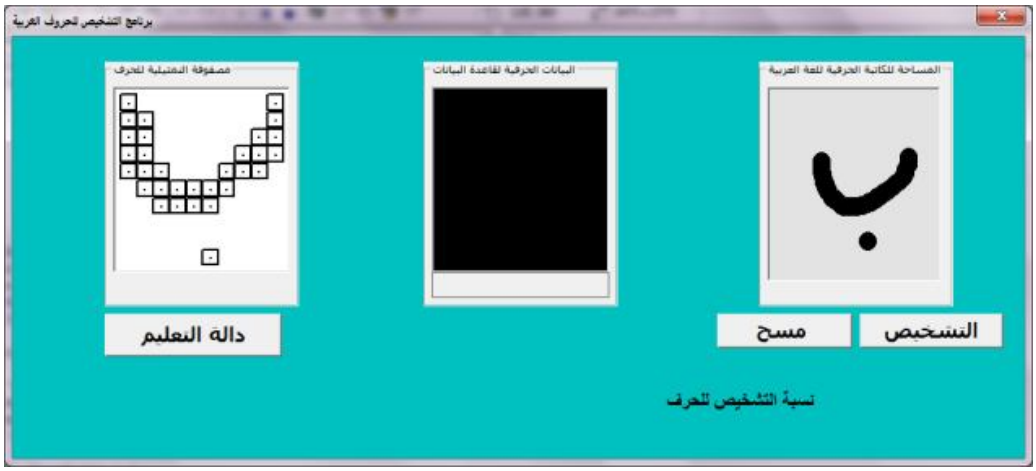

Fig.(3): Handwritten start.

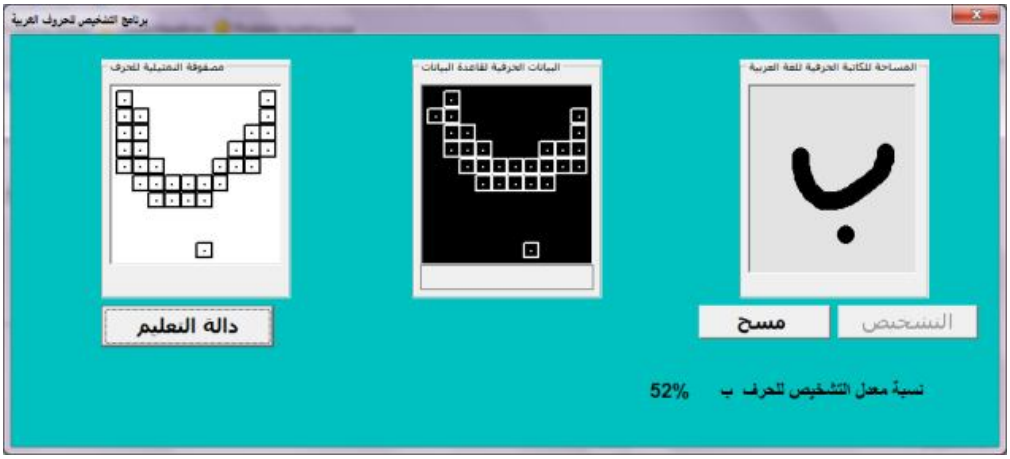

Fig.(4): Recognition step.

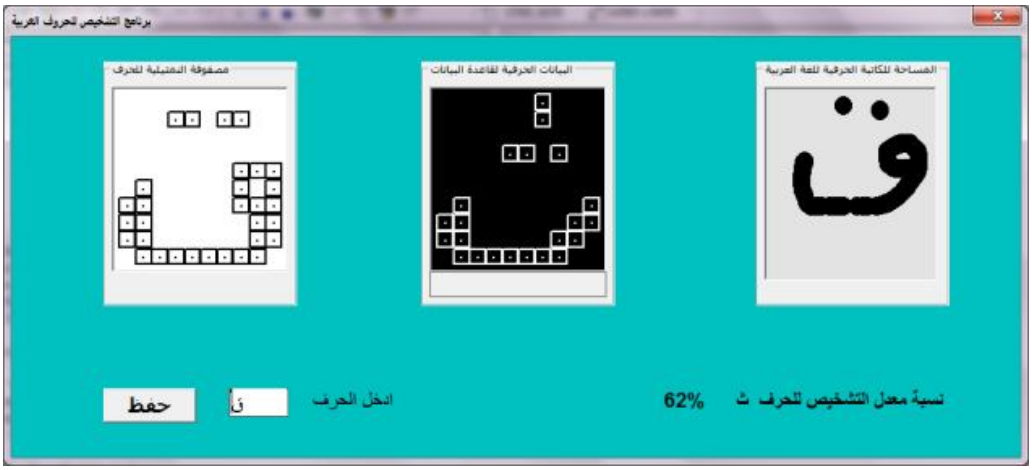

Fig.(5): Learning of character ”ق“.

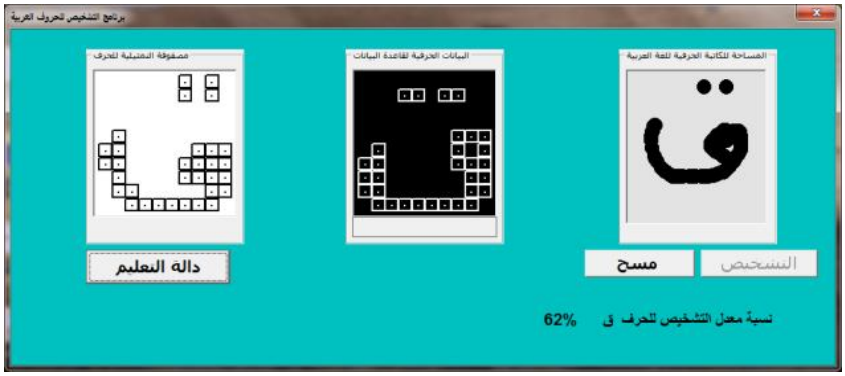

Fig.(6): Optimization of character ”ق“.

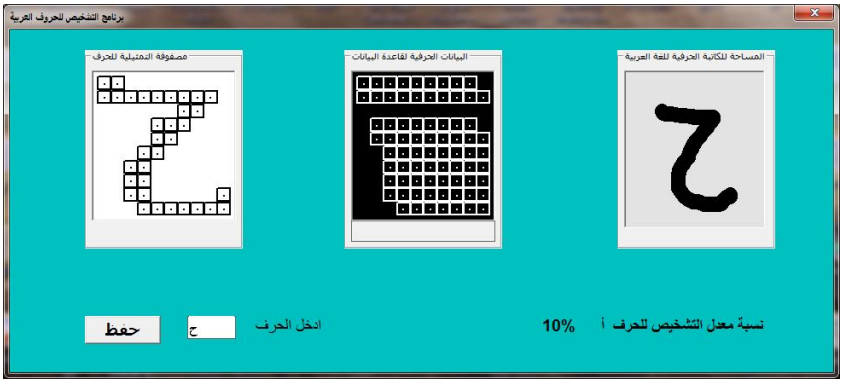

Fig.(7): Recognition Form.

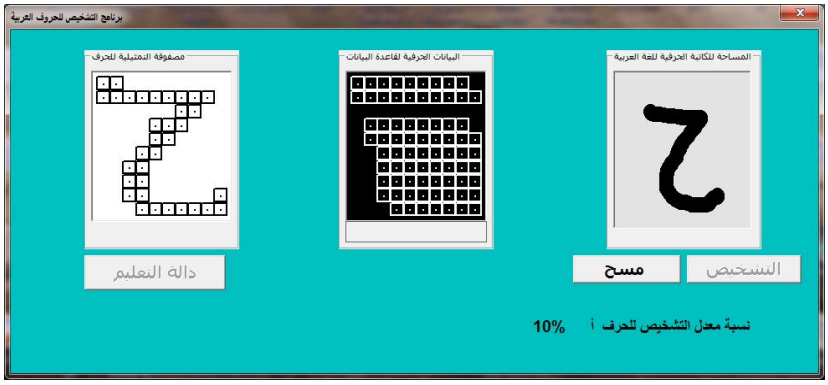

Fig.(8): Learning Form.

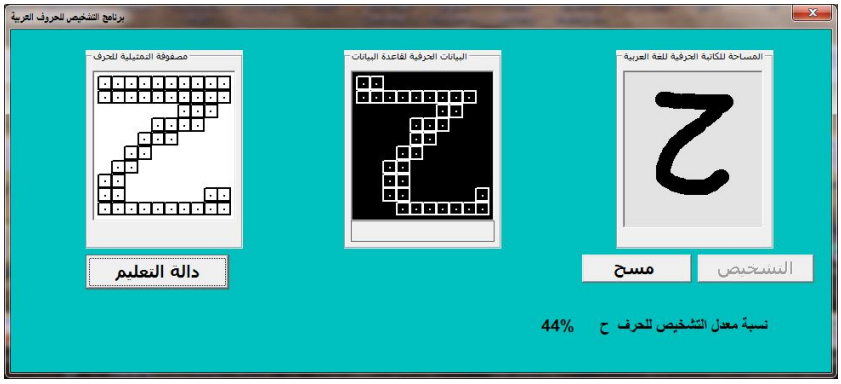

Fig.(9): Optimization and Detection Form.

تشخيص الاحرف العربية بالكتابة المباشرة إعتماداً على الخوارزمية الجينية

م.م. هيثم عبد الرحيم طه

قسم الهندسة الالكترونية، الجامعة المستنصرية

الخلاصة

ان تشخيص كتابة الاحرف العربية في الزمن الحقيقي يعتبر من اكثر التحديات في مجال التعرف على الاشكال. وهناك دراسات حول التشخيص للاحرف العربية ما زالت تعتبر من المشاكل المعقده في هذا المجال.

الهدف من البحث هو التشخيص للاحرف العربية بخط اليد واستكشاف الجوانب الثلاثة في بحثنا، اولاً ان النظام يقوم على تشخيص بشكل مجزء ضمن الاطار المتكامل للتعرف على خط اليد باللغة العربية، ثانياً، ان البارامترات الخاصة بالتصنيف لعملية البدء عند مستوى التعليم لحالة الحرف المخطوط يدوياً تقوم بتمييز مستوى الحرف عند مستوى التدريب.

تشخيص المقاطع في اللوحة المهيئة لكتابة الحرف تكون حسب التالي:كتابة الحرف المقاطع الاولية، وبعد ذلك يتم جمع المقاطع على التوالي في نموذج ترشيحي وثم يتم تصنيف جميع النماذج في شبكه عبارة عن مقاطع. وبالاستناد الى الخوارزمية الجينية (التبادل، الطفرات، والتوزيع)، فان الاخراج يكون ناتج عن المسار الامثل في للنموذج الترشيحي وهي التي تعتبر الناتج النهائي. يتم تدريب المصنف الأول على مستوى جزءا لا يتجزأ وذلك بعزل حرف على حرف ومن ثم يتم تحديث المعلمات على مستوى السلسلة على عينات تم حفظها في قاعدة بيانات لمقارنتها مستقبلاً بالادخال الجديد.