Optical Character Recognition based on Genetic Algorithms and Machine Learning

Arafat A. Muharram
Associate professor
Department of
Computer Science
Faculty of Computer
Science and Engineering
Hodeidah University,
Yemen

Khaled M. G.
Noaman
Professor
Department of
Distance Learning
Deanship of E-Learning
and Distance Learning
Jazan University, K.S.A

Ibrahim Abdulrab
Ahmed
Professor
Department of
Information System
Community College
Najran University, K.S.A

Jamil A. M. Saif
Associate Professor
Department of
Information System
Community College,
Bisha University, K.S.A

ABSTRACT

Pattern recognition is known to be one of the earliest applications of image processing. Genetic algorithm and Machine Learning have been used in this study to recognize English alphabets which are represented as matrix one and two dimensions. Genetic algorithm and machine learning were used in this paper to compare their efficiency and accuracy regarding   concrete conditions, testing and evaluation results, it has ben got 95% for Genetic Algorithms and 94% for Machine Learning.

Keywords

Genetic Algorithms, character recognition, machine learning.

1. INTRODUCTION

Optical Character Recognition   has become an important and widely used technology for pattern recognition. Artificial intelligence elements like, Machine learning , genetic algorithms, fuzzy logic, expert systems, SVM, neural networks etc. are tending to emulate the human brain and are one of the main interests of the researches nowadays[11].

By considering the image processing methods like segmentation,   which divide the image into small parts that are the constituents of the original images. In filtering, it vanishes the scattered spots as well as the noise found in the image. Thus a pure and clean image comes out. The size of the image gets minimized and is used for saving memory. the available pixels are taken specific square areas (n*n) and get the average of each of them to get a smaller area containing these averages[4] [12].

The genetic algorithms, recently and currently are used in a range of problems like scheduling, images creating, planning strategy, predicting with dynamical systems, classification etc. [9] [2]. Populations, encoding, crossover and mutation operators, fitness functions and selection are generally used to evaluate the performance of the Genetic Algorithms. However, many genetic algorithms try to optimize their population by means of a static fitness function that can be derived from performance on a fixed set of examples [8 [5].

The Machine Learning field also evolved   from the broad field of Artificial Intelligence, which aims to mimic intelligent abilities of humans by machines[13]. An important task in Machine Learning is classification. Machine Learning can recognize the shape of the character ,which depend on the black pixels taken individually, especially those shapes in the interior of the objects shown in this paper. They don't carry any information whether they are part of a representation of

an object or another. It is their spatial interrelations which make us recognize the shape and create variations such as the fonts of different boldness of the object[10] [7].

2. EXPERIMENTS DESIGN

The data are images containing 26 alphabets with different shapes and further each alphabet has a different shapes. By using image processing methods like segmentation, filtering, minimization and average, alphabets are represented as an array of one and especially two dimensions to be applied in selected methods.

Every character after image processing will be converted to an array of pixels, an array of 5 rows and 5 columns is used as a standard for our work as follow: In this study MATLAB 7.0.1 package is used for the application of genetic algorithm and machine learning to initialize application values.

Furthermore, Roulette Wheel Selection GA method is used as parents to crossover that it has been applied to a set of binary numbers (0, 1) encoding representing capital English alphabets from A to Z. Furthermore, decision tree MA method is used as a function of learned tree which was represented as a set of IF THEN rules , each node in the tree specifies concrete   feature and the leafs of this tree contains all alphabets from A to Z .

3. RESEARCH METHODS

3.1 Genetic Algorithms

For each population, two or more chromosomes are selected to be parents to crossover. The problem   faced by the researchers have to know how to select these chromosomes to create a new offspring. There are many methods to overcome this problem and for selecting the best chromosome like Roulette Wheel Selection, Boltzmann Tournament Selection for Genetic, Rank Selection and some others[6]. The chromosomes with higher fitness have a higher possibility to be selected to produce offspring for the next generation. After many generations of evolution, the optimal solution of the problem is hopefully to be found in the population [7] [8] [9]. The selected fitness function that that has been utilized   in the research is as follows:

$$F.F = \frac{1}{\sum_{i=1}^{n}(\alpha_i - \beta_i)^2 + 1}$$

Where:

α: desired gene from the Actual data;

β: fault level of a chromosome computed by genetic algorithm

n: number of genes in chromosome.

From the fitness function, if the fault level of a chromosome is equal to the desired, the fitness value of that chromosome will be equal to 1 and the process of GAs will be terminated or if the generation no. is more than 40 then the process of GAs will be terminated.
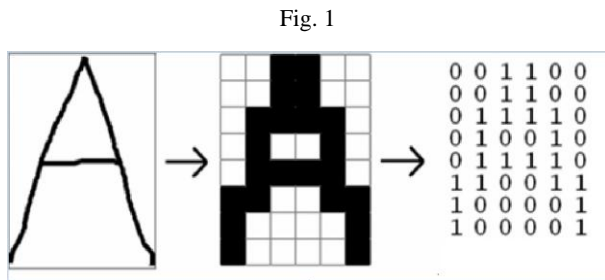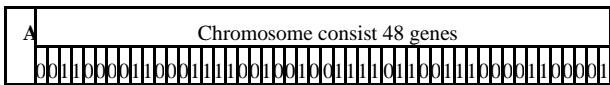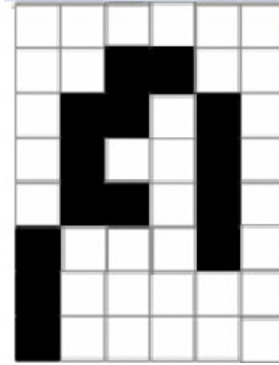
Fig. 1

Table 1

A	Chromosome consist 48 genes
	001100001100011110010010011110110011100001100001

3.2 Test for unclear alphabet A

Fig. 2

Initial chromosome for primary offspring (Table 2)

Table 2

A	Chromosome consist 48 genes
	000000011000110100100100110101000101000010000000
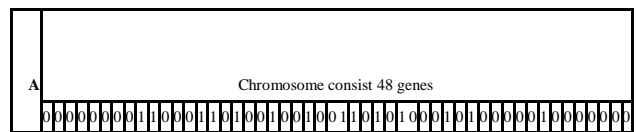
Account of generated chromosomes in one generation is 100 chromosomes, highlighted as below examples between them two chromosomes selected with high fitness values:

Table 3

0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	0	1	0	0	1	0	0	1	0	0	1	1	0	1	0	1	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	0	1	0	0	1	0	1	1	1	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	0	0	1	0	1	0	1	0	1	0	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	1	0	0	0	1	1	0	1	0	0	1	0	0	1	0	0	1	1	0	1	0	1	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	1	0	0	1	0	0	1	1	0	0	0	1	0	0	1	0	1	0	1	0	0	0	1	1	0	0	1	1	1	0	0	0	0	1	1	0	0	0	0	0	0
0	0	1	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	1	0	0	1	0	1	1	1	0	0	0	1	0	0	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0
0	0	1	1	0	0	0	1	1	0	0	0	1	1	0	1	0	0	1	0	0	1	0	0	1	1	0	1	0	1	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	1	0	1	0	0	1	1	1	0	0	0	1	0	1	1	1	1	1	0	1	0	0	0	1	0	0	0	1	1	0	0	0	0	1	1	1	0	0	0	0	0
0	1	1	1	0	0	1	0	1	0	0	0	0	1	1	1	0	0	0	1	0	0	1	0	1	0	1	1	0	0	1	0	0	0	1	1	0	0	0	1	0	1	0	1	0	1	0	
0	0	1	1	1	0	0	0	1	0	0	0	0	1	1	1	0	0	0	0	0	0	1	0	1	0	1	0	0	0	1	0	0	0	1	1	0	0	1	0	1	1	0	0	1	1	0	

Now we will choose two chromosomes with a higher fitness and applying crossover and mutation operation to get new chromosomes, the two selected chromosomes (see bold rows in Table 3) are as bellow.

Table 4 consist Chromosome 1, Table 5 consist chromosome2

Table 4

0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	0	1	0	0	1	0	0	1	0	0	1	1	0	1	0	1	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0

Table 5

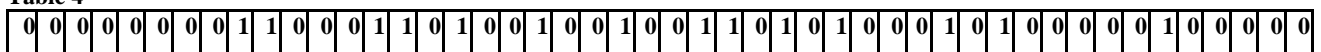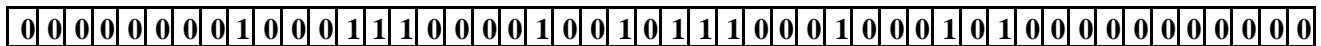| 0|0|0|0|0|0|0|0|1|0|0|0|1|1|1|0|0|0|0|1|0|0|1|0|1|1|1|0|0|0|1|0|0|0|1|0|1|0|0|0|0|0|0|0|0|0|0|0 |
|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|-|

Using fitness calculation function, it has been found that a function values are greater than the previous two bold chromosomes.

Hence, a crossover for these chromosomes are given as follows:
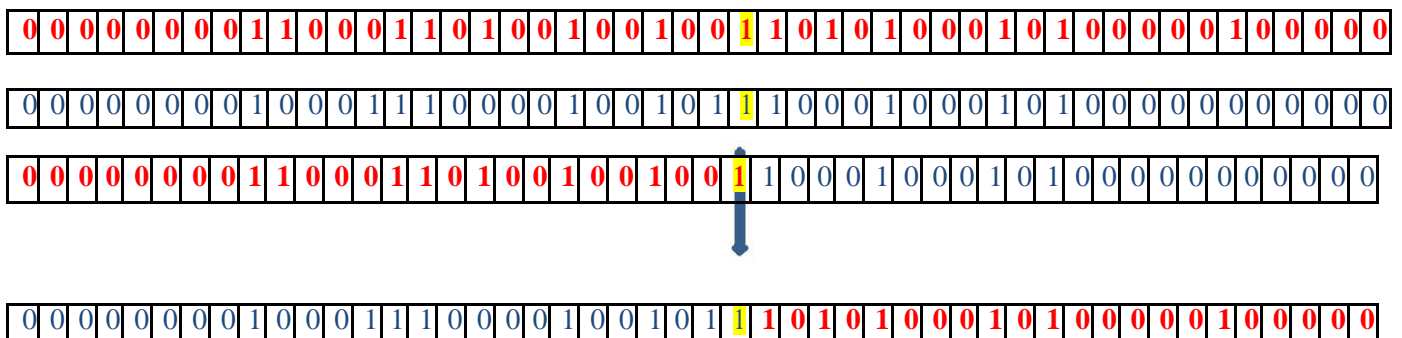
Crossover for those two chromosomes

Moreover, all configured chromosomes and crossover process or mutation are as follows:

After account of fitness value that genetic algorithm is able to identify the characters in the 40th generation

0	0	0	0	0	0	0	1	1	0	0	0	1	1	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0

0	0	0	1	0	0	0	0	1	1	0	0	0	1	1	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1

Table 6

Final Chromosome after testing

0	0	1	1	0	0	0	0	1	1	0	0	0	1	1	1	1	0	0	1	0	0	1	0	1	1	1	1	1	1	0	1	1	0	0	1	1	0	0	0	0	0	1	1	0	0	0	1

Table 6 consists final chromosome, which specify character A after testing with different only in two genes (in 25th and 37th position) comparing with Table 1 which specify normal (clear) character A as target

3.3 Machine Learning

Ml describes a variety of learning paradigms, algorithms, theoretical results and applications. The key ideas are presented from various mentioned fields by using a non-specialist vocabulary, with unfamiliar terms and concepts introduced as the need arises.

The learned function is represented by a decision tree. Learned trees can be represented as a set of if-then rules to improve human readability. Each node in the tree specifies a test of some attributes of the instance [4][10].

The researchers tried to classify English alphabet letters according to the features of their shapes and build a tree to separate the determinated alphabets group to the two groups and continue using another features to reach the final node which are the leaf of the tree which contain the alphabet letters .

English text features are:

| - \ (
| left ,middle, right
-          up ,middle, down
\   left ,right, small, large
(left , right, up ,down

Fig.3

The researchers tried by counting these features to know how to match the character by its features regarding   its shape.

The following example illustrates how how ML technique is made   to match character shape features

3.4 Test for selected letters as example : E H B M A C O Z

The table 7 below shows the number of character recognition correctly by the Machine Learning method.

Table 7 The Average Results ML

Characters	ML
A	9
B	10
C	9
D	9
E	10
F	9
G	9
H	9
I	10
J	9
K	9
L	10
M	9
N	10
O	10
P	9
Q	9
R	9
S	9
T	9
U	10
V	9
W	10
X	10
Y	9
Z	10
Average	9.38

Table 8 Testing Results of character recognition in percentage

Testing Criterion	Method	
	M L	**GA**
	Correctly activity around (%)	Correctly activity around (%)
Testing Result	**94**	**95**

The correctly classified instances of the experiments are summarized in table 8,which shows that Machine Learning and Genetic Algorithms are able to increase the specificity up to 95%.

4. RESULTS AND DISCUSSION

The performance and the ability of Genetic Algorithms and Machine Learning to recognize English alphabet letters were realized regarding the standards matrices, which represents originals alphabets letters(A..Z) and features assigned a Boolean classification of decision tree . The genetic algorithms being trained and then calculate a fitness function value. It has been noted that a character recognition process in the first iterations produces comparatively big values for specific character. So, it take a long time and more iterations training because of fitness functions convergence values and due to an unclear features of character 'A'(for example) . Training accomplish that genetic algorithm was able to identify the character in the 40th generation with 95% . Maybe ,since the Machine Learning (Decision Tree Method) which we used in this paper assigns a Boolean classification (yes or no) for the English alphabet (A-Z) , beginning a disjunction of conjunctions of constraints on the attribute values of instance. Each path from the tree root to a leaf corresponds to a conjunction of attribute test, and the tree itself to a disjunction of these conjunctions, machine learning was able to identify characters with 94%. Maybe these Satisfactory results which was used had got from training by GA and classification using ML because of the mechanisms , are highly selected methods in this paper

5. CONCLUSION

This paper signifies the important achievement of selecting and using Genetic Algorithms and machine learning methods as an interesting field of AI. accurately learning methods have been evaluated. The study and the selected methods which has been used in this research has classified instances correctly with high classification . This May be because of the strength of the selected methods of GA and ML , which are used in this paper. So These fields have better capability not only in classification or recognition but also in responding to learning faster as compared to others algorithms, which

produces powerful function in optimization and classification. The overall study shows that GA and ML generate better results by training of system and testing in a record time as compared with other algorithms.

6. REFERENCES

[1] Alpaydin Ethem, Introduction to Machin Learning , ,Istanbul,2014

[2] Chong E.K.P., Zak S.H., An introduction to optimization ,New York,1996

[3] Christopher M. Bishop, Pattern Recognition and Machin Learning , ,U.K,2012

[4] Cosmin Grigorescu, Student Member, IEEE, and Nicolai Petkov: Distance Sets for Shape Filters and Shape Recognition, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 12, NO. 10, OCTOBER 2003

[5] George F.Luger Artificial Intelligence :Structure and Strategies for Complex Problem Solving, Fourth Edition .India, 2002

[6] Glodberg David E. ,Genetic Algorithms in search optimization and Machin Learning, Addison –wesely publishing company, 2006

[7] Gunnar Ratsch, A Brief Introduction into Machine Learning, Friedrich Miescher Laboratory of the Max Planck Society , Germany, www.ccc.de/congress/fahrplan/files, 2004

[8] Haupt R.L., Haupt S.E. ,Practical Genetic Algorithms, New York , 2004

[9] Kenneth A. De Jong William M. Spears, Learning Concept Classification Rules Using Genetic Algorithms, USA

[10] Michalewicz Z., Genetic Algorithms +data structure=evolution programs, New York,1996

[11] Optical Character Recognition, RavinaMithe,SupriyaIndalkar, NilamDivekar,International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-2, Issue-1, March 2013

[12] Russ, J.C., The Image Processing Handbook., Boca Raton, Florida: CRC, Second ed. 1995

[13] Sivanandam S.N, Deepa S.N., introduction to genetic algorithms ,2007,New York

[14] TOM M. MITCHELL, Machine Learning ,McGraw-Hill, USA,1997