Efficient Genetic Algorithms for Arabic Handwritten Characters Recognition

Dr. Laheeb M. Al-zoubaidy
Associative Proff.[a]
Department of Computer Science, College of Computer Sciences & Math.,
Mosul University Iraq

Abstract. The main challenge in handwritten character recognition involves the development of a method that can generate descriptions of the handwritten objects in a short period of time. Genetic algorithm is probably the most efficient method available for character recognition. In this paper a methodology for feature selection in unsupervised learning is proposed. It makes use of a multiobjective genetic algorithm where the minimization of the number of features and a validity index that measures the quality of clusters have been used to guide the search towards the more discriminate features and the best number of clusters.

The proposed strategy is evaluated synthetic data sets and then it is applied to Arabic handwritten characters recognition. Comprehensive experiments demonstrate the feasibility and efficiency of the proposed methodology, and show that Genetic Algorithm (GA) are applied here to improve the recognition speed as well as the recognition accuracy.

Keywords. Arabic handwritten characters recognition, Genetic Algorithm (GA), Feature Extracted, Feature Selection

1 Introduction

In areas of pattern recognition, features can be characterized as a way to distinguish one class of objects from another in a more concise and meaningful manner than is offered by the raw representations. Therefore, it is of crucial importance to define meaningful features when we plan to develop a good recognizer, although it has been known that a general solution has not been found. In many cases, features are generally defined by hand based on the experience and intuition of the designer.

Depending on problems given, there are a number and variety of features can be defined in terms of extracting methods and ways of representation. In many practical applications, it is not unusual to encounter

[a]Corresponding Author.
E-mail addresses: lahmzub@yahoo.com

problems involving hundreds of features. The designer usually believes that every feature is meaningful for at least some of the discriminations. However, it has been observed in practice that, beyond certain point, the inclusion of additional features leads to worse rather than better performance. Furthermore, including more features means simply increasing processing time. This apparent paradox presents a genuine and serious problem for classifier design [5].

Arabic is a major world language spoken by 186 million people [6]. Very little research has gone into character recognition in Arabic due to the difficulty of the task and lack of researchers interested in this field. As the Arab world becomes increasingly computerized and mobile, and technology becomes increasingly ubiquitous, the need for a natural interface becomes apparent.

Classification of Arabic handwritten characters, which has been a typical example of pattern recognition, contains the same problem. Due to diversity in ones written by a single person. In order to deal with such a wide range of diversity existing in handwritings, recognizers often employ several hundreds of features.

Approaches to circumvent the feature selection problem found in the literature are: linear combination of feature vectors, principal component analysis, simple selection based on the discrimination power of features and sequential forward/backward selection. Because feature dimensions are large enough, and the solution space has the characteristics of the multi-modal function, the feature selection process takes a lot of time when most of the above mentioned approaches are adopted, and a local optimum can be chosen as the solution, instead of the global one [10,12]

In this paper, we introduce a feature selection method, which can minimize most of the problems can be found in the conventional approaches, by applying genetic algorithms(GA) which recently received considerable attention regarding their potential as an optimization technique for complex problems. Genetic algorithms are stochastic search technique based on the mechanism of natural selection and natural genetics.

In this paper we propose a methodology for feature selection in unsupervised learning for handwritten Arabic characters recognition. It makes use of the Genetic Algorithm (GA) proposed by Kim [4] which deals with multi-objective optimization. The objective is to find a set of non dominant solutions which contain the more discriminate features and the more pertinent number of clusters. We have used two criteria to guide the search: minimization of the number of features and minimization of a validity index that measures the quality of clusters. A standard K-Means algorithm is applied to form the given number of clusters based on the selected features.

Afterwards, it is applied to handwritten Arabic characters recognition in order to optimize the character classifiers. Experimental results show the efficiency of the proposed methodology

2 Genetic Algorithms

In this section we present a brief introduction about genetic algorithms. [8].

The genetic algorithm is a model of machine learning which derives its behavior from a metaphor of some of the mechanisms of evolution in nature. This is done by the creation within a machine of a population of individuals represented by chromosomes, in essence a set of character strings that are analogous to the base-4 chromosomes that we see in our own DNA.

The individuals represent candidate solutions to the optimization problem being solved. In genetic algorithms, the individuals are typically represented by n-bit binary vectors. The resulting search space corresponds to an n–dimensional boolean space. It is assumed that the quality of each candidate solution can be evaluated using a fitness function.

Genetic algorithms use some form of fitness-dependent probabilistic selection of individuals from the current population to produce individuals for the next generation. The selected individuals are submitted to the action of genetic operators to obtain new individuals that constitute the next generation. Mutation and crossover are two of the most commonly used operators that are used with genetic algorithms that represent individuals as binary strings. Mutation operates on a single string and generally changes a bit at random while crossover operates on two parent strings to produce two offsprings. Other genetic representations require the use of appropriate genetic operators.

The process of fitness-dependent selection and application of genetic operators to generate successive generations of individuals is repeated many times until a satisfactory solution is found. In practice, the performance of genetic algorithm depends on a number of factors including: the choice of genetic representation and operators, the fitness function, the details of the fitness-dependent selection procedure, and the various user-determined parameters such as population size, probability of application of different genetic operators, etc. The specific choices made in the experiments reported in this paper are summarized in Table 2, (see Figure 1) depicts a GA cycle.

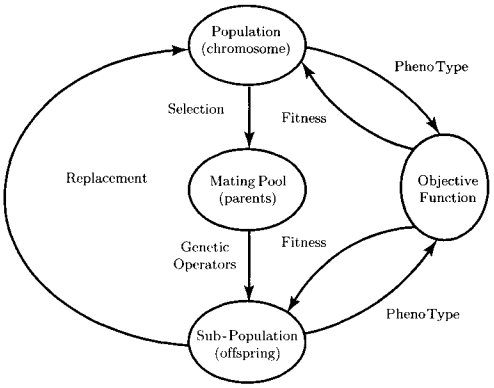

Fig. 1. Genetic algorithm (GA) Cycle

The basic operation of the genetic algorithm is outlined as follows [1]:

```
begin
 t <- 0
 initialize P(t)
 while (not termination condition)
   t <- t + 1
   select P(t) from p(t - 1)
   crossover P(t)
   mutate P(t)
   evaluate P(t)
 end
end
```

Since genetic algorithms were designed to efficiently search large spaces, they have been used for a number of different application areas, and genetic algorithms mainly deal with optimization problems. The applications include job scheduling, TSP (Traveling Salesman Problem), communication network design image restoration, data clustering, and feature selection for speaker identification, camera calibration [9], signature verification [9], medical diagnosis [11], facial modeling [2] and handwritten recognition [4].

3 Recognition System

Figure 2 briefly shows the system flow of a general handwriting recognizer. An input image is encoded for saving space and easier manipulation in the subsequent steps. Several steps of preprocessing, such as noise removal, slant correction, and smoothing are applied to the encoded image. Defined features are extracted from the processed image and the recognition process is performed using the features [7].

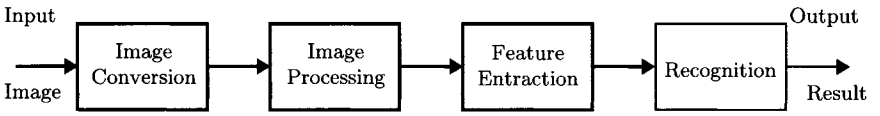

Fig. 2. Flow diagram of recognition system

Arabic handwritten character recognizer as shown in Figure 3 is used to evaluate the effect of the proposed feature selection algorithm. The recognizer has two phases (see Figure 3) training and testing.

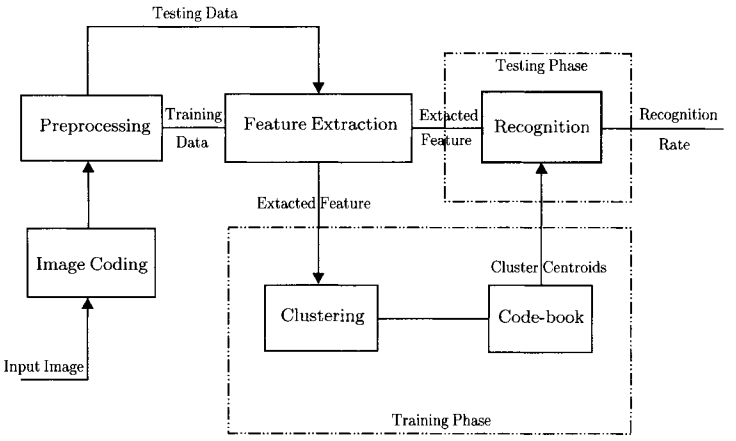

Fig. 3. Arabic handwritten character recognition system

3.1. Feature Extracted Using Genetic Algorithm (GA)

In this subsection we present the choice of a representation for encoding candidate solutions to be manipulated by the genetic algorithm. Each individual in the population represents a candidate solution to the feature subset selection problem. Let m be the total number of features available to choose from to represent the patterns to be classifier (m = 74 in our case). The individual (chromosome) is represented by a binary vector of dimension m. If a bit is a 1, it means that the corresponding feature is selected; otherwise the feature is not selected. This is the simplest and most straightforward representation scheme [4].

Feature, dimensions of 74, are defined and extracted so to see how the proposed feature extracted algorithm deals with features. The smaller set of 74 consists of 2 global features – aspect ratio and stroke ratio of the entire template, and 72 local features – distribution of the eight directional slopes for each of 9 (3*3) sub images [3]. (See Figure 4).

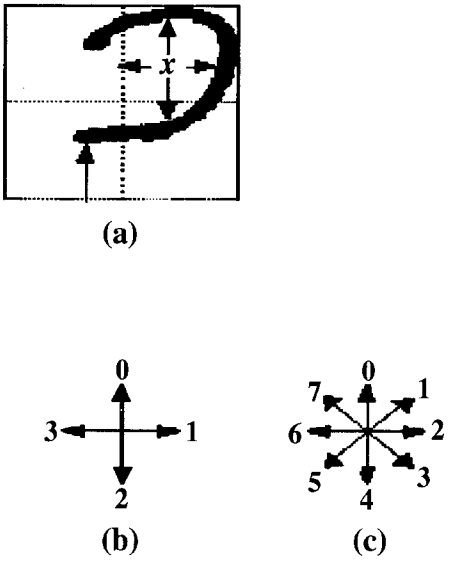

Fig. 4. Feature set for Arabic character Dal (د): (a) Concavities, (b) 4-Freeman directions (c) 8-Freeman directions

The feature extracted procedure based on Genetic Algorithm (GA). Since we are representing a chromosome through a binary string, the operator's mutation and crossover operates in the following way: Mutation operates on a single string and generally changes a bit at random. Thus, a string 11010 may, as a consequence of random mutation gets changed to 11110. Crossover on two parent strings to produce two offsprings. With a randomly chosen crossover position 4, the two strings 01101 and 11000 yield the offspring 01100 and 11001 as a result of crossover.

3.2. Selection Mechanism

The selection mechanism is responsible for selecting the parent chromosome from the population and forming the mating pool. The selection mechanism emulates the survival of the fittest mechanism in nature. It is expected that a fitter chromosome receives a higher number of offsprings and thus has a higher chance of surviving on the subsequent evolution while the weaker chromosomes will eventually die.

In this work we are using the roulette wheel selection [4] which is one of the most common and easy-to implement selection mechanism. Basically it works as follows: each chromosome in the population is associated with a sector in a virtual wheel. According to the fitness value of the chromosome, the sector will have a larger area when the corresponding chromosome has a better fitness value while a lower fitness value will lead to a smaller sector.

3.3. Feature Selection Using Genetic Algorithm (GA)

The recognizer evaluates the selected features and returns the results to the feature selection algorithm. The genetic algorithms evaluate the solution based on the recognition result. The process is repeated until the termination condition is satisfied.

The feature selection procedure, based on the simple genetic algorithm (SGA), is presented in this section. A chromosome is represented by a bit string, (see Figure 5(a)), in the Figure 5(a), '1's in the string represent the corresponding features are selected and '0's represent the corresponding ones are not selected. Figure 5(b) shows the feature selection process. A new generation is formed by selection some of the parents and offspring according to the fitness value, and rejecting others so as to keep the population size constant. Offspring are formed by either merging two chromosomes using a crossover process and modifying a chromosome using a mutation process. Fitter chromosome have higher probability of being selected. The recognizer evaluate the selected features and returns the results to the feature selection algorithm. The genetic algorithms evaluate the solution based on the recognition result. The process is repeated until the termination condition is satisfied.

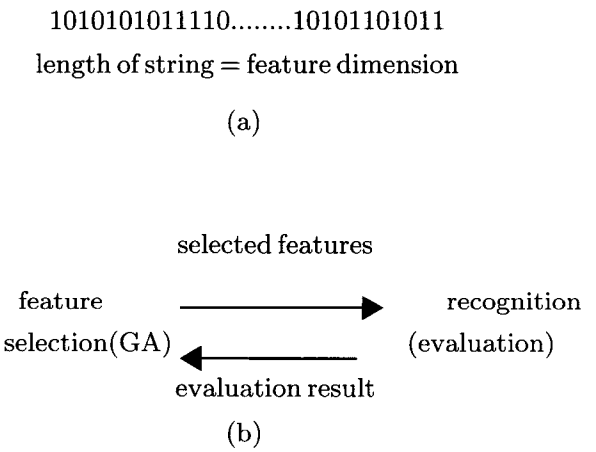

Fig. 5. (a) A chromosome (b) feature selection process

Distance between the centroid of a cluster and the feature is computed using Eqn. (1), and it is modified to Eqn. (2) in order to consider the features selected only:

$$Dist = \sum_{i=0}^{p} (x_i - y_i)^2 \tag{1}$$

$$Dist = \sum_{i=0}^{p} (x_i - y_i)^2 \quad \text{when } s_i = 1 \tag{2}$$

where, x_i is the i-th feature extracted from testing data, y_i is the i-th feature of any cluster in the codebook, p is the number of features before the selection is performed, and s_i indicates whether the i-th feature is selected (1) or not (0).

The computation complexity of the matching process in the recognition module, is reduced as much as the number of features reduced, (p - q), when q is the number of features selected. Reducing the feature dimension is important in terms of improving recognition speed because it has been known that the matching is the most time consuming process among the modules in a recognizer and the number of matching is proportional to the number of features being compared [11]. In addition, storage space required for the codebook is reduced (see Figure 6), where k is the number of clusters.

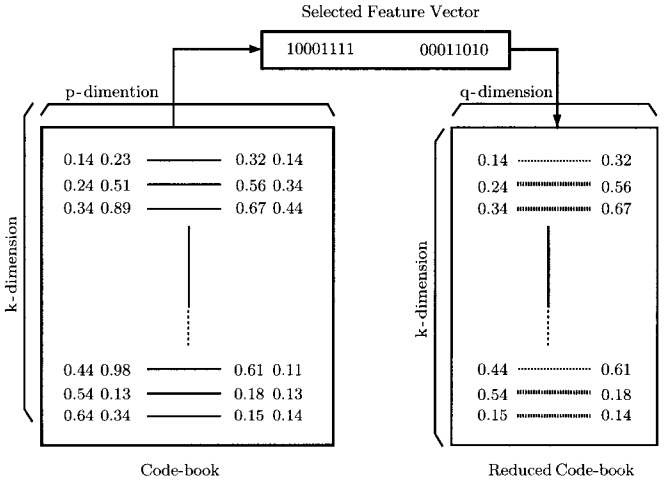

Fig. 6. Reducing Coodebook space

3.3.1 Introducing Variable Weights

As a result of the feature selection process, the number of features is reduced and slight performance degradation could be expected. Therefore, the feature selection method can be used for applications in which efficiency in terms of both speed and space is important in spite of some degree of degradation in recognition accuracy.

In this section, we use an approach to improve the recognition accuracy using the information collected while the feature selection is being performed. In the approach we do not aim for reducing the feature dimension, but we build a weight matrix by observing bits in the chromosomes. (see Figure 7) to show how to build the matrix using this approach. If, while the genetic algorithms are being performed, a bit in a chromosome is '1' (selected), the corresponding element in the matrix is increased by 1,

otherwise no change occurs. After the feature selection process is completed, the matrix is normalized and used in recognition module. The equation used for computing distance between features of an input image and the centroid of a cluster is changed from Eqn. (1) to Eqn. (3),

$$Dist = \sum_{i=0}^{p} (x_i - y_i)^2 \times \alpha_i \tag{3}$$

Where α_i is i-th normalized weight. Even though there is no reduction in the feature dimension in the approach, the weights reflect how often the corresponding features are selected during the feature selection process, and the weight is considered during the matching because more frequently selected features can be regarded as more important ones than others.

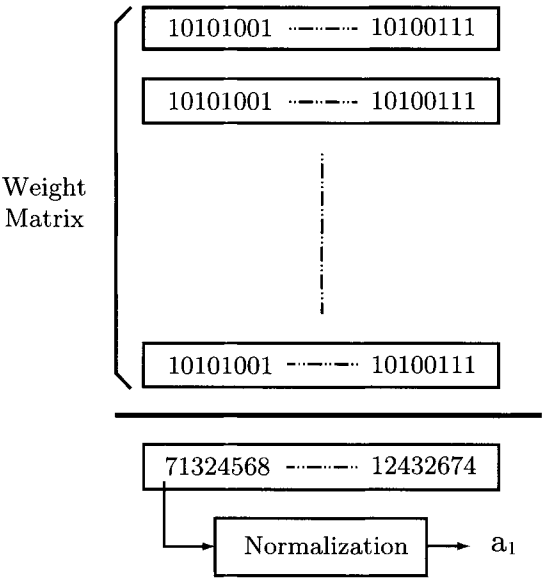

Fig. 7. Building the weight matrix

4 Experimental Results

Arabic Character handwritten Recognizer working with features extracted from image representation is used in the experiment. Handwritten Arabic characters recognizer conducted on the Arabic handwriting of 5 independent writers images, 28 images (represent 28 Arabic characters) are selected for codebook generation, evaluation of the feature selection algorithm, and performance measure of the recognizer with the selected features in the testing phase, respectively.

4.1 Feature Selection Using GA

Table 1 shows recognition accuracy of the Arabic Character handwritten recognizer with selected features using the GA. In the table, feature dimension of 74 represents the case that whole features are included without reduction.

According to the results in the table, significant reduction in the feature dimension, with trivial drop in the recognition accuracy, is observed. Parameters chosen for the GA are summarized in Table 2.

Table 1. Recognition performance with various number of selected features

Feature Dim.	74	74(Var. weight)
Recognition Rate	95%	95.3%

Tab. 2. Parameters used in the GA

Parameters	**Value**
Population Size	30
String length	74
Probability of Crossover	0.8
Probability of mutation	0.007
Selection method	Roulette wheel
Number of generations	1000

Also The last column in Table 1 represents the recognition rate with the variable weights obtained during the feature selection process. Variable weights are assigned to all features when the distance is computed using Eqn. (3). Comparing the result with the original, assigning different weight to each feature, depending on how often the feature is selected during the feature selection process, improves the recognition rate.

5. Compare Genetic Algorithm Approach with Previous Works

Recognition approach (genetic approach) using in this work compared with previous works according to the number of writers and the nature of training samples used by the system (see Table 3). The system is required to deal with as much writers as possible, training samples of each handwriting. The work presented by klassen [6] reports correct characters recognition using genetic Algorithm is an evolutionary machine learning strategy that uses cross-overs and mutations to create a program of mathematical operations on a data population to produce the "fittest" population. Genetic Programming had a positive example average of 92% for the training set, 77% for the validation set, and 72% for the test set.

Table 3. Comparison of performance and test conditions for recognition system in recent studies.

	Klassen,2001	Genetic algorithm approach
Training Rate	92%	95%
Validation rate	77%	89%
Test rate	72%	85%
Writers	25	25
Classes	15	28

The results obtained in this research were very promising and identification accuracy as high as 95% for the training set, 89% for the validation set, and 85.3% for the test set. classifier has shown a good performance.

6 Conclusion

In this paper, a feature selection method based on genetic algorithms is presented. The experimental results prove that feature selection using the GA reduces 30-50% of features with trivial drop in recognition accuracy, applying the variable weight obtained during the feature selection process improves the recognition accuracy, and the feature selection is working more effectively for features with larger dimension. The results obtained in this research were very promising and identification accuracy as high as 95% for the training set, 89% for the validation set, and 85.3% for the test set.

References

[1] Davis L. (1991), Handbook on Genetic Algorithms, Van Nostrand, Reinhold.
[2] Ho S.Y. & Huang H.L. (2001), Facial Modeling From an Uncalibrated Face Image Using a Coarse-to-Fine Genetic Algorithm, Pattern Recognition, 34(5).
[3] Kim G. & Govindaraju V. (1997), A Lexicon Driven Approach to Handwritten Word Recognition for Real-Time Applications, IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(4).
[4] Kim G. &, Kim S. (2000), Feature Selection Using Genetic Algorithms for Handwritten Character Recognition. In 7th IWFHR, Amsterdam-Netherlands.
[5] Kim B.S. & Song H.-J. (1998), An Efficient Preprocessing and Feature Extraction Method Based on Chain Code for Recognizing Handwritten Characters (in Korean), Journal of KISS(B): Software and Applications, 25(12).
[6] Klassen T. (2001), Towards Neural Network Recognition Of Handwritten Arabic Letters. Masters Thesis, Dalhousie University, Halifax, U.S.A.
[7] Madhvanath S., Kim G., & Govindaraju V. (1999), Chaincode Contour Processing for Handwritten Word Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(9).
[8] Mitchell M. (1996), An Introduction to Genetic Algorithms. MIT Press, Cambridge – MA.
[9] Ramesh V.E. & Murty N.(1999), Off-line Signature Verification Using Genetically Optimized Weighted Features, Pattern Recognition, 32(2).

[10] Shi D. (1998), Feature Selection for Handwritten Chinese Character Recognition Based on Genetic Algorithms, International Conference of Systems, Man and Cybernetics, 5.
[11] Yang J. & Honavar V. (1998), Feature Subset Selection Using a Genetic Algorithm, IEEE Intelligent Systems, 13(1).
[12] Wang Y. K. & Fan K. C. (1996), Applying Genetic Algorithms on Pattern Recognition: An Analysis and Survey. In Proceedings of ICPR'96.