A Survey on Genetic Algorithm based Classification Technique for Handwritten Character Recognition

Abhishek Phukan
Final year Engineering, CSE
Sikkim Manipal Institute of Technology
Majithar, East Sikkim

Mrinaljit Borah
Asst. Professor, MCA Deptt.
Jorhat Engineering College
Garmur, Jorhat

ABSTRACT

The paper depicts the progress achieved in the field of character recognition using genetic algorithm. Character recognition is a process in image processing where the characters fed into the system are identified and classified. The main focus of this paper is on the offline character recognition since very less work has been done in this field. The use of genetic algorithm is the basis of this paper and it focuses on the advantages of using a genetic algorithm and also a survey of the works that have been implemented so far.

General Terms

Classification Method, Character Recognition, Genetic Algorithm

Keyword

Character Recognition, Genetic Algorithm, Classification Phase

1. INTRODUCTION

Character recognition is the process of recognizing the text or documents that can be either fed into the computer via input devices or can be online. The character recognition systems can be classifiedon the basis how the text is entered. These are:

- Online character recognition
- Offline character recognition

The paper focuses mainly in the offline character recognition although the processes are very similar to a certain extent.

1.1 Optical Character Recognition

Online character recognition is the process of recognizing handwriting, recorded using a digitizer as a time sequence of the pen coordinates [1]. Digitizers are basically electronic tablets which can be either capacitive or resistive [pressure sensitive]. The digitizers send the pen coordinates as signals in regular intervals to the computer. The computer recognizes the motion and calculates the character. Online character recognition is however a lot easier than the offline handwriting recognition. Very less pre-processing is required and processes such as de-slanting and skew correction are not required. Also segmentation can be easily detected by pen lift information. Online character recognition is a real time process [1].

1.2 Offline Character Recognition

Offline handwritten character recognition as explained by S. Saha, N. Paul, S. Kunduand S.K. Das in their paper [2], is a process of automatic recognition of different characters from a document image and also provides a full alphanumeric recognition of printed or handwritten characters, text, numerals, letters and symbols into the computer process able format such as ASCII. Pulak Purkait on the [4] classified the stages of an OCR as follows:

- Pre-Processing
- Segmentation
- Feature Extraction
- Classification
- Post Processing

In the next section a brief overview of the stages in the OCR process will be discussed. A detailed study on the stages can be found in [4, 1, and 3]

2. STAGES

2.1 Preprocessing

Pre-processing aims at correcting the image before proceeding to the next phase prior to segmentation. Stages in preprocessing as given in [4] are:

- Binarization
- Noise reduction
- Stroke width Normalization
- Skew correction
- Slant removal

Binarization is the conversion of the image into binary image. Also known as thresholding, the image is converted from a colored image to dual colored image. The Binarization process can be done either locally or globally. Local Binarization segments the whole image into parts and a local threshold value is selected for each region. Global on the other hand takes into account a single threshold value for the entire document.

Noise reduction techniques are used to remove the noise from an image. Noises may occur due to technical hardware problem or even when an image is transferred to the system. Detailed study on noise has been done by Mrs. C. Mythili and Dr. V. Kavitha in their paper [5]. Filters are used for removing noises from an image. [5] Consists of details into the usage of linear, non-linear, wiener and fuzzy filters to remove noise and also a brief description about the different noise models.

Strokes in a character may be thin or thick. The idea behind stroke normalization is to correct the stroke width. Dilation can be used if there is a thin stroke and erosion if the width is thicker. [7] Describes the details of the morphological operations for stroke width normalization. Situations arise where a character may have more than one stroke and with different widths. The image needs to be subdivided to use the different operations.

Skew correction aims at correcting the alignment of the text document that is scanned.

Slant correction aims at correcting the slant of the text prior to the segmentation phase.

2.2 Segmentation

Segmentation is the process of separating the entire document into paragraphs and lines and finally to words and characters. Segmentation, as in [1] can be classified into:

- External segmentation
- Internal segmentation

External segmentation aims at isolating the writing units i.e. paragraphs, lines and words. On the other hand internal segmentation is the isolation of the letters from those segmented words. External segmentation is achieved through structural or functional analysis. Structural analysis segments the document into the components i.e. the paragraphs, rows and words. On the other hand functional analysis is used to label the functional parts of the documents i.e. title, abstract etc. A detailed study on segmentation can be studied from A. Cheung, M. Bennamoun, N. W. Bergmann in their [6]. An overview of the segmentation process is also given in [1] with classification of the different segmentation techniques. Character segmentation is classified into explicit, implicit and mixed strategy.

2.3 Feature Extraction

Features are a set of elements that categorize a particular object. Feature extraction is a method of extracting the features from the segmented characters to aid in the classification process. Another advantage of feature extraction is that it also reduces the storage space since only the extracted features are required to be stored and not the whole segmented character image which would consume a larger space. [7] Describes the various feature extraction methods that are listed below. This paper gives just an overview of the various methods. A detailed study can be made from Ms. SnehalDalal and Mrs. Lateshmalik in their paper [9].

- Horizontal and Vertical Histogram: based on the "matra" feature. A bangla or assamese script will have a longer horizontal black pixel run because of the matra. This information can be used to distinguish the other languages from English.
- Curvature information and local extreme of curvature: it is a useful in shape descriptor. The basic idea is to use the sequence of curve segments. The degree of curvature is calculated and this can be done by measuring the cumulative angle difference from all the sampling points. A minus indicates a clockwise curve.
- Topological features: topological features such as loops, endpoints, dots and junctions are also effective features that can be used for recognition. A detailed description about as to what all this features mean can be again studied from [9].
- Contour information: The whole image is traced from all sides and the transitions from black to white background are detected. Exterior contours are traced in a counter clockwise manner and clockwise for the interior pattern. A detailed study on the different contour patterns that can be used is again described in [9].

2.4 Classification

This is the phase where the features extracted are used to classify the character and to recognize the class to which it belongs. Methods used can be classified as [1]: template matching, statistical techniques, structural techniques and neural networks. Overviews of these techniques havebeen given below and will focus mainly on the use of genetic algorithms used for this purpose. The detailed study on the above methods can be read from [1].

- Template matching: based on the degree of similarity between the features extracted and the features of the characters stored in the database. There are several techniques of doing this. The simplest is the direct matching where the features are directly compared with the features in the database. Another method is to measure the dissimilarity with the contour edges. Relaxation matching uses the feature based description.
- Statistical techniques: it uses the statistical decision functions and optimality criteria. The features extracted are used to form a vector which is used to compare with the features of the original character. The approaches used here are the non-parametric and parametric recognition [1].
- Structural techniques: it uses the structural shape pattern of the objects. Methods include graphs and grammatical methods described in detail in [1].
- Neural networks: neural networks can process faster due to its parallel nature. It learns from the changes in the input signals. It consists of nodes, where the output of a node is fed to another node. The nodes interact among themselves and the final interpretation depends upon this interaction. Approaches for neural networks can be studied from [10].

2.5 Post-Processing

Since human handwriting differs from person to person hence in pre-processing stages certain information may be lost. This may cause irregular segmentation. The best method is to use a dictionary after the recognition stage [1]. Spell checkers can be used to find alternatives to the outputs generated in the classification stages. The use of lexicon words is also highly efficient in correcting these errors. The detailed version of the use of the lexicon recognizer is given in [1]

The next section discusses the achievements achieved in the classification and recognition stages using genetic algorithm. The first part gives a brief overview about the general idea of a genetic algorithm followed by advantages of its use and some of the works implemented based on this algorithm.

3. GENETIC ALGORITHM-OVERVIEW

Inspired by the processes in biological evolution, it is based on recombination, natural selection, inheritance, recombination and mutation. Random populations are generated for which the fitness function is calculated in order to find the fitness function that is the most optimal. If the first set does not contain a fitness function value that is not satisfied, then chromosomes recombine among themselves and mutate to find a collection of another random population and the process continues. The random population generated after recombination and selection is called the offspring. Genetic algorithm is basically a search technique that is faster than the classical ways of searching. They belong to evolutionary class of algorithm also known as EA. Each individual in a population of a genetic algorithm consists of properties (or traits) called the chromosome. These chromosomes can be mutated and altered. Chromosomes are generally represented in 0s and 1s. The population generation is an iterative process. Genetic operators (mutation and crossovers) are used to generate the next generation. Crossover is the process of combining the traits from two or more chromosomes in order to create a new chromosome. One point, two point are the most common crossover techniques. Other than these, techniques such as cut and splice and uniform crossover and half uniform crossover are also present. For selecting the chromosomes for crossover, again methods such as the Boltzmann selection or the tournament selection are use. There are also a lot of other methods but it is beyond the scope of this paper to go into the very detail of each and every selection method. To show for an e.g. a crossover operation, consider the chromosomes 111111 and 111111. Chromosomes are crossed over to create another chromosome say 111101. This is an example of a single point crossover method.

Mutation is to change the bits of the chromosomes to create a new chromosome. For e.g. chromosome strings can also be mutated, say 111101 to 011111 by changing just two bits and it gives a new chromosome. Different mutation types are available such as the bit string mutation, flip bit, boundary, non-uniform, uniform and Gaussian.

The algorithm terminates when:

- A solution has been found
- Time/resources drain out
- A fixed number of generation have been calculated
- There is no possibility of finding a better fitness function value.
- manually

The basic main advantages of using genetic algorithms are:

- Does not get stuck in a local optimum solution since it takes into account a population and not just one solution.
- It can be transferred to existing simulation and models.
- A large number of parameters are used and hence an accurate result can be found.
- The solution is not of fixed length.

4. GENETIC ALGORITHM BASED CLASSIFICATION AND RECOGNITION TECHNIQUES

VedguptSaraf and D. S. Rao in [8] have used genetic algorithm in character recognition of devnagari script. Without going into detail about the methods used in the pre-processing stages and the segmentation and feature extraction phases, the paper moves directly to the use of genetic algorithm in their work. The method that they have employed is to first normalize the image and then to find the number of peaks in that image. Next step is to determine a loop in the peaks and then find the complementary character. Following this is to determine the height and width of the loop along with the left and right connection. The peak's string is then sent to the Genetic algorithm. The condition that they have used at this point is that if the last peak in the sub word is found then proceed to the next condition which is that the last sub word in the word exist then again proceed to the next condition which is the last word. The algorithm terminates if the last word is found. The optimal condition is applied and if the condition is satisfied , than it is the best string and is considered as the solution. A flow chart for their entire process can be found in [8]. Through their work using genetic algorithm, they claim to have an accuracy of around 97%-98%, although there are pairs that they found confusing.

Pier Luca Lanzimet. al. [9], have used genetic algorithm for fast feature selection. The main advantage of their approach was that lesser processing time of CPU was required and the method was independent from a specific learning algorithm. They in their work have used the general representation scheme for the chromosomes, i.e. 1s represent the presence of a feature and 0s represent the absence. The change that they have made is that the fitness of the individuals is computed using inconsistency rate. The inconsistency rate is used to specify how accurately the reduced dataset of features represent the original dataset and how inconsistent a data becomes when only a subset of the data is taken into consideration. An example and the calculation method of the inconsistency rate can be found in [9]. They claim that using this method the algorithm is at least 10 times faster than a general genetic algorithm based feature selection. A discretization algorithm found in [22] was applied to a dataset. The inconsistency was used as a criteria and the genetic algorithm was applied to each dataset. The tree induction algorithm mentioned in [16] was applied to compare the subset of feature selected by the genetic algorithm and the original set of features. Accuracy of the original set and the reduced feature set have been averaged and compared using a two tail T-test mentioned in [17]. The basis for the algorithm is the GENESIS genetic algorithm. Parameters used are the default as in the GENESIS algorithm. A detail of the algorithm can be studied from [18]. The comparison revealed that although the reduced set contained a lesser number of features, it is as descriptive as the original set. References to related works in this field can be found in [9].

VedPrakashAgnihotri in [11] presents the use of genetic algorithm in character recognition of handwritten devanagari script. In the feature selection phase the image is divided into zones, and in total 54 features are extracted for use in the recognition phase. The zones algorithm used for feature extraction phase is given in [11]. Proceeding to the next phase, that is the recognition phase, he uses genetic algorithm to classify the characters. 3 bits per feature is used which results in a chromosome of 162 bits. A fitness function is used for the recognition by comparing the fitness value ofan unknown character with the dataset values. The solution has the highest fitness value. The fitness value is calculated using the given formula: fitness value=$\sum_{i=1}^{n} S - L$. Here S stands for chromosome bit string in database and L is the unknown chromosome bit string. He in his work has achieved a 85.78% match and 13.35% mismatch and the design for the interface is given in [11].

ChomptipPornpanomchai, VerachadWongsawangtham, SatheanJeungudomporn and NannaphatChatsumpun in their Paper [12]presents the use of genetic algorithm in character recognition of handwritten Thai characters. The information from the feature extraction phase is used for creating the chromosomes. The entire chromosome consists of 3 bits for the number of loops, 27 bits for the location of each loop, 17 bits for loop connected with a straight line and 12 bits for location of the lines. This chromosome is used for classifying the image. The equation that they have used for evaluating the fitness value is:$\sum_{i=1}^{66} |(S_i + 1.0) - (L_i \ + 1.0)| * w_i$. Here S is the chromosome string in dataset, L is the chromosome of testing character and w is the weight of the chromosome. The experiment was conducted on more than 111 characters. The training data set consisted of 8160 characters and testing data set consists of 840 characters. They achieved an accuracy of 88.17%. There was a 10.10% mismatch and 1.66% rejection. They achieved a speed of 0.4192 seconds per character. The experiment was also conducted for English characters but a poor percentage approximately 1.06% match was obtained.

E. K. Vellingiriraj and P. Balasubramanieet. al. [13] have done a similar work to the thai recognition. The change is only in the script. They have used ancient Tamil handwritten characters instead of thai characters. They have also implemented with the same strategy and the same fitness value function. The design for their system is given in details in [18]. Difficulties have been faced when cursive writing has been used.

ShashankMathur in [14] have also used genetic algorithm to implement character recognition. The process that he has used is to take the input string and to take a string of the current population. Both the strings are xor'ed. The dissimilarity coefficient from the resultant string is taken into consideration. The dissimilarity coefficient is calculated as number of 1s which shows the difference between the input and the resultant string. All the strings are taken into account and the coefficients are calculated. The mean value of all the coefficients is found to find chromosomes that are fit to generate the next population. The rest are discarded. Crossover is next applied between the chromosomes to find a chromosome with a lesser dissimilarity coefficient. Following is the mutation stage. The flowchart for the process is given in [14]. If error is above 10%, then the crossover stage is repeated. Else if the error is nearby 5%, the chromosome is mutated. These way new chromosomes are generated and through their work, they have obtained an overall efficiency of 79.16%.

5. CONCLUSION

The paper discusses as to what an offline character recognition system is. This paper discusses the various steps used in the process giving an overview of the details of the steps. Along with the overview, references to the detailed description of the process have been given below. Focus of this paper is mainly in the classification phase. The classification stage is explained in an overview again with the focus mainly on the genetic algorithm used for this process. In the genetic algorithm section, an overview of the genetic algorithm along with the details of the algorithm followed by its advantages has been discussed. The various methods of genetic algorithm that used in different papers have been explained along with the results and the efficiencies that they have achieved through their work.The work of genetic algorithms in the field of character recognition classification stage with an immense advancement in the Thai and Devanagiri script can be clearly noticed. Also the process has been used in other languages as well. The paper describes in detail the use of genetic operators and how they are used. There has been a massive advancement in this field and more work is still under research. The main advantage of using genetic algorithm is that it operates on a set of solutions rather than a single solution at a time. This prevents the algorithm from getting stuck in local minima.

6. REFERANCES

[1] Nafiz Arica, Fatos T. Yarman-Vural, "An overview of character recognition focused on off-line handwriting"

[2] SandeepSaha, Nabarag Paul, Sayam Kumar Das, SandipKundu, "optical character recognition using 40-point feature extraction and Neural Network"

[3] Gaurav Y. Tawde, Mrs. Jayashree M. Kundargi, " an overview of feature extraction techniques in OCR for indian scripts focused on offline handwriting "

[4] Pulak Pukait, "9th North-East Workshop on computational information processing"

[5] Mrs. C. Mythili, Dr. V. Kavitha, "efficient technique for color Image noise reduction"

[6] A. Cheung, M. Bennamoun, N.W. Bergmann, "an Arabic optical character recognition system using recognition based segmentation"

[7] Ms. SnehalDalal, Mrs. Latesh Malik, "A survey for feature extraction methods in handwritten script identification"

[8] VedguptSaraf, D.S. Rao, "Devnagiri script character recognition using genetic algorithm for better efficiency"

[9] Pier Luca Lanzi, Politecnico di Milano, "Fast feature selection with genetic algorithm: A filter approach"

[10] A. K. Jain, J. Mao, and K.M. Mohiuddin, "Artificial Neural Networks:A Tutorial", IEEE Computer, pp.31-44, 1996.

[11] VedPrakashAgnihotri, "offline handwritten Devanagiri script recognition"

[12] ChomtipPornpanomchai, VerachadWongsawangtham, SatheanpongJeungudomporn, NannaphatChatsumpun, " Thai Handwritten Recognition by gentic algorithm (THCRGA)"

[13] E. K. Vellingiriraj, P. Balasubramanie, "Recognition of ancient Tamil handwritten characters in palm manuscripts using genetic algorithm"

[14] ShashankMathur, "self-evolving character recognition using genetic operators"

[15] Lu H., Sung S. Y. and Lu Y.: On Preprocessing Data for EffectiveClassification. Workshop on Research Issue on Data Miningand Knowledge Discovery in Databases. (1996)

[16] Richeldi M., Rossotto M.: Supervised Quantization of ContinuousPredictor Variables. Seminars on New Techniques and Technologyfor Statistics. Bonn 20-22 November. (1995)

[17] Dietterich T. G.: Statistical Tests for Comparing SupervisedClassification Learning Algorithms. Tech. Report. Department ofComputer Science. Oregon State University. (1996)

[18] Grefenstette J. J.: Technical Report CS-83-11 ComputerScienceDept., Vanderbilt Univ.