A Survey on Recognition and Analysis of Handwritten Document

Remya Vinayakumar[#1], Vince Paul[#2]

#Computer Science And Engineering Department, Calicut University
Kerala, India

Abstract— Handwriting differs from person to person. Some may be legible while some others are difficult to read or understand. Hence this project aims at recognizing the handwritten text and understanding what it is with the help of a neural network and fuzzy logic. It involves segmentation, feature extraction and classification.Here the method used is Canny Edge Detection Algorithm and the Histogram Of Gradients for the feature extraction. The neural network is trained on to a 50 set samples for each of the 26 alphabets and 10 numbers for recognition. The fuzzification can be applied along with this inorder to get more accurate results by giving the questionnaires, ie, by giving the conditions to check if it satisfies a particular character which is to be determined. This would thus yield 80 percentage accuracy and reliability in recognition of the handwritten text.

Keywords— **Neural network, Histogram of Gradients, Canny algorithm, fuzzy logic**

I. INTRODUCTION

Artificial intelligence is one of the most interesting and fascinating area in the field of research. The word Intelligence is derived from the Latin word Intelligere which means to comprehend or perceive. Intelligence is the ability to perceive information and retain it as knowledge so as to apply towards the adaptive behavior within an environment. Generally we say that only human beings have the ability to think and understand. But it is also being observed in non-human animals and in plants. It is now being created artificially and that is how the term Artificial Intelligence has come into existence. Artificial intelligence is the intelligence associated with machines or software. It is a branch of computer science which studies how to create computers that can think like human beings by considering the environment in which it is. It is mainly being applied in the field of robotics, pattern recognition etc.

Handwriting can be defined as a person's writing using the pen or any possible instrument. It is unique for each person. Even identical twins have different styles of writing. Handwriting is like fingerprint. People may be able to copy it but can never write it in the same identical way. The circumstances in which a person grew up, his environment and the forces , the first language he learnt all contribute to shaping words and the uniqueness in the handwriting of a person.

Handwriting Recognition is now an engrossing field of research. Recognition of the handwritten document is not an easy task since each person has his/her own style of writing. Some might write very legibly whereas some others write in such a way that it is very difficult to understand what the person has written. This plays a very important role when it comes to the prescription written by the doctors , amount written of a cheque etc. Hence the understanding of the handwritten document is very important. In this project, we recognize the handwritten document using the Neural Network , Image Processing and the Fuzzy logic. It can be performed in either offline or online. Here in the latest method we do it offline because online is more time consuming and faces high difficultly in training a huge amount of data for character recognition. In offline handwritten character recognition, it focuses on documents that have been written on papers. The data is presented to the system as an image, requiring a segmentation of the writing from the image background before recognition can be done.

II. LITERATURE SURVEY

Researchers have been studying on the recognition of handwritten documents since 1960s. A lot of papers were published on this area since then. A survey on the handwritten document recognition by Mr. R. Plamondon and S. N. Srihari started in early 2000. In India, the Technology Development for Indian Languages (TDIL), under the Department of Information Technology,Government of India, were the ones who took the initiative in this research. Some practical applications of character recognition include recognition of vehicle number plates for identification and security concern information , to sort mails according to the postal code , in banks for processing checks, for easy-search of the scanned documents in database etc. Research papers on this topic have been published by the researchers since 1960s and the latest one in 2015.

In the 2011 Anshal[1],Plamondon[2],Arica[4] focuses especially on offline recognition of handwritten English words by first detecting individual characters. They considered two approaches for the word recognition. It was holistic approach and segmentation based approach. The holistic approach is used in recognition of limited size vocabulary where global features extracted from the entire word image are considered. As the size of the vocabulary increases, the complexity of holistic based algorithms also increases and correspondingly the recognition rate decreases rapidly. The segmentation based strategies, on the other hand, employ bottom up approaches, starting from the stroke or the character level and going towards producing a meaningful word. After segmentation the problem gets reduced to the recognition of simple isolated characters or strokes and hence the system can be employed for unlimited vocabulary.

Here they have incorporating artificial neural networks, hidden Markov models and statistical classifiers to extract segmentation rules based on numerical data. The segmentation is done using a heuristic algorithm. The segmentation is based on locating the minima or arcs between letters, common in handwritten cursive script. For this a histogram of vertical pixel densities is examined which may indicate the location of possible segmentation points in the word. A multilayer feedforward neural network trained with the back propagation algorithm is used in training the artificial neural network. Feature extraction is done using the Fourier transform for the recognition of numerals and Gradient features have been widely used in character recognition for machine and hand printed binary character images. But this method was not considered the best because it remained invariant to deformation in the character thus forming errors in the output. So, in 2012 researchers in this field came up with a better method to eliminate these limitations.

In the 2012 Neeta[3],Shabana [5] a Diagonal based feature extraction technique was used along with the neural networks for the recognition of handwritten documents. A feed forward artificial neural network is being used for character classification. segmentation of words from text is based on pixel density between upper and lower base line with multiple expert base validation for character recognition and classification.
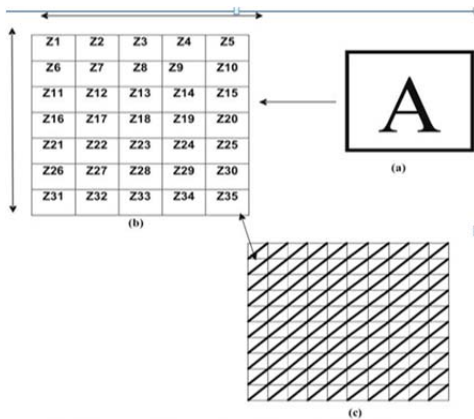

Fig. 1 (a). Normalized character image (b). Image divided in to 35 zones (c). Diagonal feature extraction

In the Diagonal based Feature extraction algorithm, every character image of size 90x60 pixels is divided into 54 equal zones, each of size 10x10 pixels. The features are extracted from each zone pixels by moving along the diagonals of its respective 10x10 pixels. Each zone has 19 diagonal lines and the foreground pixels present along each diagonal line is summed to get a single sub-feature, thus 19 sub-features are obtained from each zone. These 19 sub-feature values are averaged to form a single feature value and placed in the corresponding zone. This procedure is sequentially repeated for the all the zones. Finally, 54 features are extracted for each character. In addition, 9 and 6 features are obtained by averaging the values placed in zones row wise and column wise, respectively. As a result, every character is represented by

69 (i.e 54 + 15) features. The limitation of this approach was that these are several techniques used for segmentation and segmentation is done in parts. Also,for cursive handwriting full character segmentation is further more complicated as the characters are mostly closely connected making perfect segmentation point detection very difficult. So a more advanced technique had to be adopted for the recognition of handwritten document which was considered in the next years.

In 2013 Kandula[6], Takumi[7]two techniques were used to identify the handwritten character and they were the Active Character Detection and the Contour Algorithms. Contour tracing is one of many preprocessing techniques performed on digital images in order to extract information about their general shape. Once the contour of a given pattern is extracted, it's different characteristics will be examined and used as features which will later on be used in pattern classification.

Active Character Detection algorithm uses an active heuristic function. It uses small set of features that can be computed from a pattern image. Features are also computed from different sub-images. Sub-images are defined by the quad tree rule. The bounding box of a character image is divided into four rectangular regions. The center of mass of the contour is first compute. A vertical and horizontal line through the center of mass determine the four regions.

The two Contour Algorithms used here are the Square Tracing Algorithm and the Mooreneighbour algorithm. In square tracing algorithm, given a group of black pixels on a backround of white pixels,ie a grid, declare one of the black pixel as the start pixel. Now, imagine that we are searching on the start pixel . In order to extract the contour of the pattern,every time we find ourselves standing on a black pixel, turn left, and every time we find ourselves standing on a white pixel, turn right, until we encounter the start pixel again. The output will be a sequence of boundary pixels, ie, the contour.
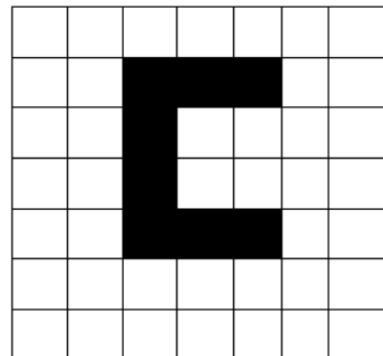

Fig. 2 Square tracing algorithm for character recognition

The Moore neighborhood of a pixel, P, is the set of 8 pixels which share a vertex or edge with that pixel. These pixels are namely pixels P1, P2, P3, P4, P5, P6, P7 and P8. Every time you hit a black pixel, P, backtrack i.e. go back to the white pixel you were previously standing on, then, go around pixel P in a clockwise direction, visiting each pixel in its Moore neighborhood, until you hit a black pixel. The algorithm terminates when the start pixel is visited for a second time.
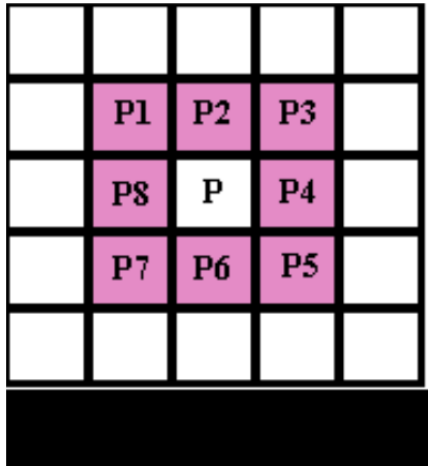
Fig. 3 Moore neighborhood algorithm for character recognition

These techniques were found to be good but however there may be few mismatches if there is completely different style of hand written character due to dissimilar region. The percentage of accuracy level obtained in the recognition was not very satisfactory. Hence improvements were made by adopting new algorithms and new technologies inorder to obtain better results.

The 2014 Disha[8], Sandeep[9],Prerna[10], were based on recognition of handwritten characters in the presence of noise. Noise has been considered as one of the major issue that degrades the performance of character recognition system. In order to overcome this limitation, the back propagation of artificial neural network is designed for the English character recognition in presence of noise. The recognition system is designed and tested in MATLAB under different noise levels.

At first a new graphical window interface is created. Input the characters to be recognized in the matrix form. Design the neural network with the desired parameters. Then set the noise level. Train the network with noise and without noise. Here supervised type of learning is used and finally the trained network is tested to see whether it has created a good balance

b/w memorization and generalization. There was another novel approach considered in the recognition of the handwritten text proposed by Disha et.el.[8] as shown in Fig (4, 4.1, 4.2, 4.3, 4.4) and the steps they included are as follows:

- Load the input image in IMG.
- Convert IMG to gray-scale image.
- Convert gray-scale IMG to binary
- Start a loop to isolate every line in IMG as LINE.
- Start a loop to isolate every word in LINE.
- Label characters in every word.
- Crop each and every character.
- Resize each and every character.
- Classify each character using correlation.
- Display the closest matched output

Fig. 4 Image input as reference (Source: A Novel Approach for Character Recognition, International Journal of Engineering Trends and Technology (IJETT) – Volume 10 Number 6 - Apr 2014, Disha Bhattacharjee#1, Deepti Tripathi#2, Rubi Debnath#3, Vivek Hanumante#4, Sahadev Roy*5 # Student *Assistant Professor, Department of Electronics and Communication Engineering National institute of Technology, Arunachal Pradesh Yupia, INDIA)

Fig. 4.1 Image after binarization (Source: A Novel Approach for Character Recognition, International Journal of Engineering Trends and Technology (IJETT) – Volume 10 Number 6 - Apr 2014, Disha Bhattacharjee#1, Deepti Tripathi#2, Rubi Debnath#3, Vivek Hanumante#4, Sahadev Roy*5 # Student *Assistant Professor, Department of Electronics and Communication Engineering National institute of Technology, Arunachal Pradesh Yupia, INDIA)

Fig. 4.2 Line wise segmenatation of the word "The"(Source: A Novel Approach for Character Recognition, International Journal of Engineering Trends and Technology (IJETT) – Volume 10 Number 6 - Apr 2014, Disha Bhattacharjee#1, Deepti Tripathi#2, Rubi Debnath#3, Vivek Hanumante#4, Sahadev Roy*5 # Student *Assistant Professor, Department of Electronics and Communication Engineering National institute of Technology, Arunachal Pradesh Yupia, INDIA)

Fig. 4.3 Character wise segmenatation ,isolating "T" (Source: A Novel Approach for Character Recognition, International Journal of Engineering Trends and Technology (IJETT) – Volume 10 Number 6 - Apr 2014, Disha Bhattacharjee#1, Deepti Tripathi#2, Rubi Debnath#3, Vivek Hanumante#4, Sahadev Roy*5 # Student *Assistant Professor, Department of Electronics and Communication Engineering National institute of Technology, Arunachal Pradesh Yupia, INDIA)
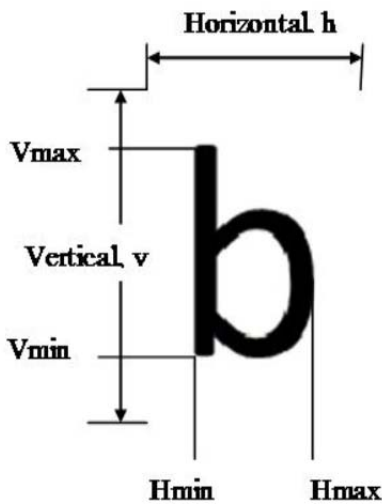

Fig. 4.4 Feature extraction (Source: A Novel Approach for Character Recognition, International Journal of Engineering Trends and Technology (IJETT) – Volume 10 Number 6 - Apr 2014, Disha Bhattacharjee#1, Deepti Tripathi#2, Rubi Debnath#3, Vivek Hanumante#4, Sahadev Roy*5 # Student *Assistant Professor, Department of Electronics and Communication Engineering National institute of Technology, Arunachal Pradesh Yupia, INDIA)

The deficiency of this method is that the detection of the white space between the letters cannot be recognized. The smaller case letters like i is recognized as 2 letters and this is one drawback of the method. The similar looking alphabets and letters like 9 and q cannot be recognized properly. Thus these problems can be eliminated in the further enhancement inorder to make the system more reliable.

Haswadi[11],kimura[12], in 2015 adopted a method to reduce the dimentionality of the feature space in which data has to be processed. Thus the feature reduction technique has been divide into two, the feature extraction and the feature selection. The two kinds of features are statistical features and the structural features. Here the major statistical features used for the character representation are zoning, crossing and distances and projections. Structural features are based on topological and geometrical properties of the character . In feature selection, the most relevant features to improve the classification accuracy must be searched. Feature selection is finding a subset of features which improve the recognition accuracy.This process has two main phases. First phase includes a search strategy to select one feature subset among all possible. Second phase includes a method for evaluating selected subsets with assigning a fitness value to them generally divided in two: filter and wrapper method. Approaches as search strategy for feature selection: heuristic and metaheuristic. Metaheuristic algorithm is an approach to solve the optimization problems and to find the best of all possible of solutions for example the Genetic algorithm .
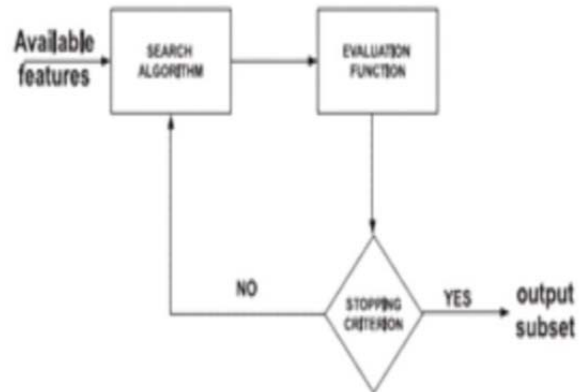

Fig. 7 Feature selection process

So far by 2015, it has been found that as a future work, the harmony search algorithm (HAS) ,a metaheuristic, inspired from a musician performance can be used for the feature extraction as it would help to find the better state of harmony. Recently , HSA has been applied to many engineering optimization problems including structural engineering , structural material, hydraulics , cost optimization and construction management , and structural vibration control. A feature selection method may also be proposed based on HAS. It does not require initial value settings for decision variables. It uses stochastic random searches and does not require derivation information. It can be easily adopted in various types of optimization problems.

III. CONCLUSIONS

Recognition of the handwritten text is a very interesting area of research. The paper gives a useful methods for the recognization of hand written characters. By using these algorithms character can be recognized and fuzzy logic can be used to reduce the time complexity..Character Recognition of varying font and style can also be done with the use of Neural Network during classification. It is further extended for hand-written text by incorporating artificial intelligence. A successful survey is done on the topic and also the best possible ways to implement the algorithms have been found out. Thus by the proper recognition of the character the better the text conversions become and more error free the document becomes.

REFERENCES

[1] Anshul Gupta,Manisha Srivastava, Chitralekha Mahanta "A.Offline Handwritten Character Recognition Using Neural Network"AI 2011 International Conference on computer applications and industrial electronics, April 2011.

[2] Plamondone ,S.Chaturved, N.R.Sondhiya and R.N.Titre, Izhikevich "Model Based Pattern Classifier for Hand Written Character Recognition "– A Review Analysis, International Conference on Electronic Systems, Signal Processing and Computing Technologies (ICESC), pp.346-349, 9-11 Jan. 2014.

[3] Neeta Nain Subhash Panwar,"Handwritten Text Recognition System Based onNeural Network", Journal of computer and information technology ,vol2,no-2,june 2012.

[4] Arica, N., and Yarman-Vural. F.T., "An Overview of Character Recognition Focused on Off-line Handwriting", IEEE Trans. On Systems, Man, and Cybernetics, Vol. 31. No. 2, 2001, pp. 216-233.

[5] Intelligent Systems for Off-Line Handwritten Character Recognition: A Review, International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 4, April 2012, Shabana Mehfuz,

[6] .Hand Written Character Detection by Using Fuzzy Logic Techniques, International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 3, March 2013 , Kandula Venkata Reddy1, D. Rajeswara Rao2, K. Rajesh3

[7] BoF meets HOG: Feature Extraction based on Histograms of Oriented p.d.f Gradients for Image Classification, 2013 IEEE Conference on Computer Vision and Pattern Recognition, Takumi Kobayashi

[8] A Novel Approach for Character Recognition, International Journal of Engineering Trends and Technology (IJETT) – Volume 10 Number 6 - Apr 2014, Disha Bhattacharjee#1, Deepti Tripathi#2, Rubi Debnath#3, Vivek Hanumante#4, Sahadev Roy*5 # Student *Assistant Professor, Department of Electronics and Communication Engineering National institute of Technology, Arunachal Pradesh Yupia, INDIA

[9] Handwritten Character Recognition System Using Neural Network, Volume 2, Issue 2, February 2014 International Journal of Advance Research in Computer Science and Management Studies, Amrita hirwanil, Sandeep Gonnade, MATS school IT, MATS university, Raipur

[10] Prerna Kakkar1, Umesh Dutta2. Dept. of ECE, Manav Rachna International University, Faridabad, India "A Novel Approach to Recognition of English Characters Using Artificial Neural Network", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 3, Issue 6, June 2014,

[11] Dewi Nasien, Muhammad Arif Mohamad, Haswadi Hassan, Habibollah Haron,A Review on Feature Extraction and Feature Selection for Handwritten Character Recognition, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 2, 2015

[12] Y. Kimura, A. Suzuki, K. Odaka, "Feature selection for character recognition using genetic algorithm," IEEE Fourth International Conference on Innovative Computing, Information and Control (ICICIC), Kaohsiung , pp. 401-404, Dec. 2009.

[13] A Review on Character Recognition Techniques, International Journal of Computer Applications (0975 – 8887) Volume 83 – No 7, December 2013, Richa Goswami, O.P. Sharma

[14] U.Bhattacharya , B.B.Chaudhri, Handwritten Numeric Databases Of Indian Script And Multi Stage Recognition Of Mixed Numericals , IEEE Transaction On Pattern Analysis And Machine Intelligence , 2009,VOL.3, No.3,pp 444-457