Language Specific Offline Character Recognition Using Neural Network Classifier

Preethi N[1],Dr T Gunasekaran[2]

[1]*PG scholar, Department of ECE,Vivekanandhe College of Engineering for Women*
[2]*HOD & Professor, Department of ECE,Vivekanandha College of Engineering for Women,*

***Abstract*— Character recognition is one of the most fascinating and challenging researches currently in the area image processing. The area of character recognition has been receiving considerable attention due to its versatile range of real-time application which includes reading aid for the blind, postal automation, processing of cheque and digitization of historical documents. Nowadays different methodologies for different language are in widespread use for character recognition. The work carried out here mainly concentrates on the recognition accuracy of the language. The proposed scheme consists of 4 stages: preprocessing, segmentation, feature extraction and recognition. Preprocessing techniques resulting for better accuracy such as noise removal, gray scale conversion and binarization are employed to enhance the character before classification. The neural network classifier methodology is implemented to recognize the character.**

***Index Terms*— character recognition, feature extraction, preprocessing, neural network classifier**

I. Introduction

In today's world character recognition plays important role to make everything automotive. Optical Character Recognition (OCR) is the process of converting handwritten or printed documented into machine readable form. In today's fast growing technology, digitization of the documents are important for future use which gives scope for the researches to perform Optical Character Recognition. to perform character recognition ,a document scanner with digitization of document software have been implemented in today's fast growing technology. OCR software allows you to scan a printed document and then convert the electronic text in word format. OCR receives its attention in the area of digitization in library and digitization of historical documents. In this paper an efficient approach for digitization of the Tamil character have been proposed. Tamil is one of the accepted language which is currently used by Tamil people. Character recognition in Tamil language is very less because defining of features for the Tamil language is difficult due to its complex character style and large data sets.

Recognition of Tamil handwritten scripts is complicated compared to other language scripts. Even the extraction of features and the segmentation of the individual characters is difficult in Tamil language. This vast difficulty in the area of character recognition motivates many researches to perform the recognition of character in Tamil language.

Character recognition is the process of designation the software that translates the handwritten/printed characters of document to machine accessible form. Businesses utilize character recognition software to keep track of paper invoices and other financial documents. Many different types of character recognition exist, with some having the ability to scan handwritten documents.OCR is generally classified into two types online character recognition and offline character recognition .The offline character recognition is most commonly in practice because online character recognition have its own difficulty in the area of character recognition.

II. Related Work

The off-line character recognition is an interesting and active area of research in the field of image processing. The work done by researches in printed characters is high compared to handwritten character because of certain common difficulties faced .The following are some of the effective character recognition papers in different languages. In the year 2011 Anshul Gupta, Manisha Srivastava, Chitralekha Mahanta used Fourier descriptor with magnitude for feature extraction and SVM classifier for recognition and attained accuracy of 86.66%. In the year 1997 Yuk Ying Chung, Man to Wong used Fourier descriptor and topological properties of feature extraction algorithm and used MLP with back propagation classifier and attained the accuracy of 96%.

A. Statistical Analysis Approach

The statistical approach uses the features of the different data sets measured from the input image. The features may be skeleton matrix vector obtained or may be zoning of input image. This approach is limited and cn be used only for the certain languages.

B. Structural Approach

The structural approach uses the structural features of the image such as number of horizontal, vertical or slanting line. This approach is called geometric feature extraction method. This is the effective approach for the feature extraction in large data sets and complex data features.

C. Neural Network Approach

Neural Network is one of the classification algorithm which provides the accurate range of classification between the different features of the character. The training phase in the character recognition should be more efficient so that the during testing any inputs can be given .

III. Types Of Optical Character Recognition

Generally OCR is classified into different types. The most common types are: offline and online character recognition.

ISSN: 2278 – 909X

A. Off line Character Recognition

Off-line handwriting recognition involves the automatic conversion of text in an image into letter codes which are usable within computer and text-processing applications. The data obtained by this form is regarded as a static representation of handwriting. Off-line handwriting recognition is comparatively difficult, as different people have different handwriting styles. And, as of today, OCR engines are primarily focused on machine printed text and ICR for hand "printed" (written in capital letters) text. There is no OCR/ICR engine that supports handwriting recognition as of today. The text produced by a person by writing with a pen/ pencil on a paper medium and which is then scanned into digital format using scanner is called Offline Handwritten Text.

B. Online Character Recognition

On-line handwriting recognition involves the automatic conversion of text as it is written on a special digitizer or PDA, where a sensor picks up the pen-tip movements as well as pen-up/pen-down switching. That kind of data is known as digital ink and can be regarded as a dynamic representation of handwriting. The obtained signal is converted into letter codes which are usable within computer and text-processing applications.

C. Intelligent Character Recognition

There is also another type of character recognition called Intelligent Character Recognition (ICR) .ICR is an advanced optical character recognition that allows fonts and different styles of handwriting to be learned by a computer during processing to improve accuracy and recognition levels. Most ICR software has a self-learning system referred to as a neural network, which automatically updates the recognition database for new handwriting patterns. It extends the usefulness of scanning devices for the purpose of document processing, from printed character recognition (a function of OCR) to hand-written matter recognition. Because this process is involved in recognizing hand writing, accuracy levels may, in some circumstances, not be very good but can achieve 97% accuracy rates in reading handwriting in structured forms. Often to achieve these high recognition rates several read engines are used within the software and each is given elective voting rights to determine the true reading of characters.

IV. Preprocessing

Preprocessing is defined as process of removing unwanted pixels present in the input image so that the segmentation will be more accurate. The process of preprocessing takes several steps according to the type of input taken and type output the user should get. The common preprocessing steps are

i. Noise removal
ii. Binarization
iii. Skew Correction

The noise is mainly due to optical scanning devices in the input, leads to poor system performance. These imperfection must be removed prior for recognition. Noise can also be included in the image during image acquisition. There are several types of noise such as Gaussian noise, Rayleigh noise, salt and pepper noise. The noise present in the input image can be removed by using filters. The input image of handwritten character in from the local data base is converted to gray scale format. Binarization is the important image processing step in which pixel value is separated into two groups, white as foreground and black as background. The goal of binarization is to remove unwanted information and thus protecting the useful information from the image. The skew correction is the process of aligned the document image to the base line of that document. The methodologies for the Tamil character recognition described in this paper is shown in figure 1.

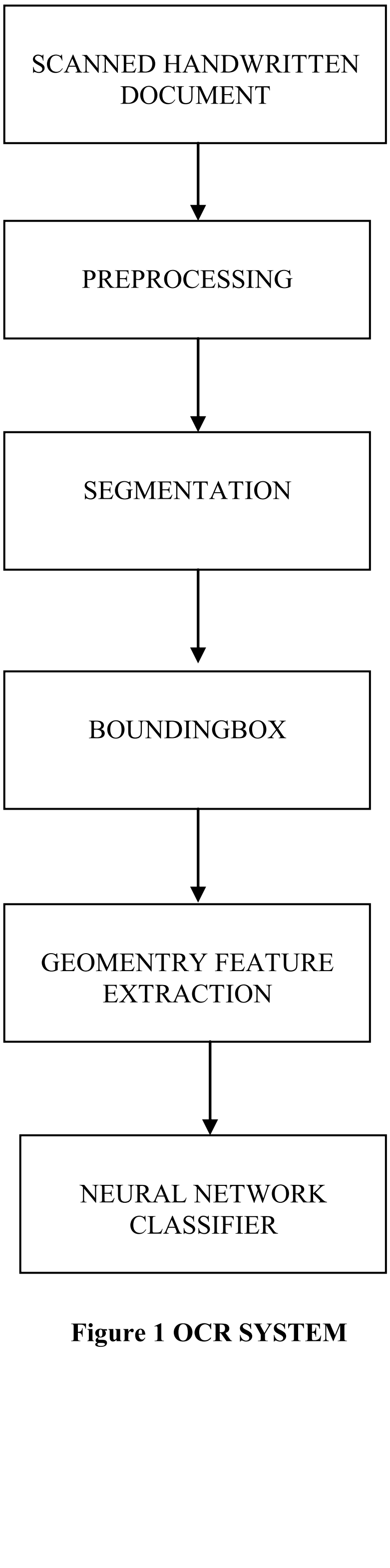
SCANNED HANDWRITTEN DOCUMENT
PREPROCESSING
SEGMENTATION
BOUNDINGBOX
GEOMENTRY FEATURE EXTRACTION
NEURAL NETWORK CLASSIFIER

Figure 1 OCR SYSTEM

V. SEGMENTATION

Script segmentation is an important task for character segmentation system. Segmentation is the process of splitting the document image into text lines and then splitting lines to word and then to individual character. This task is somewhat difficult for handwritten character due to various writing styles of different people. The process of segmentation mainly follows identifying the text lines in the page by considering the gap between each line. Then identifying the words in the individual segmented line whose gap is little higher than the gap between the letters. Finally each character is segmented by considering the pixel wise gap between the characters. Each character is input to the next stage of the classifier system for further processing. The algorithm used for segmentation is projection profile algorithm. The projection profile is horizontal projection profile for line segmentation and vertical segmentation for word and character segmentation. The segmented image is cropped and resized to constant size before the feature extraction process and this process of resizing is called as bounding box. The figure 2 shows the horizontal and vertical projection graph of the input image

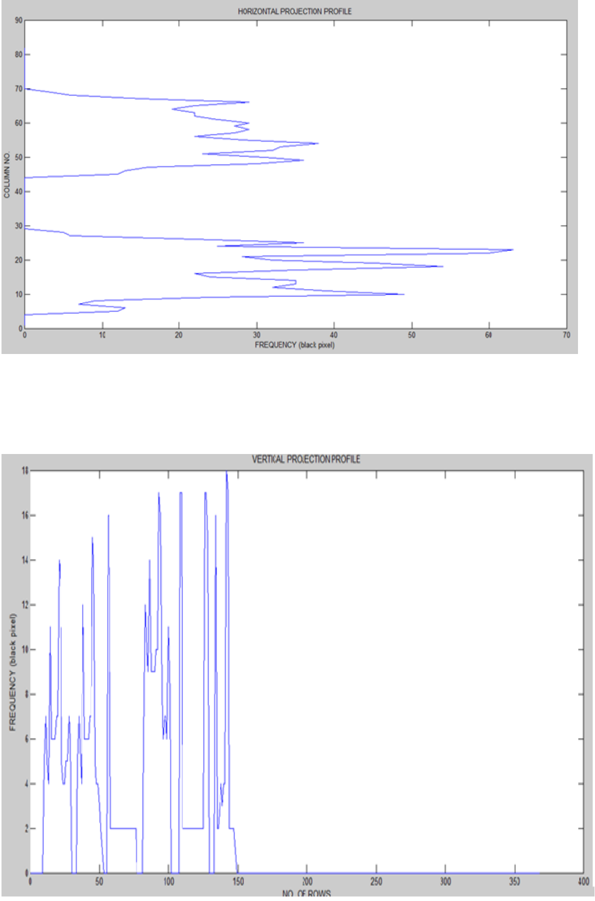

Figure 2 SEGMENTATION GRAPH

VI. FEATURE EXTRACTION

Feature extraction is the process of finding the unique shape of the character that defines the parameter of the character. Feature extraction methods are mainly of 3 types:

i. Statistical feature
ii. Global transformation
iii. Geometric feature

In this paper the geometric based feature extraction is proposed. The first step before applying the algorithm is universe of discourser. The universe of discourse is the fitting of entire character skeleton in the smallest matrix. The entire character skeleton is included in that matrix. The extra white space or the pixel is cropped out in order to avoid the large amount of execution time to execute the unwanted area. The second step is Zoning .The input image is divided into zones of several windows. The number of zones in the input image is determined by the user. After that the number of horizontal, vertical and slanting lines in each zone is determine. The nine features for the each zone is determined. If the number of zones in the image is 6 and 9 features are determined then total feature determined is 54 for each character. These 54 characters for the different character is different which helps to classify the unique character of the input character. The set of 54 features are given as an input to the neural network classifier stage.

Another type of feature extraction technique is based on '1' present in character skeleton. The segmented character will be of the form 0 s and 1 s, where 0 represents the background and 1 represents the presence of the character. Each individual character is analyzed to extract the specific feature corresponding to that particular character. The whole character is considered as a single input and the feature for individual character is extracted. Here the character binary image is converted into row first manner, so that all the characters are combined to form 248 X input image size. This input is given to neural network for training the individual characters. Many such characters inputs with various styles are considered for training, which greatly helps in improving the accuracy of the testing phase.

VII. NEURAL NETWORK CLASSIFIER

The classification is the processes of finding each character and assigning it to each character class. The classification technique is categorized into two techniques namely classical and soft computing technique .The various classical techniques are template matching, statistical techniques and structural techniques. Whereas the various soft computing techniques are Neural networks, Fuzzy logic, Evolutionary computing techniques. Here the neural network classifier is used because of its back-propagation network, which automatically assigns its weights according to which the desired output is obtained. The neural network classifier consists of three layers namely Input layer, Hidden layer and Output layer. The neural network has two phase namely the Training phase and Testing phase.

In training phase, the Neural Network is trained by using the Feature Vector which is extracted for the each independent character. During training the network updates its weights according to the input pattern. At the end of training phase, the neural network reaches a steady state where its weights do not change. The weights attain a final value such that any pattern similar to those samples which are presented in the training phase can be recognized.

In testing phase, the Feature Vector of the character is given as input to the Neural Network. The Neural Network processes these inputs and recognizes the character according to the similarities it attained during the training phase. During this phase no weight change takes place.

Each character pattern presented at the input layer will put a '1' at only one neuron of the output layer in which there is the highest confidence. A '0' is put at all the remaining neurons. For every character pattern at the input, the output is a 248 **x** 1 column matrix, in which a '1' is present at a single place only and the remaining 247 entries are all '0' e.g. character 'ah' results in (1, 0, 0 …, 0).

The sample neural network classifier is depicted in figure 3. The number of neurons in different stage is taken according to the kind of inputs considered for training.

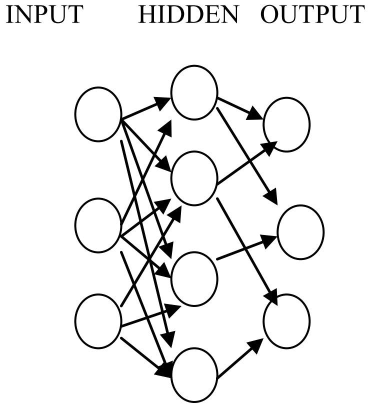

FIGURE 3 SIMPLE NEURAL NETWORK LAYER

VIII. CONCLUSION

Script identification is an important step for multi-lingual OCR development. The document was scanned and input was taken into system. Then the various Preprocessing techniques were done. Feature extraction is the important steps in character recognition system as each character has different feature that distinguish from each other characters. In this paper, an effective approach for Tamil character recognition is proposed. The different kinds of data sets for Tamil characters are considered. This work can be further improved with other different classifiers such as SVM, SOM, tree classifier, Fuzzy classifiers and comparing the accuracy.

.

REFERENCES

[1] Amit Choudharya, Rahul Rishi, Savita Ahlawat "Off-Line Handwritten Character Recognition using Features Extracted from Binarization Technique " AASRI Conference on Intelligent Systems and Control 2013 Elsevier

[2] Jafaar Al Abodi , Xue Li " An effective approach to offline Arabic handwriting Recognition " Compter and Electrical Engnineering 2014 Elsevier

[3] Jomy John ,Pramod K.V,Kannan Balakrishnan"unconstrained Handwritten Malayalam Character Recognition using wavelet transform and support vector machine classifier"International conference on communication Technology and system Design, 2012 Elsevier

[4] Choudhary, A., Rishi, R., and Ahlawat . S, "Handwritten Numeral Recognition Using Modified BP ANN Structure", Communication in Computer and Information Sciences (CCIS-133) 2010, Advanced Computing, Springer-Verlag, pp. 56-65

[5] S. Thadchanamoorthy, Umapada Pal et.all, "Tamil Handwritten city name database development and recognition of postal automation", 2013 12th International Conference on Document Analysis and Recognition

[6] Nisha Sharma, Tushar Patnaik, Bhupendra Kumar ,"Recognition for Handwritten English Letters: A Review" ,International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 7, January 2013

[7] Vijay Patil and Sanjay Shimpi ,"Handwritten English character recognition using neural network", Computer Science and Enginerring , 41 (2011) 5587-5591 Elixir

Authors Profile

Ms.N.PREETHI, Completed Under Graduate in Electronics and Instrumentation Engineering from Kongu Engineering College, Erode in 2013. She is currently pursuing Master of Engineering in Applied Electronics in Vivekanandha College of Engineering for Women. Her area of interest includes Digital Image Processing and Instrumentation & control.

Dr.T.GUNASEKAREN, received his Bachelor's Degree in Electronics and Communication Engineering from Kongu Engineering College, Erode in 2000. He completed his Master's Degree in Communication Engineering from Birla Institute of Technology and Science (BITS), Pilani, Rajasthan in 2003.He obtained his Doctorate in Micro strip Array Antennas , in Anna University, Chennai in 2013. Presently he is working as HOD in ECE department of Vivekananda College of Engineering for Women. His area of interest is Microwave Antenna and Digital Signal Processing.