RESEARCH ARTICLE OPEN ACCESS

Character Recognition (Devanagari Script)

Ankita Karia*, Sonali Sharma**, Reevon Rodrigues***, Maitreya Save****
*(Department of Computer Science, St. Francis Institute of Technology, Mumbai-103)

ABSTRACT
Character Recognition is has found major interest in field of research and practical application to analyze and study characters in different languages using image as their input. In this paper the user writes the Devanagari character using mouse as a plotter and then the corresponding character is saved in the form of image. This image is processed using Optical Character Recognition in which location, segmentation, pre-processing of image is done. Later Neural Networks is used to identify all the characters by the further process of OCR i.e. by using feature extraction and post-processing of image. This entire process is done using MATLAB.

Keywords - Character Recognition, Devanagari, feature extraction, MATLAB, Neural Networks, Optical Character Recognition, pre-processing, post-processing segmentation.

I. INTRODUCTION

The technology used nowadays to implement character recognition is by using Optical Character Recognition. It recognizes characters through optical mechanisms. The output of the OCR should ideally be the same as input in formatting. The process involves pre-processing of the image file and then acquisition of important knowledge of written text [1]. The current OCR systems use an image as an input and then that is used to be converted as an output to OCR.

We have seen many implementations of optical character recognition from the early seventies, where the input is taken in an image format using a scanner or any other image device. This image is then used to digitize the whole document so as to store as a virtual backup of the original document. Character recognition is majorly implemented for the English Script. We are proposing a design to create a character recognizer for 'Devanagari Script'.

This paper presents the idea of creating system for recognizing characters using Neural Networks. Here the user writes the character which is then processed using image processing and then the processed image is given and trained using neural networks. All of the processing and training is done using MATLAB coding.

II. PREVIOUS WORK

The origins of character recognition can actually be found back in 1870. This was the year that C.R.Carey of Boston Massachusetts invented the retina scanner which was an image transmission system using a mosaic of photocells. The first true OCR reading machine was installed at Reader's Digest in 1954. This equipment was used to convert typewritten sales reports into punched cards for input to the computer [2].

In recent years attempts have been made to develop text recognition systems in almost all the languages and scripts of the world. Devnagari script, the attempts were made as early as 1977 in a research report on handwritten Devnagari characters with a limited success [3]. Devanagari script is written by joining the characters, even merging characters to have compound characters and also putting "matras" in various forms, this makes the interpretation or recognition extremely difficult.

Arora S., Bhattacharjee D., Nasipuri M., Basu D. K., Kundu M. and L. Malik, proposed a Devnagari Character Recognition system which uses different feature extraction and recognition algorithms. Their proposed system was tested on approximately 1500 handwritten Devnagari character database collected from different people. It was observed that the proposed system achieved 98.16% recognition rates. Singh Raghuraj, Yadav C. S., Verma Prabhat and Vibhash Yadav have presented a scheme to develop complete OCR system for different five fonts and sizes of printed Devanagari characters and the accuracy is found to be quite high. Dongre Vikas J and Vijay H Mankar have proposed a simple histogram based approach to segment Devanagari document with accuracy of 100% in case of line segmentation and 91% in case of word segmentation [3].

III. WHAT IS OCR?

OCR is acronym for Optical Character Recognition. It is a concept that can recognize characters using optical mechanisms. Optical character recognition belongs to the family of techniques performing automatic identification. OCR can be used in recognizing both handwritten as well as printed characters. Optical Character Recognition deals with the problem of recognizing optically processed Characters. Optical recognition is

performed off-line after the writing or printing has been completed, as opposed to on-line recognition where the computer recognizes the characters as they are drawn. Both hand printed and printed characters may be recognized, but the performance is directly dependent upon the quality of the input documents [2].

OCR is a field of research in pattern recognition, artificial intelligence and machine vision. There are different approaches that can be used for designing OCR systems like Matrix Matching, Fuzzy Logic, Feature Extraction, Neural Networks etc [1].

Optical Character Recognition systems is transforming large amount of documents, either printed alphabet or handwritten into machine encoded text without any transformation, noise, resolution variations and other factors.

IV. ARCHITECTURAL DESIGN

The basic OCR system consists of 5 stages as shown in figure 1,

A. Optical Scanning
B. Location & Segmentation
C. Pre-Processing
D. Feature Extraction
E. Post-Processing

Let us consider each stage separately.

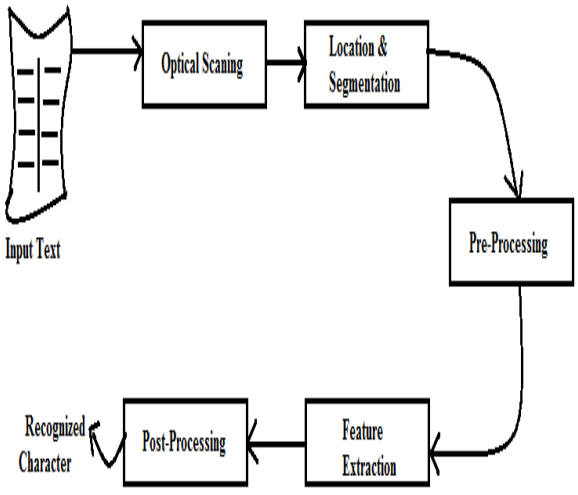

Figure1: Stages of OCR

A. Optical Scanning

The main idea of using optical scanning is convert the text that is either handwritten or printed into a digitized format which can be ultimately used by the computers for their further processing.

Nowadays scanning can be done using scanners. As in figure 2, the user is using a mouse to write the character in the given space which will be captured and stored as image. The image is stored in .tiff format which is then used for further stages in the system for processing and recognizing characters.

Figure2: Optically Scanned Image

B. Location & Segmentation

As the name implies location is the process of identifying the start of the first pixel. i.e. to locate where the character starts. This is done by scanning the image from left to right and from top to bottom.

Since the paper is about Devanagari Script, involves single characters with 'Matras' or compound characters with 'Hallant'. Location also helps in identifying the regions for 'Matras' i.e. certain characters might have "Matras' above the main character or below the character.

Segmentation is needed to distinguish the actual character and 'Matras' associated with it. Here as shown in figure 3, we have divided the entire region into three main zones [3]. Upper zones for 'Matras', middle zone for actual character and lower zone for other 'Matras' and 'Hallant'. It isolates the individual characters that are preset in the entire word. Functions like regionprops, bwlabel, imcrop are used in Matlab for segmenting the image

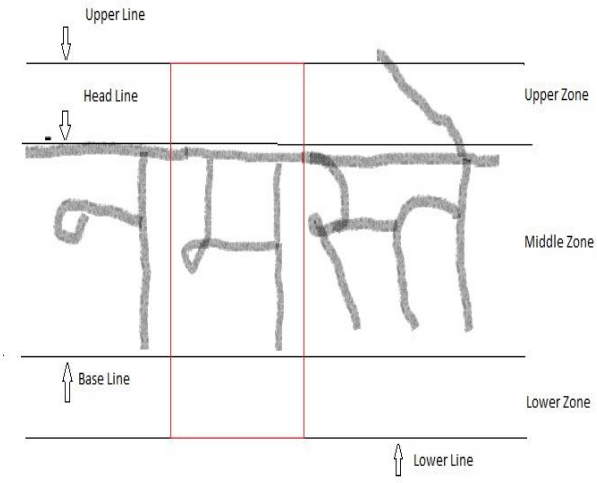

Figure3: Location & Segmentation of Image

C. Pre-Processing

In this step series of operations are performed on the scanned segmented images. Various tasks are performed in image processing stage.

As shown in figure 4, binarization is done that converts gray scale into binary image i.e. an image with zeros and ones this is done by using global threshold technique [4][6][7]. Edge detection is done using Sobel technique. Sobel is used because of its better results as compared to other techniques like canny or Prewitt. To enhance the quality of image dilation and filling of holes is done using simple MATLAB functions like imdilate and imfill. Also functions like noise reduction are used to remove disconnected line, bumps, gaps etc. Image smoothening includes normalization of image. Whereas thinning of image can be done to reduce the width of the character.

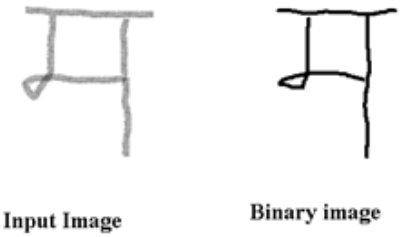

Figure4: Pre-Processed Image

In this paper we require the size of the image to be predefined i.e. 5x7 which is done by using function resize. This resized image is then used for extracting features for further process.

D. Feature Extraction

Feature extraction is the main part of the entire system. It defines characteristics of each character by absence or presence of key features. The techniques for extraction of such features are often divided into three main groups, where the features are found from

- The distribution of points.
- Transformations and series expansions.
- Structural analysis.

The different groups of features may be evaluated according to their sensitivity to noise and deformation and the ease of implementation and use. These key features can be defined by user and they can be height, width, density, loops, lines etc.

Feature extraction can be done by matrix matching i.e. by taking the input image matrix and den directly matching with matrix set of defined class. This is the simple technique which is used in hardware. Whereas other technique is to extract features based on statistical distribution of points i.e. the zone in which the pixels are populated. In structural analysis the no of loops or lines is taken as features i.e. the ones which give the geometrical and topological structure of symbol.

In this paper we are using neural networks to extract the features out of every character. In neural networks there are again various methods that can be used to train the system like Genetic Algorithm, Multi layer feed forward network, Artificial Neural Fuzzy Inference Model etc.

For error back propagation method trainbr command is used in Matlab to train the network. Considering a back-propagation network, this network is composed of several layers of interconnected elements. A feature vector enters the network at the input layer. Each element of the layer computes a weighted sum of its input and transforms it in to an output by a nonlinear function. During training the weights at each connection are adjusted until a desired output is obtained. This type of network is known as supervised learning method where the user is aware about the desired output and based on input and desired output the error is calculated which is again given back to input and trained again. This helps in minimizing the error after every iteration and this process stops when the input and output vector both have same values.

While training the network:

Number of neurons in input layer: 35

Number of hidden layer: 2

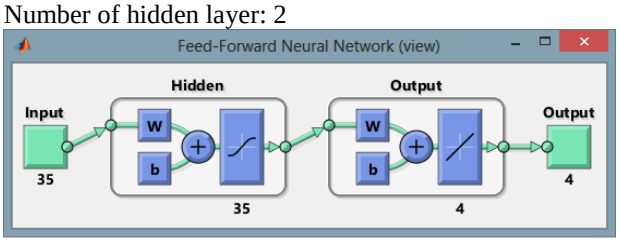

Figure5: Network Representation

E. Post Processing

This is the last stage of the OCR process. After the character is recognized it is a set of individual characters. These characters individually might not contain enough information i.e. these single characters might belong to the same string with each other making up words in that particular string. This process of performing association of these single symbols into string is referred as grouping. This grouping of symbols is based on the location of the individual characters i.e. how close the characters are from each other and which character comes next in line with the previous one.

Finally figure 6, shows the sequence of recognized characters in the string and display to the user in the standard format [8] [9].

Figure 6: Post Processed Image.

V. RESULTS

The current recognition system is been implemented in MATLAB R2014a. The scanned image is stored in the database and stored in .tiff format. The structure of neural network includes an input layer with each input to be resized to 5x7 i.e. 35 inputs, one hidden layer with 10 neurons and output later with 49 neurons. Neural network has been trained using known dataset. After training the network, the recognition system was tested using several unknown dataset and the results obtained are presented in this section. Figure 7 shows neural network training state. While figure 8 shows the best Training Performance graph and thus the desired performance goal has been achieved in 53 epochs.

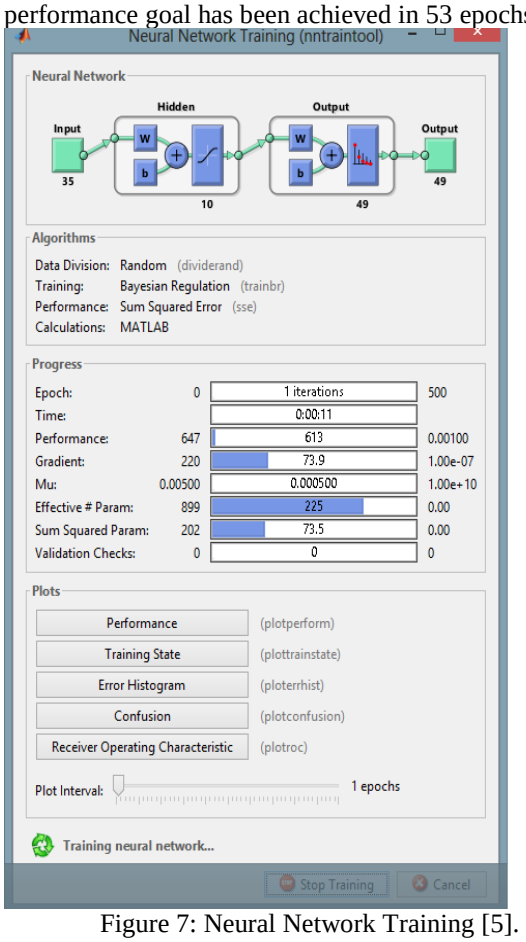

Figure 7: Neural Network Training [5].

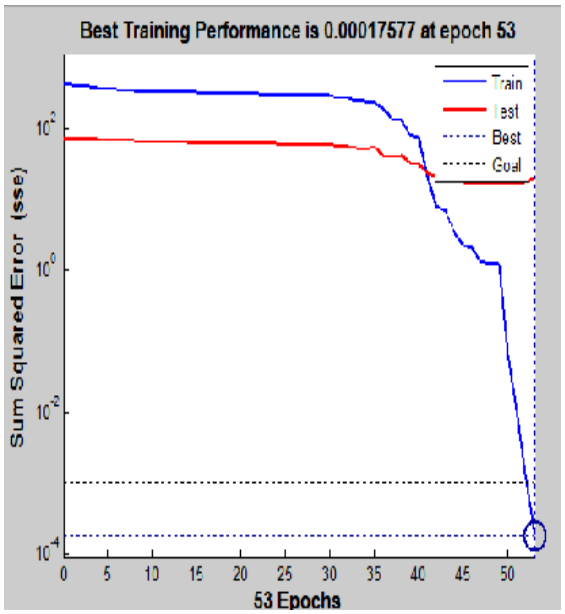

Figure 8: Training Performance of Network [5].

VI. CONCLUSION

Character recognition techniques associate with the symbolic identity with image of character. It is commonly referred as the OCR i.e. Optical Character Recognition that deals with recognition of optically processed characters. Since OCR techniques based on neural network provide more accurate and précised results than the other techniques.

In OCR system input characters will be given in digitized format. Each character will be then located and segmented and the resulting character will be then fed to the pre-processor for noise reduction and normalization. Sobel technique is used to reduce the noise and give a proper normalized image. The normalized image is then given to feature extraction block. In feature extraction the characters will be uniquely identified using neural networks and then the final output will be displayed to the user after post-processing it. Thus figure 9 shows how the written character is then recognized and displayed in the standard format in a notepad file.

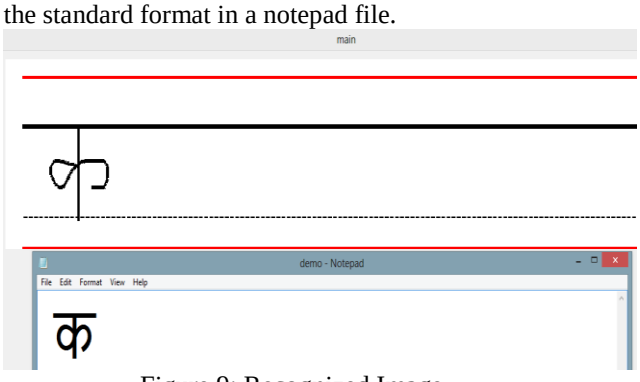

Figure 9: Recognized Image

In the training set small set of Devanagari characters with the help of back propagation neural

network is trained and accordingly it is tested. The characters that are difficult to identify are further tested and analyzed with different training set to get a better accuracy to the given character [10]. Thus as the number of training set increases, the accuracy to recognize characters increases. And this approach has been proven effective and accurate in identifying characters in digital format.

REFERENCES:

[1] Sukhpreet Singh," *Optical Character Recognition Techniques: A Survey*", 2013..

[2] Line Eikvil,"*Optical Character Recognition*", 1993..

[3] M.s. Neha Sahu, Mr.Nitin Kali Raman, "*An Efficient Handwritten Devnagari Character Recognition System Using Neural Network*", *978-I-4673-5090-7/13/$31.00 ©2013 IEEE,*2013.

[4] B.Indira, M.Shalini 1, M.V. Ramana Murthy, Mahaboob Sharief Shaik, "*Classification and Recognition of Printed Hindi Characters Using Artificial Neural Networks*", *I.J. Image, Graphics and Signal Processing*, 2012

[5] http://in.mathworks.com/products/neural-network

[6] Rinku Patel1, Avani Dave, "*A Survey Paper on Character Recognition*", *IJSRD - International Journal for Scientific Research & Development| Vol. 1, Issue* 9, 2013 | *ISSN (online):* 2321-0613.

[7] Rohit Verma and Dr. Jahid Ali, "*A-Survey of Feature Extraction and Classification Techniques in OCR Systems*" *International Journal of Computer Applications & Information Technology, Vol. I, Issue III,* November 2012 *(ISSN: 2278-7720)*

[8] Ankit Sharma, Dipti R Chaudhary, "*Character Recognition Using Neural Network*", *International Journal of Engineering Trends and Technology (IJETT) - Volume4Issue4-* April 2013.

[9] Prof. Mukund R. Joshi, Vrushali V. Sabale, "*Recognition of Devanagari Printed Text Using Neural Network and Genetic Algorithm*", *International Journal of Advance Research in Computer Science and Management Studies Volume 3, Issue 2,* February 2015 pg. 279-282.

[10] Raghuraj Singh, C. S. Yadav, Prabhat Verma, Vibhash Yadav, "*Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network*" *International Journal of Computer Science and Communication, Vol. 1, No. 1* January-June 2010,pp.91-95.