Oriya Character Recognition using Neural Networks

Soumya Mishra[1], Debashish Nanda[2] , Sanghamitra Mohanty [3]

P.G.Department of CSA

Utkal University

soumyalitun@gmail.com [1], emailtodebasish@gmail.com [2], sangham1@rediffmail.com[3]

***Abstract*—The biggest challenge in the field of image processing is to recognize documents both in printed and handwritten format. Optical Character Recognition (OCR) is a type of document image analysis where scanned digital image that contains either machine printed or handwritten script input into an OCR software engine and translating it into an editable machine readable digital text format. Development of OCRs for Indian script is an active area of research today. We are making an attempt to develop the OCR system for Oriya language, which is the official language of Orissa. Oriya language present great challenges to an OCR designer due to the large number of letters in the alphabet, the sophisticated ways in which they combine, and the complicated graphemes they result in. In this paper, we argue that a number of automatic and semi-automatic tools can ease the development of recognizers for new font styles and new scripts. We discuss briefly and show how they have helped build new OCRs for the purpose of recognizing Oriya script. We have used the Back propagation Neural Network for efficient recognition where the errors were corrected through back propagation and rectified neuron values were transmitted by feed-forward method in the neural network of multiple layers, i.e. the input layer, the output layer and the middle layer or hidden layers.**

Keywords-Optical Character Recognition, Artificial Neural Network, Histogram, Binarization, Back Propagation.

I. INTRODUCTION

Optical character recognition (OCR) systems have been under research for few decades. But still it remains a highly challenging task to implement an OCR that works under all possible conditions and gives highly accurate results. Optical character recognition of Oriya characters is a challenging field. Being an Indian language Oriya is ornamented with fabulous literature and sophisticated grammar. Identification of Oriya scripts though is not an easy task; still many fruitful attempts have been made. This proposed approach is very much different from the previous approaches. Artificial Neural Network technique is totally based on feature matching. Therefore various geometric features and character features are extracted from segmented text. Features are collectively evaluated and a combined score is obtained.

OCR works by first pre-processing the digital page image into its smallest component parts with layout analysis to find text blocks, sentence/line blocks, word blocks character blocks. The character blocks are then further broken down into components parts, pattern recognized and compared to the OCR engines large dictionary of characters from various fonts and languages. The approach used here is Artificial Feed Forward Neural Network, which fuses different extracted geometric such as height, width, number of pixels in columns and rows and textual features such as histogram, centroid etc to get a valid epoch or score. If the score is within the desired range then the character is recognized, else it suggests the user to back propagate and balance the input parameters by adding or subtracting weight values and bias values with overall score.

Verification of provided information and validation of data item depends on a threshold value previously imposed on the knowledge base system. Proposed technique is efficient under idle circumstances and limited corpus size.

II. MOTIVATION

Orissa has a rich heritage of manuscripts and novels, which are need to be preserved in Oriya language and Oriya scripts. The invaluable ancient works are in the process of being lost to termites in libraries in India, due to the lack of proper OCR systems. Automatic recognition of alphabetic characters through computers is a basic need for text recognition in different languages .OCR system for different foreign languages like Japanese, Chinese and Korean are already been developed. For Indian Languages attempts are made for Devanagari, Gurmukhi, Marathi, Bengali, Telugu and Tamil. We are making an attempt to develop the OCR system for Oriya language. In order to read those scripting we have made a sincere attempt to develop Oriya OCR. Therefore, our attempt here is a humble effort towards this noble goal of developing an efficient OCR system, in public domain, for Oriya characters.

III. METHODOLOGY

There are two basic methods used for OCR: Matrix matching and feature extraction. Of the two ways to recognize

characters, matrix matching is the simpler and more common. But still we have used Feature Extraction to make the product more robust and accurate. Feature Extraction is OCR without strict matching to prescribed templates. This method is much more versatile than matrix matching. Matrix matching works best when the OCR encounters a limited repertoire of type styles, with little or no variation within each style. Where the characters are less predictable, feature, or topographical analysis is superior.

The Process of Optical Character Recognition of the document image mainly involves six phases:

1. Acquisition of Grayscale Image

2. Digitization/Binarization

3. Thinning and Edge Detection

4. Feature Extraction

Such as: Dimension Measurements

Storage Details

Centroid and Histogram

Pixel Counts in Total, Rows, and Columns.

5. Feed Forward Artificial Neural Network based Matching.

6. Recognition of Character based on matching score.

The acquisition phase uses a scanner or digital camera that catches photocopy of the text document as an image. The scanned image must be a grayscale image or binary image, where

binary image is a contrast stretched grayscale image. That grayscale image is then undergoes digitization. In digitization a rectangular matrix of 0s and 1s are formed from the image.

Where 0-black and 1-white and all RGB values are converged into 0s and 1s.The matrix of dots represents two-dimensional array of bits. Digitization is also called binarization as it converts grayscale image into binary image using adaptive threshold·

After that thinning is done to narrow edges in the image and along with eliminate the unwanted noise pixels. Thinning is a morphological operation that is used to remove selected foreground pixels from binary images. Edge detection is the process of identifying points in a digital image at which the image brightness changes sharply or more formally has discontinuities.

Feature Extraction such as dimension measurement like height and width of the image, storage details like file size and vector size. Centroid of the character image in horizontal and vertical coordinates, Histogram of luminance for each gray level value. Lastly pixel counts in the image in total and in each row and columns separately are calculated. After all the relevant features are extracted the corresponding numeric values are stored in the database.

Feed Forward Neural Network approach is used to combine all the unique features, which are taken as inputs, one hidden layer is used to integrate and collaborate similar features and if required adjust the inputs by adding or subtracting weight values, finally one output layer is used to find the overall matching score of the network. If the score is within the predefined range then the character is recognized else the system is trained again.

This OCR system is developed to recognize individual Oriya alphabets and provides a base to build complete OCR system for composite characters and entire page readability. Along with it aims to embed text-to-speech system and spellchecker to give better understanding and error correction. The Corpus size taken here is standardized to make the concerned system work smooth but intend to be enhanced in future.

Features to be incorporated in future:

There are ten structural features can be extracted from the 16x16 pixel matrix which are given below:

1. Upper Part Circular

2. A vertical line on the right most part

3. Holes

4. Horizontal Run code

5. Vertical Run code

6. Number of holes

7. Position of hole

IV. ARTIFICIAL NEURAL NETWORK

An Artificial Neural Network (ANN), usually called "neural network" (NN), is a mathematical model or

computational model that tries to simulate the structure and/or functional aspects of biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. An Artificial Neural Network usually consists of one input layer,more than one output layers,except the last output layer all other intermidiate layers are called hidden layer.Outcome of input layer is fed as input to the hidden layers and outcome of hiden layers are fed as input to the final output layer.

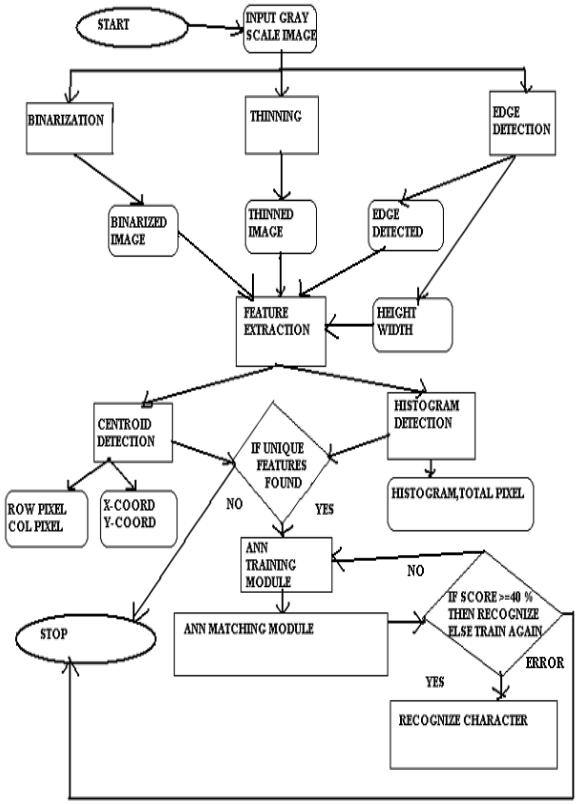

Fig-01 (Process Flow Chart)

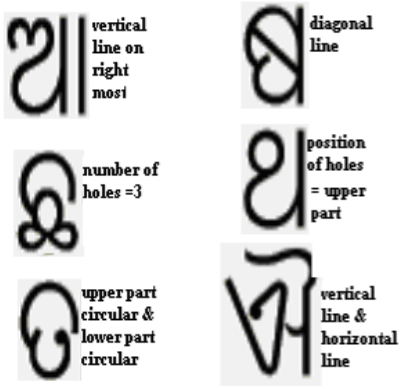

Fig-02 (Structural Features of Oriya Character)

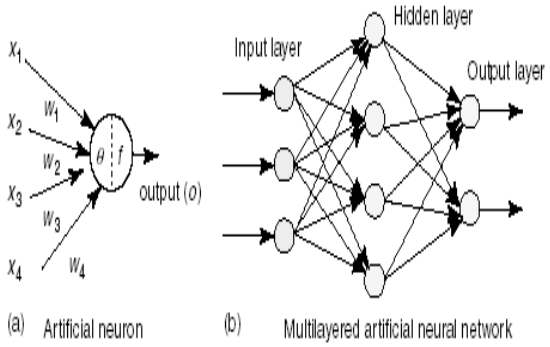

Fig-03 (a)Artificial Neural Network (b)Multilayer ANN

Algorithm of Feed Forward Neural Network:

1. Form network according to the specified topology parameters
2. Initialize weights with random values within the specified ± weight bias value.
3. Load trainer set files (both input image and desired output text)
4. Analyze input image and map all detected symbols into linear arrays
5. Read desired output text from file and converts each character to a binary Unicode value to store separately
6. For each character:
   a. Calculate the output of the feed forward network
   b. Compare with the desired output corresponding to the symbol and compute error
   c. Back propagate error across each link to adjust the weights
7. Move to the next character and repeat step 6 until all characters are visited
8. Compute the average error of all characters
9. Repeat steps 6 and 8 until the specified number of epochs
   a. Is error threshold reached? If so abort iteration
   b. If not continue iteration

V. CONCLUSION

Due to no availability of 100% noise freed digitized image and presence of similar shaped characters the accuracy rate is affected. To tackle this type of problem the system has been integrated to spell checker with the help of dictionary and a huge corpus. During scanning the document may be slanted leading to problem at the time of line extraction. Noise arises due to so many reasons namely old document, low paper quality, and dust particles on the surface of scanner, low quality ink and low quality printing machines.

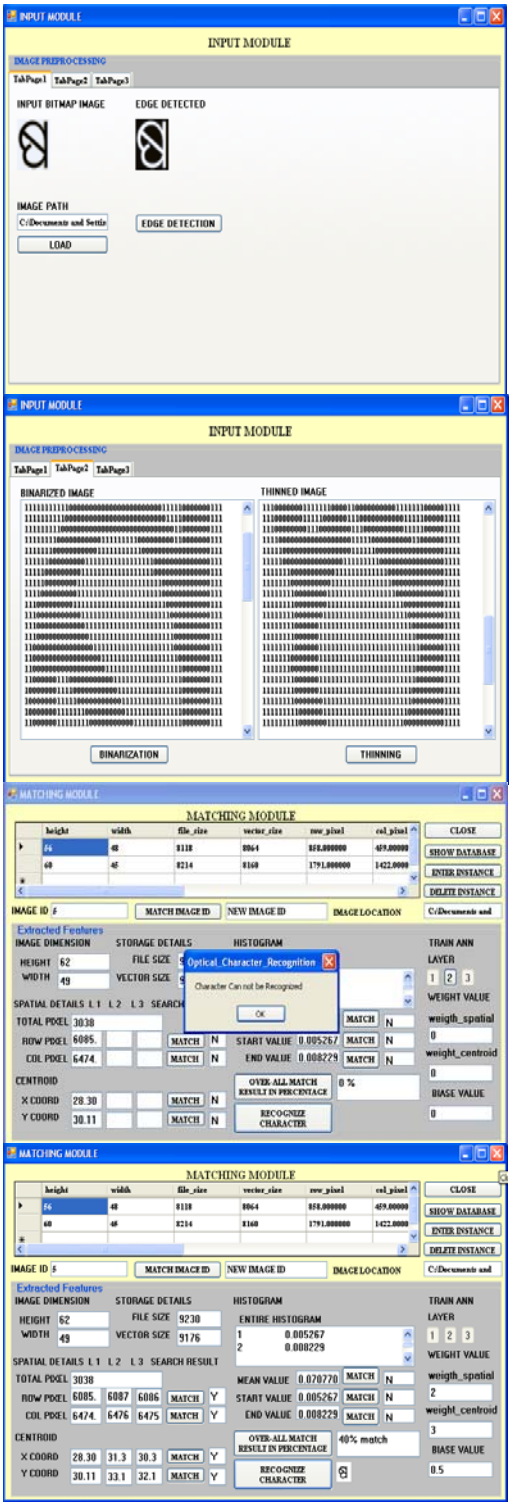

Fig-04 (Screenshots of Oriya OCR being developed)

Oriya alphabets are complex in nature and consist of 115 alphabets among which around 60 characters are very difficult to recognize because they are composite characters (conjuncts). Fixed font style and font size are the biggest problem of OCR.

Of even greater concern is the problem of misreading a character (substitutions). In particular, if the system does not accurately balance dollar data, customer dissatisfaction will occur. The success of any OCR device to read accurately without substitutions is not the sole responsibility of the hardware manufacturer. Much depends on the quality of the items to be processed.

However, these limits are not objectionable to most applications, and dedicated users of OCR systems are growing each year. But the ability to read a special character is not, by itself, sufficient to create a successful system.

ACKNOWLEDGEMENT

The authors sincerely thank DIT,MCIT, Government of India for providing financial assistance for this work.

REFERENCES

[1] Digital Image Processing, by Rafael C. Gonzalez, Richard E. Woods.

[2] A Complete Development System for Oriya Script, *Sanghamitra Mohanty, Hemanta Kumar Behera RC-ILTS-Oriya,Dept. of Comp. Sc. And Appl.,Utkal University,Bhubaneswar*

[3] S.Mohanty, K.Sahoo and H. K. Behera," *A New Algorithm for the restoration of characters in old noisy document with varying level of intensities*", ISC Conference, India, Jan'2003.

[4] S. Kahan, T. Pavlidis and H. S. Bairb, "On the Recognition of Printed characters of any font and size", IEEE Trans. Pattern Analysis Machine Intelligence, Vol-9, 1987, pp. 274 –287.

[5] LTG (Language Technologies Group), 2003. Optical Character Recognition for Printed Kannada Text Documents. SERC, IISc Bangalore. VijayaKumar, B., 2001.

[6] Gonzalez, R.C., Woods, R.E., Eddins, S.L., 2004. Digital Image Processing Using MATLAB. PHI Pearson. Unicode, 2000. The Unicode Standard Version 3.0. Addison Wesley.

[7] Rajavelu, M. T. Muaavi, and M. V. Shirvaikar, "A neural network approach to character recognition," *Neuml Networks*, Vol. 2, No. 5, 1989, pp. 387-389.

[8] Chaudhuri, B.B., Pal, U., "A complete Bangla OCR system", Pattern Recognition, Vol. 31, pp. 531–549, 1998.

[9] Complete OCR for Printed Hindi Text in Devnagari Script", Sixth International Conference on Document Analysis and Recognition, IEEE Publication, Seatle USA, 2001. Page(s):800-804.

[10] Nallasamy Mani and Bala Srinivasan, "Application of Artificial Network Model for Optical Character Recognition", System, *Man and Cybernetics*, 1997,"Computational Cybernetics and Simulation". *International Conference* on 12-15 Oct. 1997 page(s): 2517-25203.

[11] Veena Bansal and R.M.K. Sinha, "A Devnagari OCR and A Brief Overview of OCR for Indian Script", *PROC Symposium on Transaction support System (STRANS 2001),*Feb. 15-17, 2001, Kanpur, India.

[12] Bansal, V., Sinha, R.M.K., "Partitioning and Searching Dictionary for Correction of Optically Read Devnagari Character Strings", *Document Analysis and Recognition*, 1999. ICDAR'99, Proceedings of the Fifth International.

[13] A. A. Chaudhary, E.A.S. Ahmad, S. Hossain, C. M.Rahman, "OCR of Bangla Character Using Neural Network: A better Approach", *2nd International Conference on Electrical Engineering (ICEE 2002)*, khuln, Bangladesh.

[14] Grain U and Choudhuri B B 1998 compound character recognition by run number based metric distance. *Proc. SPIE Annual Symposium on Electronic Imaging, San Jose, USA, pp 90-97.*

[15] S. Kahan, T. Pavlidis and H. S. Bairb, "On the Recognition of Printed characters of any font and size", IEEE Trans. Pattern Analysis Machine Intelligence, Vol-9, 1987, pp. 274 -287