International Journal of Information Technology and Knowledge Management
July-December 2010, Volume 2, No. 2, pp. 639-643

OCR for Telugu Script Using Back-Propagation Based Classifier

Rinki Singh[1] & Mandeep Kaur[2]

This paper deals with the theory and implementation of an Optical Character Recognition (OCR) system for printed Telugu script, which exploits the inherent characteristics of Telugu scripts, one of the major scheduled language of India, spoken by more than 66 million people, especially in South India. The principle idea is to convert images of text documents such as those obtained from scanning a document into editable text. The system consider a images as input, separates the lines, words and then characters step by step and then recognizes the character using artificial neural network approach, in which creating a character matrix and a corresponding suitable network structure is key. The features detection methods are simple and robust. The various features that are considered for classification are the character height, character width, the number of horizontal lines (long and short), the number of vertical lines (long and short), number of slope lines, special dots. The glyphs are now set ready for classification based on these features. The extracted features are passed to neural network where the characters are classified by supervised learning of Back Propagation algorithm which compromises training, calculation of error, and modifying weights and then testing the given image. These classes are mapped onto Unicode for recognition. Once the characters are recognized they can be replaced by the standard fonts to integrate information from diverse sources.

Keywords: OCR, Supervised Learning, BPN Algorithm, ANN, MLP

1. Introduction

Optical Character Recognition abbreviated as OCR means that converting some text image into computer editable text format. For example the encoding scheme can be ASCII code. But in this thesis Unicode-16 is considered as converted text, due to its ability to represent all known symbols in a single encoding. In Unicode, first 256 codes of these new sets are reserved for the ASCII set in order to maintain compatibility with existing systems [3] [16]. The Telugu Unicode range is U+0C01 to U+0C47. Telugu language has the second largest number of speakers mainly concentrated in South India. It is the official language of Andhra Pradesh and second widely spoken language in Tamilnadu, Karnataka. Lots of recognition systems are available in computer science and also OCR plays a prominent role in computer science. Recognition system works well for simple language like English. It has only 26 character sets. And for standard text there are 52 numbers of characters including capital and small letters. But a complex but organized language like Telugu, OCR system is still in preliminary level. The reason of its complexities are its characters shapes, its top bars and end bars more over it has some modified, vowel and compound characters and also one of the important reasons for poor recognition in OCR system is he error in character recognition[1] [17].

[1]PG Student, Dept. of Comp. Sc. & Engg, Lingaya's University, Faridabad, India

[2]Assistant Professor, Dept. of Comp. Sc. & Engg., Lingaya's University, Faridabad, India

Email: rinkisingh18@gmail.com[1], mandeephanzra@gmail.com[2]

Unlike Latin script, Chinese script, Devnagri script or Bangla script, Telugu characters are rarely containing horizontal, vertical or diagonal lines. Basically the Telugu characters are obtained by joining circular shapes (full or partial) or of different sizes with some modifiers. Not much work has been reported on the development of Optical Character Recognition (OCR) systems for Telugu text [1] [17]. Therefore, it is an area of current research. The Telugu characters consists of 60 symbols, of which 16 are vowels, 3 vowel modifiers, and 41 consonants as some of them shown in Fig1.

Fig. 1: Some of Telugu Alphabets

2. Related Work

Optical character recognition (OCR) has been one of the most well studied problems in Pattern Recognition. Today, reasonably efficient and inexpensive OCR packages are commercially available to recognize printed texts in widely used language such as English, Chinese, and Japanese, etc. A verities of techniques of Pattern Recognition such as Template matching, Neural Networks, Syntactical Analyses, Wavelet Theory, Hidden Markov Models, Bayesian theory, etc. have been explored to develop robust OCRs for different languages such as Latin, Chinese, Hangul scripts, Arabic scripts also. There have also been some attempts to develop

OCRs for some Indian language like Devnagri, Bengali, Telugu and Gujarati [14].

The first reported work on OCR of Telugu Character is done by Rajasekharan, B.I. Deekshatulu. This work was generation and recognition of printed Telugu characters by using Computer graphics and image processing techniques. It was able to identify 50 primitive features and proposes a two-stage syntax-aided character recognition system. Primitives are joined and superimposed appropriately to define individual characters. Firstly, these primitives are recognized by a Sequential Template Matching mechanism [1].The basic letters are recognized by a process called On the Curve Coding. Though the experimental results are sufficiently promising, its validity in a practical OCR is not investigated. The next report is done by M.B. Sukhaswami, P. Seetharamulu, A.K. Pujari for Recognition of printed Telugu characters using neural networks [1] [2]. An extensive study is undertaken to identify the structural characteristics of Telugu script and the distinct symbols of the Telugu language are categorized based on their relative size. The authors propose neural network architecture and investigate different learning techniques to explore the recognition capability of such the network. It is an Image Mapped system and it is demonstrated that the proposed network can yield extremely efficient recognition. The work is a demonstration of robustness of a hierarchical Hopfield Network for the purpose of recognition of noisy Telugu characters. The next reported work was presented by Arun K. Pujari,C. Dhanunjaya Naidu, M Sreenivasa Rao and B.C. Jinag a for recognition of an intelligent character recognizer for Telugu scripts using multiresolution analysis and associative memory. In this report very encouraging experimental results on certain fonts are reported. But techniques was not applicable to other Indian Languages also but it was not clear whether the specific features of Telugu scripts are exploited or not.

3. THE PROPOSED METHOD

There are many algorithms and ways to accomplish the recognition of optical character. Although none of the can claims to provide cent percent in character recognition, many of them provide good results. The proposed OCR system will support various functionalities in terms of proposed specifications. These are divided into 3 OCR Function phases as shown in Fig 2.

Phage I

This phase includes three functions: collection of documents, scanning of documents, segmentation and feature extraction.

Collection of Data

First of all proposed system require a large number of raw data or collected data will be processed and later trained the system. It is very important to collect the data. Later on there is a need to compare with similar kind of data. And also consider the complexity of data because the next step will be dependent on data type.

Scanning the Documents

In second step the collected data will be placed over the scanner. A scanner software is invoked which scans the document. Here the text is free of mathematical symbols, figures or tables. The document will be sent to program that saves it in preferably JPG or GIF format, so the image of the document can be obtained when needed.

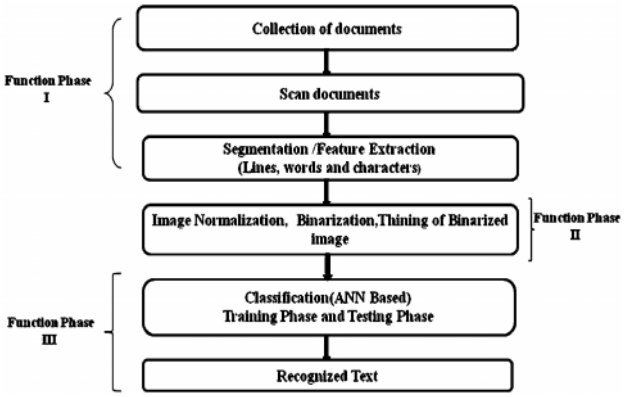

Fig. 2: Basic Block Diagram of Proposed System

Segmentation

After scanning the documents, image is passed to the segmentation phase, where the image is decomposed into individual characters. It consists of three steps.

Line Segmentation: Generally the problems occurs during the detection process are of two types

1. All words in a text lines are not aligned properly, and;
2. Gap between text lines is not uniform. It can be possible that at some places the gaps between the lines may be zero. It is essential to count the number of character lines in a character image in order to delimiting the boundary within which the detection can proceed. Thus detecting the next character in an image does not necessarily involve scanning the whole image all over again [4].

Word Segmentation: The word segmentation is based on looking for the vertical gaps in the segmented line, and checking them to identify the beginning and end of words.

Character Segmentation: After detection of reference line it is removed from the word to separate out the characters again by looking for vertical gaps in the segmented word, and checking them to identify the beginning and end of character [7].

Feature Extraction

Feature extraction is vital part of any recognition system and it is followed by segmentation process. In feature extraction stage each character is represented as a feature vector, which becomes its identity and also make the same character assume different appearance. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate with the least amount of elements. After detecting the individual symbols it is able to extract the general features such as width, height, closed shapes, diagonal lines, line intersections, special dots, etc [9].

Phase II

This second phase of the OCR function consists of normalization, image binarization and thinning of binarized image.

Normalization/Windowing

Normalization provide the uniformity in the input and store pattern that moves towards the recognition system, thus the windowing the character means to bring the character to a standard image window size. It is also helpful in minimizing the processing [13]. This is required because after segmentation each character may have a different window size thus giving different features for the same character. Since the height and width of individual image vary, an adaptive sampling algorithm is implemented.

Image Binarization

In Image binarization, the text image which is gray scale image is converted into a binary image with each pixel taking a value of 0 or 1 represent an individual pixel of image. Here consider the background pixels have a value of 1 and the foreground pixels have value of 0. Actually the main target is finding a vector from the image. So image is processed and then binary image is created. Here, the Otsu's threshold algorithm is used to binarize the gray scale image [13].

Image Thinning (Skeletonization)

The characters of the text page have to be thinned prior to recognition. After the binarization process the image is thinned so only the skeleton remains. Thinning or skeletonization is the process of extracting border pixels from the image matrix, by preserving the connectedness of the object and its end points. Hence, this process can be seen as a conditional deletion of boundary pixels [15]. Thus the basic approach of skeletonization algorithm is to delete from the object y simple border points, that have more than one neighbor in y and whose deletion does not locally disconnects y as shown in Table 1. A number of skeletonization algorithms have been proposed and are being used. For this paper, the thinning of binarized image is done by one of the most widely used algorithms that are Hilditch algorithm and its variants.

Table1
Thinning of Binarized Image

Y1	Y2	Y3
Y8	P	Y4
Y7	Y6	Y5

Phase III

The last phase of the proposed system consists of classification and recognition strategies that are based on the feature extraction on the previous phase I.

Classification

Classification that is nothing but Character recognition will be performed on the based on feature extraction done on the previous steps, which consist of two stage that is training and testing phase that will discuss in the following stage.

Back Propagation based Classifier

Artificial Neural Network (ANN) is to solving problems that are the most difficult to solve by traditional computational methods. The ANN can be trained into two main groups that are supervised (or associative learning) and unsupervised (self-organization) learning. The supervised learning means the network learn by example where in unsupervised method is not given any target value (example).Unsupervised learning is more complex and difficult to implement. Unsupervised learning is based on clustering of a data [5, 6].

In this paper Back Propagation (BP) algorithm or propagation of error is one of the well known algorithm in neural network, which is efficient learning of Multi-Layer Perceptron (MLP) and that is based on supervised learning mode. The introduction of BP algorithm has overcome the drawbacks of previous NN algorithm in 1970s where single layer perceptron fail to solve a simple XOR problem. The architecture consist of three layer structure: first layer is called input layer, hidden layer between, and the last layer is called output layers as shown in Fig3. The hidden layer

is responsible to provide efficiency in the output. Fig4 shows BP's structure.

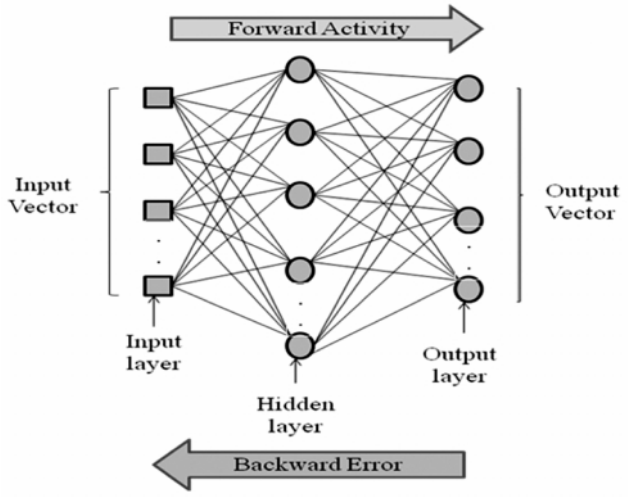

Fig. 3: Back-propagation Neural Network

The Back Propagation algorithm is having two phases as given below:

1. Forward Phase: In forward phage the features extracted from image are propagated from input layer to layer until the output pattern is generated by the output layer. Actually each neuron receives a signal from the neurons in the previous layer, and each of those signals is multiply by a separate weight value [7] [14]. The weighted inputs are summed, and passed through an activation function which scales the output to a fixed range values. Then the output is compared to the known-good output and a mean-squared error signal will be calculated.

2. Backward Phase: In case of this phase weights are adjusted to reduce the error by propagating the output error signals back to the network. This process is where the Back Propagation neural network gets its name and is known as the backward pass. The training set is repeatedly presented to the network and the weight values are adjusted unit the overall error is below a predetermined tolerance [7] [14].

BPN algorithm:

1. Begins by constructing a network with inputs n_i, hidden units n_h and outputs units n_o, according to the specified topology parameters.

2. Initialize weights with random values within the specified weight bias value. Until

   Satisfied or termination condition is met, Do

   For each training example, do

   a. Input the training example to the network and compute the network outputs.

   b. For each output units k, calculate

      $\delta_k = o_k (1 – o_k) (t_k –)$

      The $\delta_k = o_k (1 – o_k)$ term is necessary in the equation because of the sigmoid function. The sigmoid function is $f(x) = (1/e^{-x})$.

   c. For each hidden unit h, calculate

      $\delta_h = o_h(1 – o_h)\Sigma_{k \in outputs} w_{hk}\delta_k$

   d. Update each network weights $w_{i,j}$

      $w_{i,j} = w_{ij} + \Delta w_{ij}$

      Where $\Delta w_{ij} = \eta \phi_j x_{ij}$

Notations

The notation described below were used in algorithm:

w_{ij}	Weight on the connection from the i^{th} input unit to j^{th} hidden unit
w_{hk}	weight on the connection from the j^{th} hidden unit to the k^{th} output unit
o_h	Actual output for the h^{th} hidden unit
o_k	Actual output for the k^{th} output unit
t_k	Desired output for k^{th} output unit
δ_k	Signal error term for the k^{th} output unit
δ_h	Signal error term for the h^{th} hidden unit
η	learning rate

4. CONCLUSION

In this paper an OCR system is proposed for Telugu characters using artificial neural network. It is efficient in character recognition of mass standardized document via Back-propagation algorithm. The proposed system involves various phases like segmentation, recognition etc. The proposed character recognition algorithms operate on input image and efficiently recognize the individual characters. Here, the design, approach and implementation are driven by the need for a practical OCR system for printed Telugu character recognition.

REFERENCES

[1] Arun K Pujari, Prof. C Dhanunjaya Naidu, AI Lab, University of Hyderabad "An Adaptive Character Recognizer for Telugu Scripts using Multiresolution Analysis a Associative Memory".

[2] C. Vasantha Lakshmi, C. Patvard, Department of Electrical Engineering Dayalbagh Educational Institute, Agra, India, "A High Accuracy OCR System for Printed Telugu Text".

[3] Sheetalashmi R. Abnikant Singh, Department of Electrical Engg, IIT Kanpur, "Optical Character Recognition for Printed Tamil Text using Unicode".

[4] R. Jagadeesh Kannan, RMK Engg College, Chennai, India, "A Comparative Study of Optical Character Recognition for Tamil Script".

[5] Sang Sung Park,Won Gyo Jung,Young Geun Shin, Dong-Sik Jang, Department of Industrial System and Information Engineering, Korea University, South Korea, "Optical Character System Using BP Algorithm".

[6] Ahmad M. Sarhan, and Omar I. Al Helalat, "Arabic Character Recognition using Artificial Neural Networks and Statistical Analysis".

[7] Princess Summaya University for Science and Technology, Amman, Jordan, "Online Handwritten Character Recognition Using an Optical Backpropagation Neural Networks".

[8] Mansoor Al-A'ali and Jamil Ahmad, "Computer Science Department, College of Information Technology, University of Bahrain, IQRA University, Pakistan," Optical Character Recognition System for Arabic Text Using Cursive Multi-Directional Approach.

[9] School of Computer Science and Engineering University of New South Wales, Sydney 2052, Australia, "Optical Character Recognition: Neural Network Analysis of Hand-Printed Characters".

[10] Simon Tanner, King's Digital Consultancy Services, "Deciding Whether Optical Character Recognition is Feasible".

[11] Gail Hodge, CENDI Secretariat, Information International Associates, Inc. Oak Ridge, Tennessee, "Cendi Analysis of Scanning/optical Character Recognition Position Descriptions"

[12] Deepayan Sarkar, Department of Statistics, University of Wisconsin Madison, "Optical Character Recognition using Neural Networks (ECE 539 Project Report)".

[13] Adnan Md. Shoeb Shatil,Center for Research on Bangla Lanuage Processing BRAC University, Dhaka, Bangaladesh, "Research Report on Bangla Optical Chracter Recognition Using Kohonen Netwok".

[14] K.Y. Rajput and Sangeeta Mishra, Mumbai India, "Recognition and Editing of Devnagari Handwriting Using Neural Network".

[15] Sanghamitra Mohnaty, Hemanta Kumar Behera, RC-ILTS-Oriya, "A Complete OCR Development System for Oriya Script".

[16] "Telugu Unicode", http://www.xenotypetech.com/samplepdfs/TL_Sample.pdf

[17] "Telugu language", http://www.nriol.com/telugu-page.asp.