IJCSN International Journal of Computer Science and Network, May 2014
ISSN (Online) : 2277-5420
www.IJCSN.org

Devanagari Handwriting Recognition and Editing Using Neural Network

Sohan Lal Sahu

RSR Rungta College of Engineering & Technology (RSR-RCET) , Bhilai 490024

Abstract- Character recognition plays an important role in the modern world. It can solve more complex problems and makes humans' job easier Cost effective and less time consuming, businesses, post offices, banks, security systems, and even the field of robotics employ this system as the base of their operations. Whether you are processing a check, performing an eye/face scan at the airport entrance, or teaching a robot to pick up and object, you are employing the system of Character Recognition. One field that has developed from Character Recognition is Optical Character Recognition (OCR). Character Recognition has even advanced into a newer field – Handwritten Recognition, which of course is also based on the simplicity of Character Recognition.

In this paper, a system for recognizing hand written Indian Devnagari script is presented. The system considers handwritten images as an 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. In addition, knowledge of how one is Deriving the Input from a Character Matrix must first be obtained before one may proceed. Afterwards, the Feed Forward Algorithm gives insight into the entire working of a neural network; followed by the Back Propagation Algorithm which compromises Training, Calculation of Error, and Modifying Weights. Once the characters are recognized they can be replaced by the standard fonts to integrate information from diverse sources.

Keywords- *Image Processing, Character Matrix, Artificial Intelligence, Feedforword -Back propagation Algorithm,Neural Network Training* .

1. Introduction

Character Recognition plays advance role in Handwritten Recognition. The new idea for computers, such as Microsoft's new Tablet PC, is pen-based computing, which employs lazy recognition that runs the character recognition system silently in the background instead of in real time. India is a multilingual country of more than 1 billion population with 18 constitutional languages and 10 different scripts. Although since 1965 English had been officially recognized as an "associated language" in India, according to the latest census report, less than 10 percent of the Indian population can either read or write English. In a developing country like India there is an urgent need for the research and development of its own language technologies. More than 300 million people around the world use Devnagari script. It is the base script of many languages in India, such as Hindi and Sanskrit. And there are other languages that use variants of this script. Devnagari script is a logical composition of its constituent symbols in two dimensions. It has eleven vowels and thirty three simple consonants. A horizontal line is drawn on top of all characters which is referred to as the header line or *shirorekha.* A character is usually written such that it is vertically separate from its neighbors. Characters are joined by a horizontal bar that creates an imaginary line by which Devnagari text is suspended, and no spaces are used between words. A single or double vertical line called a Danda was traditionally used to indicate the end of phrase or sentence. Devnagari script has many multi-stroke characters.

2. Problem Formulation

2.1 Handwriting Style Variations

A writing style is based on the alignment and variable forms of characters. Both on-line and off-line handwritten characters have enormous variety in shape compared to machine printed characters. Handwriting style variations occur mostly between different writers but also within the handwriting of any individual writer.

2.2 Variations of Characters

Handwritten characters can vary in both their static and dynamic properties. Static properties are the underlying, ideal models of the characters, the allographs, and the geometrical properties such as relative positions and sizes of the strokes, corners, retraces, ornamentals, sizes and aspect ratios of the characters, and the general slant of the writing. Dynamic properties are more involved with the generative aspects of the characters. Characters can look

similar although their number of strokes, and the drawing order and direction of the strokes may vary considerably.

2.3 Situational Factors

Some examples of situational factors are stress, haste, motivation, distractions from the writing task, and the method of presentation.

2.4 Material Factors

The writing instrument, surface, and form constitute the material factors of the writing style.

3. Proposed Solution

In internal approaches, the segmentation and recognition of handwriting are performed simultaneously. Internal segmentation approach is used in the recognition of whole words written in mixed style. Typically, the recognition methods are based on

1 Hidden Markov Models (HMMs)
2 Time Delayed Neural Networks (TDNN)
3 Multi-State Time Delayed Neural Networks (MS-TDNN)
4 Convolutional Neural Networks (CNN)

First, these models are used for evaluating character probabilities, or some other matching scores, for highly overlapping segments of the handwriting data.

The ideas suggested behind these problems are:
1. Separation of line from the given text.
2. Separation of word from the line detected.
3. Separation of character from the word detected.
4. Separation of conjunct (joined character) characters if they are.

4. Steps Involved

The Devnagari script is used in Sanskrit, Hindi, Marathi and Nepali languages and OCR developed can be used for applications to these languages .The text is also assumed to consist of simple characters along with the headline. Various blocks of the project are shown in the fig.1 and are as follows:

1. Binarisation
2. Thinning
3. Windowing
4. Feature Extraction
5. Segmentation into lines/ characters
6. Feed forward network
7. Back propagation algorithm for training and testing

Testing phase and evaluating the probability of occurrence of various characters

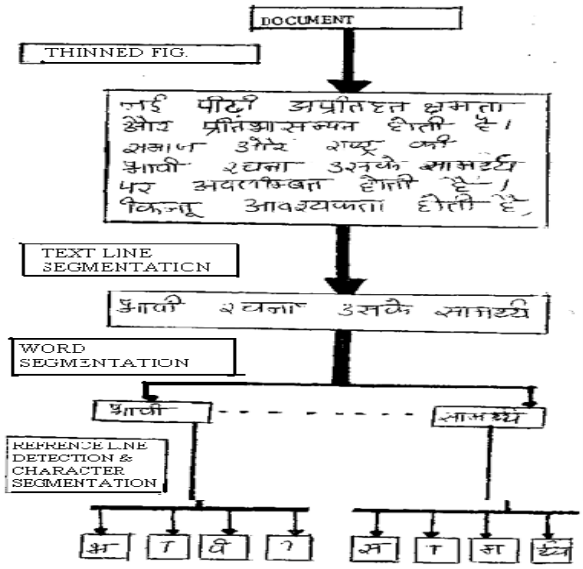
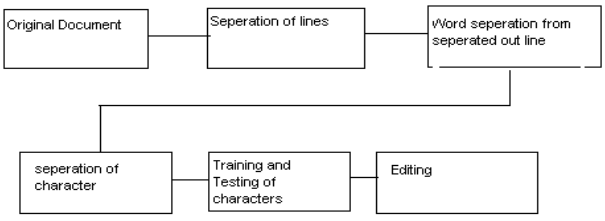


Fig 1. Testing Phrase



Algorithm

Fig 2. Algorithm

4.1 Image Binarisation

In image binarisation, the text image which is gray scale image is converted into a binary image with each pixel taking a value of 0 or 1 depending on threshold value of the image. The technique is most commonly employed for determining the threshold involves analyzing the histogram of gray scale levels in the digitized image.

$$I(x,y) = 0 \quad I(x,y) < t$$
$$= 1 \quad I(x,y) >= t$$

4.2 Thinning of Binarised Image

The characters of the text page have to be thinned prior to recognition. Thinning removes the points in such a way that only the skeleton of a branch pattern remains. Thinning algorithms transform an object into a set of

simple digital arcs the structure is not influenced by small contour inflections. The basic approach of thinning algorithm is to delete from the object x simple border points, that have more than one neighbor in x and whose deletion does not locally disconnect x.

4.3 Windowing

Windowing the character means to bring the character to a standard image window size. This is required because after segmentation each character may have a different window size thus giving different features for the same character.
Windowing is done in two ways. First the character is fitted in a tightest possible bounding box so that the background area surroundings and which does not contain any useful information can be removed .Secondly the character size is increased to a standard size in such a way that the connectivity as well as shape of the characters are preserved. This is done by expanding the characters to standard size, where intermediate pixels can be easily interpolated.

4.4 Feature Extraction

Effective feature extraction is important in all pattern recognition tasks. Final character recognition is closely related to

1. Feature used to represent the character.
2. Comparison/decision module used to compare the feature vectors of the two characters.

These two are the most crucial and the most difficult parts of the recognition process.

Line Segmentation: Detection of text lines in a hand written script is not an easy job. Generally the problems encountered in the detection process are of two types 1) all words in a text lines are not aligned, and 2) gap between text lines is not uniform; At some places in lines gap may be zero. Approach used in this system is based on the histogram at an inclination of the binary image. First we find the horizontal density histogram of a few rows in the image.

Word Segmentation: Word boundaries are detected by looking for the vertical gaps in the segmented line, and checking them to identify the beginning and ending of words.

Character Segmentation: After detection of reference line it is removed from the word to separate out characters.

5. Neural Networks

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process.

5.1 Neural Network Character Recognition

In order to have a learning task that is reasonably workable, a great amount of pre-processing of the digits is carried out using conventional Artificial Intelligence (AI) techniques. This is done before the Characters are fed to the ANN. Handwritten Character recognition is a created system that is used to recognize handwritten Character. But once these tasks have been carried out, the Characters are available as individual items. But the Characters are still in different sizes. Therefore a normalization step has to be performed so we can have to have Characters in equal sizes.

After the characters are normalized, they are fed into the ANN. This is a feed forward network with three hidden layers. The input is a 30 x 30 array that corresponds to the size of a normalized pixel image. The handwritten Character images get transformed into histograms and these histograms are fed into a neural network.

This neural network outputs scores for matching the input Character against the forty nine possible Characters. The data is trained and tested and it outputs the accuracy rate. The results can show us which Character needs more training to reach high accuracies and which Character the system had a difficulty to identify. The difficult task is there are some handwritten Characters that often run together or not fully connected.

6. Training Phase and Database of Trained Neural Networks for Each Character

The neural network classification techniques such as multilayer perceptrons trained by Error Back Propagation (EBP) algorithm.

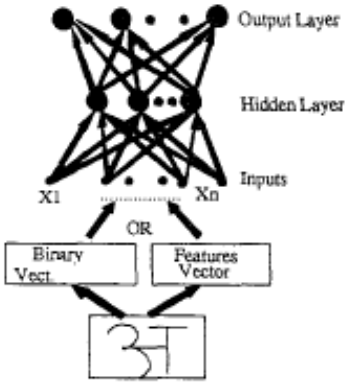


Figure3. Multilayer perceptron network

7. Editing

For editing purpose we have created a new database with fixed font size .When the character is recognized then it is replaced with this standard font.

8. Results and Discussions

The original image is taken as following Fig. 4:

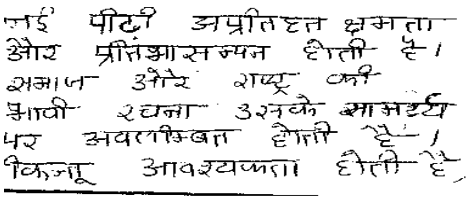


Fig.4. Sample of Image containing Devnagari hand writing.

The very first step is to separate out lines. This work is done by counting the number of black pixels per row and by plotting number of black pixels per row number is as shown in the fig 5 .

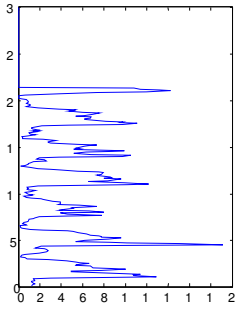


Fig.5 Histogram of image containing Devnagari handwriting.

After this step we can apply an appropriate threshold and then we can separate out the lines. The result of Line separation is shown in Fig. 6

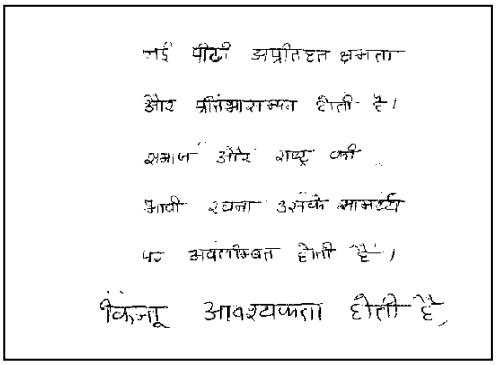


Fig. 6. Result of Line Separation

Then it word separation is carried out by vertical threshold. The result of word separation is shown in Fig 7.

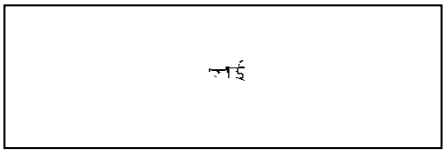


Fig. 7 Result of Word Separation.

Finally a character is separated from each of the word by applying slightly higher threshold vertically.

Fig. 8 shows database created for all the devnagari characters.

Fig. 8. Database of Devnagari character.

Character recognition is then done using neural network.

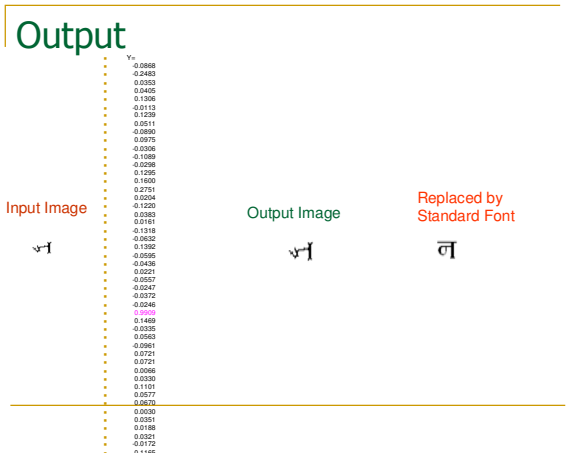


Fig. 9. Outcome of code

References

[1] Krishnamachari Jayanthi ,Akihiro Suzuki,Hiroshi Kanai,Yoshiyuki Kawazoe,Masayuki Kimura and Keniti Kido , "Devnagari character recognition using structure analysis," IEEE-1989.CH2766-4/89/0000-0363.

[2] S.Hewavitharana ,H.C.Fernando , " A two stage classification approach to Tamil handwriting" ,Tamil Internate 2002 , California USA.pg.118-124.

[3] RAJIV KAPOOR,*$ DEEPAK BAGAI,$ T. S. KAMAL ,"Skew angle detection of a cursive handwritten Devanagari script character image", *J. Indian Inst. Sci.,* © Indian Institute of Science **82**, 161–175 ,May□Aug. 2002,.

[4] Reena Bajaj, Lipika Dey and Antanu Chaudhuri, "Devnagari numeral recognition by combining decision of multiple connectionist classifiers," *S¯adhan¯a* Vol. 27, Part 1, , pp. 59–72 February 2002.

[5] Dileep Kumar , "An AI approach to hand written Devnagari script recognition",IIT Delhi.

[6] B B Chaudhuri, U Pal and M Mitra , "Automatic recognition of printed Oriya script" , *S¯adhan¯a* Vol. 27, Part 1, pp. 23–34, February 2002.

[7] Yi Li,Yefeng Zheng ,and David Doermann, " Detecting text lines in handwritten documents ",The 18th International Conference on Pattern Recognition (ICPR'06).

[8] U. Bhattacharya and B. B. Chaudhuri, "Databases for Research on Recognition of Handwritten Characters of Indian Scripts ," Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05).

[9] K.H. Aparna, Vidhya Subramanian, M. Kasirajan, G. Vijay Prakash, V.S. Chakravarthy, "Online Handwriting Recognition for Tamil" , Proceedings of the 9th Int'l Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004).

[10] Kamesh Madduri, K.H. Aparna, V.S. Chakravarthy, "PATRAM - A Handwritten Word Processor for Indian Languages, Proceedings of the 9th Int'l Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004).