International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-3 Issue-7, December 2013

Optical Character Recognition of Printed Text in Devanagari Using Neuro - Fuzzy Integrated System

Ankush A.Mohod, Nilesh N.Kasat

Abstract-In this paper, we deal with the recognition of printed Devanagari characters using Neuro-fuzzy integrated system. The paper shows measurement of the effectiveness of classifier in terms of precision in recognition. An attempt is made to adopt Neuro-fuzzy integrated system for classification purpose. In this paper, we have considered sample test image and the characters in the test image are recognized relative to the database created by the user using Neuro-fuzzy integrated system.

Keywords- Character Recognition, Printed Devanagari text, Histogram, GLCM

I. INTRODUCTION

Machine simulation of human reading has become a topic of serious research since the introduction of digital computers. The main reason for such an effort was not only the challenges in simulating human reading but also the possibility of efficient applications in which data present on paper documents has to be transferred into machine-readable format. With increasing the interest of computer applications, modern society needs the input text into computer readable format. This research is the simple approach to implement that dream as the initial step to convert the input text into computer readable form. Digital document processing is gaining popularity for application to office and library automation, bank and postal services, publishing houses and communication technology. English character recognition (CR) is extensively studied by many researchers and various commercial systems are available for it. But in case of Indian languages, the research work is limited due to the complex structure of the language. Recognition of printed characters is itself a challenging problem since there is variation of same character due to change of fonts or introduction of different types of noises. In OCR domain, it has been observed that a single feature extraction method and a single classification algorithm can't give the better recognition rate. Neural networks and Fuzzy logic are two complimentary technologies which are used in pattern recognition process. It is therefore, a compound feature extraction approach based on soft computing for recognition of printed Devanagari script. We have implemented the steps of PCR(Printed Character Recognition) System like Preprocessing, Segmentation, Feature extraction and Classification. After finding out feature of the segmented characters Artificial Neuro-fuzzy interference system (ANFIS)[1][2] &[3] will be used for classification purpose.

Manuscript published on 30 December 2013.
*Correspondence Author(s)
Ankush A.Mohod, Department of Electronics & Telecommunication Engineering, Sipna College of Engineering & Technology, Amravati.
Prof. Nilesh N.Kasat, Department of Electronics & Telecommunication Engineering, Sipna College of Engineering & Technology, Amravati.

II. DEVANAGARI OPTICAL CHARACTER RECOGNITION

Devanagari word is derived from Sanskrit words Deva(god) and Nagari(City) jointly for "City of gods"[3]. Devanagari script is derived from ancient Brahmi script emerged something around 11th century AD.

Devanagari was initially developed to write Sanskrit but was later adopted to write many other languages. Devanagari is the mother of all most all Indian scripts. It is used to write languages such as Hindi, Marathi Marwari, Bhojpuri, Kashmiri, Konkani and Sindhi. The script has a complex composition of its constituent symbols. Devanagari has 13 vowels [Fig.1(a)] and 34 consonants[Fig.1(b)] along with 14 modifiers of vowels and of "rakar," as shown in Fig.1(a) symbols. Apart from the vowels and consonants, there are compound (composite) characters in most of Indian scripts including Devanagari, which are formed by combining two or more basic characters. The shape of a compound (composite) character is usually more complex than its constituent characters. A vowel following a consonant may take a modified shape, which depending on the vowel is placed to the left, right, top, or bottom of the consonant, and are called modifiers or "matras." Text, characters, and digits are written from left to right in Devanagari. There is no concept of upper or lowercase characters.

Fig.1 (a) Vowels and modifiers of Devanagari script. (b) Consonants and their corresponding half forms in Devanagari script. (c) Some combinations of consonants with themselves

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Every Indian script has its own specified composition rules for combining vowels, consonants, and modifiers. Some of them can be combined with their type, as shown in Fig.1.c.
A modifier can be attached to a vowel or to a consonant. Consonants may have a half form when they are combined with other consonants as depicted in Fig.1.c Except for some characters, the half forms of consonants are the left part of original consonants with the right part removed. Some special combinations are also shown in Fig.1.c,where a new character or the half forms of consonants may appear in the lower half of the new composite forms. Another distinctive feature of Devanagari is the presence of a horizontal line on the top of all characters. This line is known as header line or "shirorekha"(Fig.2) The words can typically be divided into three strips: top, core, and bottom. The header line separates the top and core strips and a virtual base line separates the core and lower strips. The top strip generally contains the top modifiers, and bottom strip contains lower modifiers.

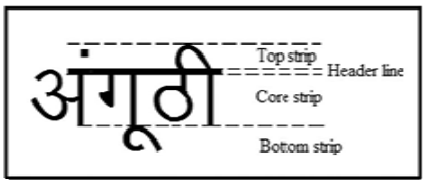


Fig.2. Three strips of a word in Devanagari script

III. PROCESSING STEPS

Character recognition is one of the important tasks in the pattern recognition. There are four different phases in the optical character recognition system, namely:

- Preprocessing Stage
- Segmentation
- Feature Extraction
- Classification

3.1Preprocessing Stage

Preprocessing is an important step of applying a number of procedures for smoothing, enhancing, filtering, etc. for making a digital image usable by subsequent algorithm in order to improve their readability for optical character recognition software. The various stages involved in the preprocessing stage are:
Binerization
Noise elimination
Size Normalization
Thinning

3.1.1Binerization

Conversion of a gray-scale image into binary image is called as Binerization or Thresholding. There are two approaches for conversion of grayscale image to binary form,i.e Global Threshold Local or adaptive Threshold Global threshold selects single threshold value based on estimation of background level from intensity histogram of image. Local or adaptive threshold uses different values for each pixel according to local area information.

3.1.2Noise Elimination

Noise that exists in images is one of the major obstacles in pattern recognition tasks.Noise can occur at different stages like image capturing,transmission.Noise elimination is also called as smoothing.It can be used to reduce fine textured noise and to improve quality of image.

3.1.3Size Normalization

Normalization is applied in order to get characters of uniform size. It provides a tremendous reduction in data size. Each segmented character is normalized to fit within suitable matrix like 32x32 or 64x64 so that all characters have same data size [5].

3.1.4Thinning

Thinning is a morphological operation that is used to remove selected foreground pixels from binary images.Thinning extracts shape information of the characters.Thinning is also called as skeletonization.Skeletonization refers to the process of reducing the width of a line from many pixels to just single pixel. Various standard functions are now available in MATLAB for above operation.

3.2Segmentation

It is one the most important process that decides the success of character recognition technique. It is used to decompose an image of a sequence of characters into sub images of individual symbols by segmenting lines and words.
Step i: **Locate the header line:** We compute the horizontal projection of the word image box. The row containing maximum number of black pixels is considered to be the header line. Let this position be denoted by *hLinePos.*
Step ii: Segmentation of Line*:* Text lines are detected by horizontal scanning. For segmentation of line, we scan scanned document page horizontally from the top and find the last row containing all white pixels, before a black pixel is found. Then we find the first row containing entire white pixel just after the end of black pixels. We repeated this process on entire page to find out all lines.
Step iii*:* Segmentation of Words*:* After finding a particular line we separate individual words. This is done by vertical scanning.
Step iv*:* Segmentation of Individual Characters: Once we get the words we segment it to individual characters. Before segmenting words to individual characters, we locate the head line. This is done by finding the rows having maximum number of black pixels in a word. After locating head line we remove it i.e. converts it in white pixels. After removing head line our word is divided into three horizontal parts known as upper zone, middle zone and lower zone. Individual characters are separated from each zone by applying vertical scanning.
Step v: Separate character/symbol boxes of the image below the header line: To do this, we make vertical projection of the image starting from the hLinePos to the bottom row of the word image box.The colums that have no black pixels are treated as boundaries for extracting image boxes corresponding to characters.
Step vi: Separate symbols of the top strip: To do this,we compute the vertical projection of the image,starting from the top row of the image to the hLinePos.The columns that have no black pixels are used as delimiters for extracting top modifier symbol boxes.Because of the presence of various modifiers,Devanagari character segmentation is very complex.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Fig.3. Devanagari Test Image

Fig.4..Devanagari Segmented Image

3.3 Feature Extraction

This step is the heart of the OCR system.Feature extraction is a set of procedures for extracting or measuring the most and relevant shape information content in the character or pattern. This step simplifies the process of classification. D.Trier et.al [5] discussed various feature extraction methods for character recognition. Here,we have extracted features as Histogram,Graylevel Co-occurance Matrix(GLCM) and Color domain of the created database and the test image.

3.4 Classification

Classification is performed based on the extracted features. Printed Character Recognition(PCR) systems extensively use the methodologies of pattern recognition,which assigns an unknown sample to a predefined class. Numerous techniques for PCR are investigated by the researchers.OCR classification techniques can be classified as follows:

1.Template Matching
2.Statistical Techniques
3.Neural Networks
4.Support vector Machine (SVM)algorithms
5.Combination classifiers

The above approaches are neither necessarily independent nor disjoint from each other. Various classification methods have their superiorities and weaknesses.Hence many times multiple classifiers are combined to solve a given classification problem.

Hybrid approaches could be considered as one of the main contributions of soft computing,with neuro-fuzzy systems being the first and probably the most successful hybrid approach till now. Neuro-fuzzy systems incorporate the elements from Fuzzy logic(FL) AND Neural Networks(NN).This idea of hybridization originates from two observations:

1. Fuzzy systems neither capable of parallel computation, whereas these charactristics are clearly attributed to NNs.
2 .NNs lack flexibility, human interaction which lies at the core of FL. Thus we have used ANFIS for classification purpose.

In Fuzzy logic and Neural Networks, we have to adjust weights and Number of hidden layers in order to achieve approximate 100% accuracy. So, Neural network and Fuzzy logic are not capable to get approximate 100% accuracy. So, we are using Neuro-fuzzy integrated system to achieve approximate 100% recognition rate.

Adaptive Neuro-Fuzzy Inference System(ANFIS)

Adaptive neuro fuzzy inference system (ANFIS) is a kind of neural network that is based on Takagi–Sugeno fuzzy inference system. Since it integrates both neural networks and fuzzy logic principles, it has potential to capture the benefits of both in a single framework. Its inference system corresponds to a set of fuzzy IF–THEN rules that have learning capability to approximate nonlinear functions. Hence, ANFIS is considered to be universal approximator.

Here we have used ANFIS for recognition. We used two parameters x and y as input and output by using segmentation and feature extraction data. Then we set Membership function(MF) and type of Membership function. After setting the epoch number, we train data and get output as recognized character.

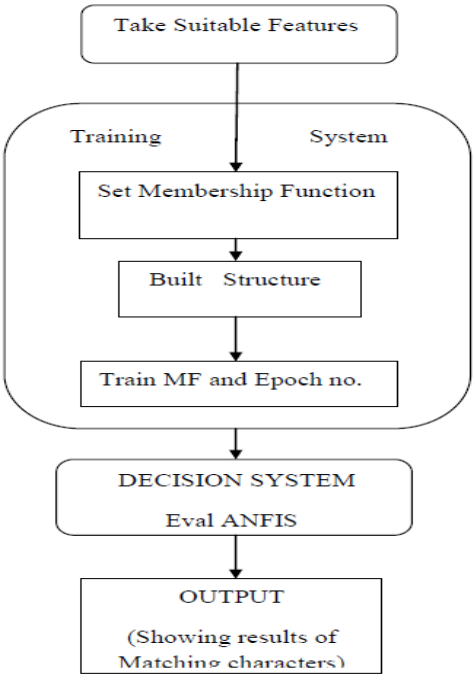
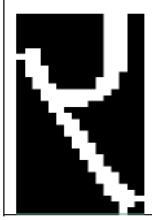


Fig.5 Flow-chart of ANFIS

IV. RESULT

Table 4.1

INPUT (Characters)	OUTPUT	ACCURACY (%)
		Approximately 100%
		Approximately 100%

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

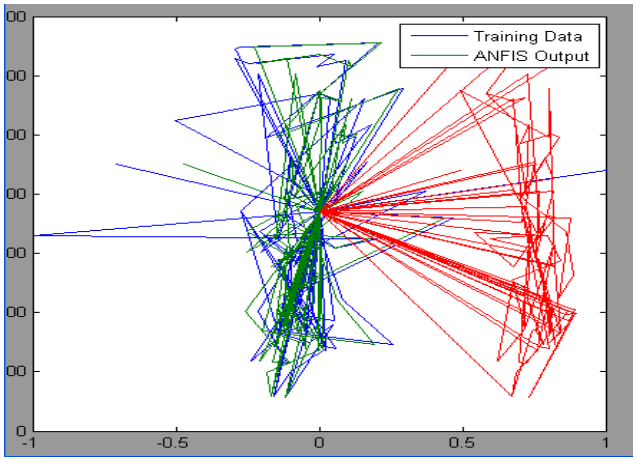


Fig. 6. Graph after training

From the graph which we obtain after training, it is observed that their is maximum overlapping between the training data and the ANFIS output. So, the accuracy of recognition we obtain is nearly 100%.

V. CONCLUSION

In this paper, we have proposed an Neuro-fuzzy integrated system based approach for the recognition of Printed Devanagari characters. The results of our approach are promising. The accuracy reported in this paper is approximately 100%.

REFRENCES

1. Jang J S. R.,C.-T. Sun, E. Mizutani, (1997) Neuro-Fuzzy and Soft Computing A Computation Approach to Learning and Machine Intelligence, Matlab Curriculum Series, Prentice Hall.
2. Zadeh, Lotfi A., "Fuzzy Logic, Neural Networks, and Soft Computing," Communication of the ACM, March 1994, Vol. 37 No. 3, pages 77-84.
3. U. Pal and B. B. Chaudhuri, “Indian script character recognition: A Survey”, Pattern Recognition, Vol. 37, pp. 1887-1899, 2004.
4. José antónio barros vieira1, fernando morgado dias2, alexandre manuel mota comparison between artificial neural networks and neurofuzzy Systems in modeling and control: a case study
5. D.Trier ,A.K.Jain ,T.Taxt , “Feature Extraction Method for Character Recognition-A Survey” ,Pattern Recognition,pp.641-662, Vol.29,No.4,1996.
6. Neural Networks, A Comprehensive Foundation by Simon Haykin
7. Introduction to Artificial Intelligence And Expert Systems by Dan W. Patterson.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)