*I.J. Image, Graphics and Signal Processing,* **2012, 6, 15-21**
Published Online July 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2012.06.03

# Classification and Recognition of Printed Hindi Characters Using Artificial Neural Networks

B.Indira [1]
Kasturba Gandhi Degree & PG College for Women, Secunderabad, A.P, India
indsneha@rediffmail.com

M.Shalini [1]
Kasturba Gandhi Degree & PG College for Women, Secunderabad, A.P, India
shalini_praveenkumar@yahoo.co.in

M.V. Ramana Murthy [2]
Chairman, Department of Computer Science, Faculty of Science, Osmania University, Hyderabad, India.
mv.rm50@gmail.com

Mahaboob Sharief Shaik [2]
Faculty of Computing & Information Technology, King Abdul Aziz, University, Jeddah, KSA
mshaik@kau.edu.sa

***Abstract*—** Character Recognition is one of the important tasks in Pattern Recognition. The complexity of the character recognition problem depends on the character set to be recognized. Neural Network is one of the most widely used and popular techniques for character recognition problem. This paper discusses the classification and recognition of printed Hindi Vowels and Consonants using Artificial Neural Networks. The vowels and consonants in Hindi characters can be divided in to sub groups based on certain significant characteristics. For each group, a separate network is designed and trained to recognize the characters which belong to that group. When a test character is given, appropriate neural network is invoked to recognize the character in that group, based on the features in that character. The accuracy of the network is analyzed by giving various test patterns to the system.

***Index Terms*—** Pattern Recognition, Character Recognition, Artificial Neural Network, Feature extraction, Thinning

## I. INTRODUCTION

Pattern Recognition is defined as the field concerned with machine recognition of meaningful regularities in noisy and complex environments [1]. There are various applications of pattern recognition such as character recognition, online signature verification, and face recognition and so on. Character Recognition is the electronic conversion of scanned images of printed or handwritten text into machine readable text. Character recognition system is the base for many different types of applications in various fields, many of which we use in our daily lives. Hindi character recognition is the challenging problem in Pattern Recognition and Neural Networks is one of the most commonly used techniques for character recognition and classification due to their learning and generalization abilities. This paper describes and discusses the classification and recognition of printed Hindi characters using Artificial Neural Networks. Some of the previous approaches related to this work are given in section II. The entire recognition process is explained in section III. Section IV gives the training procedure of neural networks. Testing and results were discussed in section V and finally concluding remarks were given in section VI.

## II. REVIEW OF PREVIOUS APPROACHES

A good text recognizer has many commercial and practical applications such as processing cheques in

banks, documentation of library materials, extracting data from paper documents, searching data in scanned book, automation of any organization like post office, which involve lot of manual task of interpreting text. The problem of text recognition has been attempted by many different approaches; some of them are Template matching, Feature extraction, Geometric approach and neural networks.

Template matching approach is one of the most simplistic approaches. This is based on matching the stored data against the character to be recognized. Template matching involves determining similarities between the given template and stored database and output the image that produces the higher similarity measure. This technique works effectively with recognition of standard fonts, but gives poor performance with handwritten characters, noisy characters and deformed images.

The objective of feature extraction is to capture the essential characteristics of the symbols and this is one of the most difficult problems of pattern recognition. In this approach, statistical distribution of points is analyzed and orthogonal properties are extracted. For each symbol a feature vector is calculated and stored in database, and recognition is performed by finding distance of feature vector of input image with those stored in the database and giving the symbol with minimum deviation. This is very sensitive to noise and edge thickness, but performs well on handwritten character set.

In geometric approach an attempt is made to extract features that are quite explicit and can be very easily interpreted. These features depend upon the physical properties, such as number of joints, relative position; number of end points, aspect ratio etc. Classes formed on the basis of these geometric features are quite distinct, with not much of overlapping. The main draw back with this approach is that this approach depends heavily on the character set.

Neural network techniques are more popular to perform Character Recognition. It has been reported that Neural Networks could produce high recognition accuracy. Neural Networks with various architectures and training algorithms have been applied successfully for Character recognition. In this, neural network is first trained by the multiple sample images of each alphabet. Then, in the recognition processes, the neural network recognizes the given input symbol. Neural networks are capable of providing good recognition even at the presence of noise but the draw back is they require a lot of training time.

Character recognition remains a highly challenging task. Hindi character recognition is one of the most difficult tasks of optical character recognition. This section gives a brief overview of related research work. The research work pertaining to character recognition of Indian languages is very limited.

Dr. P.S. Deshpande et.al, proposed a novel approach on character encoding and regular expressions for shape recognition in their paper [2]. The method is independent of the specific aspect of individual shapes, such as thickness of line, size of character and shapes. In this, features are extracted in the form of regular expression. They achieved an accuracy of 90%.

A Devanagari text recognition system was designed by Veena Bansali [3] in her research work by integrating knowledge sources, features of characters such as horizontal zero crossings, moments, aspect ratios, position of vertex points and pixel density, with structural description of characters.

Aditi Goyal, Kartikay Khandelwal, Piyush Keshri [4], in their paper discussed about various image pre-processing, feature extraction and classification algorithms, to design high performance OCR software for handwritten Hindi alphabets. Image preprocessing included Median filtering, Background removal, Threshold and sparsity removal. In feature selection and extraction, histograms of oriented gradients were used.

This provides a flexible feature and helps to deal with high bias and high variance issues. The basic back-propagation algorithm is used to determine the weight matrix. Features were tested on a reduced training set using naïve Bayes and support vector machines. They observed that SVM gave better results than naïve bayes. The performance obtained with handwritten letters is 98 %.

Pooja Agarwal, Hanumandlu and Brijesh, in their paper "Coarse Classification of Handwritten Hindi characters" [5], described a system for the categorization of complete handwritten Hindi character set into subgroups based on some similarity measure. They proposed an algorithm for finding and removal of header line and identification of present position of vertical bar in handwritten Hindi character. Experimental results demonstrate that t heir algorithm is effective and achieved a classification rate of 97.25%.

Optical character recognition for printed Devanagari script using Artificial Neural Networks was presented by Raghuraj Singh et. Al [6]. The paper [7,8,9,10] proposed a technique for OCR system for different fonts and sizes of printed Devanagari script and achieved a high recognition rate.

## III. RECOGNITION PROCESS

Character recognition is one of the important tasks in pattern recognition. The complexity of the character recognition problem depends on the character set to be recognized. Character recognition process is dependant upon number of factors like various font sizes, noise, broken lines or characters etc. and these factors influence the results of recognition system [11]. Artificial Neural Network is one of the techniques widely used for character recognition problem and considered as a powerful classifier on account of their high computation rate accomplished by massive parallelism [12, 14]. There are four different phases in character recognition processes namely Character acquisition, preprocessing stages, grouping of characters and Character Recognition.

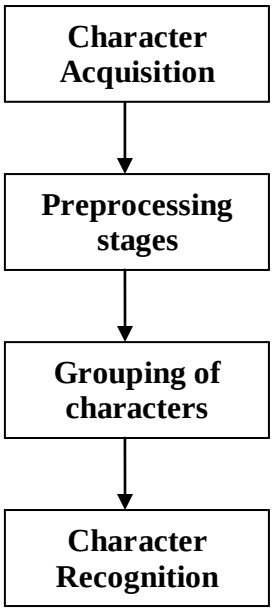



Figure 1: different phases in character recognition process

### *A. Character Acquisition*

Character acquisition is the first phase in any image processing or pattern recognition task. In this paper the images of Hindi characters, in tiff, jpg, bmp, and gif format are obtained through a scanner. After obtaining the digital image, the next step is to apply preprocessing in order to improve the image clarity and also the accuracy of recognition rates.

### *B. Preprocessing Stages*

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 are:

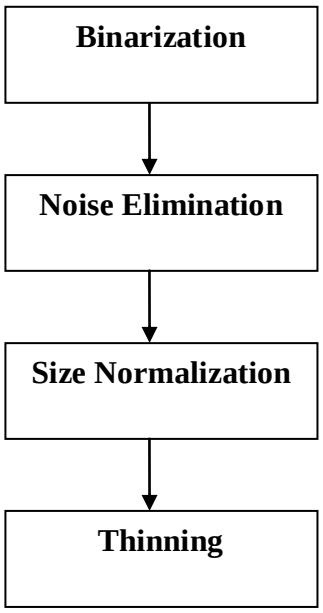



Figure 2: preprocessing stages

1) Binarization

Linearization (thresholding) refers to the conversion of a gray-scale image into a binary image. This is also generally referred to as thresholding. There are two approaches for conversion of gray level image to binary form. First one is global threshold which picks one threshold value for the entire image, based on estimation of the background level from the intensity histogram of the image. The other one is local or adaptive threshold which uses different values for each pixel according to the local area information.

The purpose of binarization is to identify the extent of objects and also to concentrate on the shape analysis, in which case the intensities of pixels are less significant than the shape of a region.

2) Noise Elimination

Noise that exists in images is one of the major obstacles in pattern recognition tasks. The quality of image degrades with noise. Noise can occur at different stages like image capturing, transmission and compression. Various standard algorithms, filters and morphological operations are available for removing noise that exists in images. Gaussian filter is one of the popular and effective noise removal techniques. Noise elimination is also called as smoothing. It can be used to reduce fine textured noise and to improve the quality of the image. The techniques like morphological operations are used to connect unconnected pixels, to remove isolated pixels, and also in smoothening pixels boundary.

3) Size normalization

Normalization is applied to obtain characters of uniform size. It provides a tremendous reduction in data size. The character patterns have different sizes. The input to the neural network is an array of fixed size. Hence to make the image suitable to this size, size normalization is required. Normalization should reduce the size of the image without getting the structure of the image altered. In this paper, the sizes of Hindi characters are reduced to the size of 32 x 32.

### *C. Grouping of Characters*

After preprocessing of character, features of character are extracted. This step is heart of the system. This step helps in classifying the characters based on their features. The vowels and consonants of Hindi character set are divided into sub groups based on certain significant characteristics. The vertical bar feature and its position in the character is used to group the vowels and consonants in to sub groups. The characters are classified in to 3 sub groups. The first sub group consists of character without any vertical bar. Characters with vertical bar at right side of the character are in second sub group and the third group includes the characters having a vertical bar in the middle of the character.

### *D. Character Recognition*

Character recognition system is the base for many different types of applications in various fields, many of which we use in our daily life. Cost effective and less time consuming businesses, post offices, banks, security systems, number plate recognition system and even the field of robotics employ this system as the base of their operations. Recently, neural network became very popular as a technique to perform character recognition

[15, 16, 17]. It has been reported that neural networks are capable of providing good recognition rate even at the presence of noise where other methods normally fail. The inherent pattern recognition abilities of layered neural networks lend itself perfectly to this type of task, by autonomously learning the complex mappings in high dimensional input data. In this work, the problem of character recognition is solved using neural networks. Neural network maps the set of input values to set of output values. The multi layer feed forward connectionist model trained by back propagation (gradient – descent) is used to recognize the given input character.

## IV. THINNING

Thinning is a morphological operation that is used to remove selected foreground pixels from binary images. Thinning extracts the shape information of the characters. Thinning is also called skeletonization. Skeletonization refers to the process of reducing the width of a line from many pixels to just single pixel. This process can remove irregularities in letters and in turn, makes the recognition algorithm simpler because they only have to operate on a character stroke, which is only one pixel wide. It also reduces the memory space required for storing the information about the input characters and also reduces the processing time too. The final stage in preprocessing is thinning. Image thinning extracts a skeleton of the image without loss of the topological properties [13]. The thinning algorithm consists of both boundary pixel analysis and connectivity analysis. The binary image before and after thinning is given in the figure 3.

The above preprocessing steps are applied to all vowels and consonants of Hindi characters.

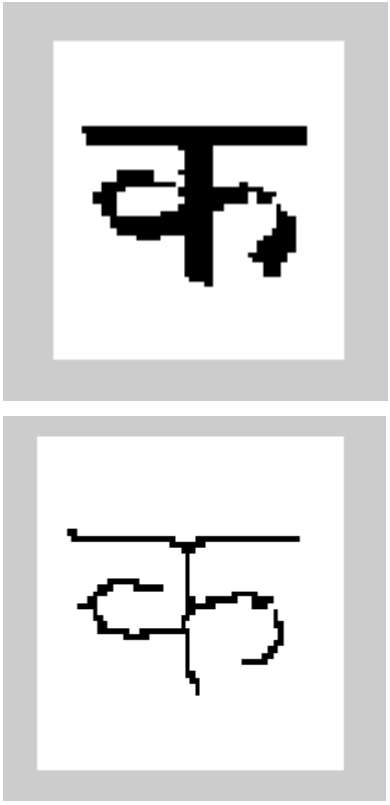


Figure 3: Original and the thinned image

## V. TRAINING THE NEURAL NETWORK

Recognition of printed Hindi character is performed by giving the input image of the character. The given image is first converted into a gray scale image. Then the gray level image is converted into a binary image using threshold. Afterwards noise is eliminated by using filters. The next step is size normalization followed by thinning which extracts the skeleton of the image without any loss of the topological properties. After preprocessing of character, features of character are extracted. This step helps in classifying the characters based on their features.

In this work, Hindi characters can be classified into three subgroups. Hence three feed forward neural networks are designed to recognize the characters in each sub group. The back propagation learning algorithm is used to train each network with the characters in that group as input examples to that network. This network takes input-output vector pairs during training. During training the weights of the

network are iteratively adjusted to minimize error. The input image, number of neurons in each layer, learning rate, momentum and error value is given as input. The integrated module takes its input from the output of any one of the three networks and with the help of the lookup table of that subgroup, it recognizes and classifies the given character.

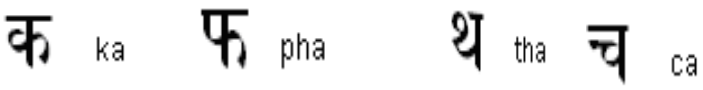
## VI. TESTING AND RESULTS:

The vowels and consonants of Hindi character set are divided into 3 subgroups based on certain significant characteristics. For each subgroup, a separate feedforward neural network is designed to recognize the character which belongs to that group. Back propagation algorithm is used to train each network with examples. Finally, after training the neural networks with proper set of examples of each sub group, the performance of the system is tested with various test patterns with and without noise. The system recognized the character which had a noise up to 40%. Overall performance of network is tested with test samples. It achieved a recognition rate in the range of 76% - 95% for various samples. The results also show that the recognition accuracy and efficiency of the network increases with more number of training samples.

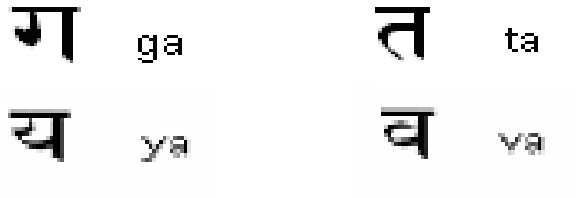
## VII. CONCLUSION

Character recognition is one of the important applications of pattern recognition. Instead of using only one neural network for recognizing and classifying Hindi vowels and consonants, we divided the characters into three subgroups based on certain significant features and three feed forward neural networks are designed and trained to recognize the character in each subgroup. It is observed that recognition accuracy is increased by using the concept of subgroups instead of single network.

This work is limited to recognition of Hindi vowels and consonants. Good recognition rate is achieved for the following characters since these characters are of simplistic in nature.




Poor recognition rate of character is achieved for the following characters since these characters have close resemblance with ya and va.




## RFERENCES

[1] Jie Liu, Jiguisun, Shengshenq wang – "Pattern Recognition – An Overview" International journal of computer science and network security, Vol.6, No. 6, June 2006.

[2] Dr. P.S. Deshpande, Mrs. Latesh Malik, Mrs. Sandhya Arora "Characterizing Hand written Devanagari Characters using Evolved Regular Expressions", 1-4244-0549-1/06, IEEE 2006.

[3] Veena Bansali "Integrating with source in Devanagari text recognition", PhD thesis , 1999.

[4] Aditi Goyal,Kartikay Khandelwal, piyush Keshri, "Optical Character Recognition for Handwritten Hindi", Stanford University, CS229 Machine Learning, Fall,2010.

[5] Pooja Agarwal, M. Hanmandlu, Brejesh Lall, "Coarse classification of Handwritten Hindi Characters", International Journal of Advanced Science and Technology, Vol. 10, September, 2009.

[6] 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 & Communication Vol. 1, No. 1, January – June 2010, PP. 91-95.

[7] Latesh Malik, P. S. Deshpande "Recognition of Printed Devanagari Characters with Regular

Expression in Finite State Models", proceedings of International workshop on Machine Intelligence Research, 2009.

[8] Veena Bansal and R.M.K. Sinha "Segmentation of Touching and Fused Devanagari Characters" IIT, Kanpur.

[9]. Swamy Saran Atul and Swapneel Prasanth Mishra " Hand-Written Devnagari Character Recognition", NIT, Rourkela, 2007.

[10]Line Eikvil "Optical Character Recognition", December, 1993.

[11]Dyashankar Singh, Sajay Kr. Singh, Dr. (Mrs) MitreyeeDutta, Hand Written Character Recognition Using Twelve Directional Feature Input and Neural Network – 2010 International Journal of Computer Applications(0975 – 8887) vol. 1 – No. 3.

[12]A.K.Jain, Mohiuddin "Artificial Neural Networks: A Tutorial", IEEE Computers, 29, 31-44, 1996.

[13]Anil K. Jain, Orivind Due Trier and Torfinn Taxt "Feature Extraction Methods For Character Methods- A survey", Pattern Recognition, Vol. 29. No 4, PP 641-662, 1996.

[14] B. Indira, "Artificial neural networks and its use in Automatic Recognition of Vehicle Registration Numbers", Ph.D.thesis, 2008.

[15] Brain Clow – "A comparison of neural network training Methods for Character Recognition" – Carleton University April, 2003.

[16] Dave Anderson and George MCneil – " A DACS state of the Art report on Artificial Neural Network Technology" – International Conference on Computer Vision – New York-2003.

[17] E.Barnard, "Optimization for Training Neural Nets " IEEE transactions on Neural Networks, Vol. 3, No. 2 , March 1992, PP 232-240.

**M.Shalini** completed her MCA from Osmania University and M.Phil from sri padmavathi mahila university in the area of Artificial neural networks She has more than 17 years of teaching experience. she is currently working as an associate professor, department of computer science (PG Courses) in kasturba Gandhi college for women ,affiliated to osmania university, Hyderabad, Andhra Pradesh, India. Her research interests include artificial neural networks, image processing, AI and web designing.

**Dr. B.Indira** completed her MCA from Kakatiya University and Ph.D from sri padmavathi Mahila University in the area of Artificial neural networks. She has more than 16 years of teaching experience. she is currently working as an associate professor, department of computer science (PG Courses) in kasturba Gandhi college for women ,affiliated to osmania university, Hyderabad, Andhra Pradesh, India Her research interests include artificial neural networks, image processing , AI and cloud computing.

**Dr. M.V.Ramana Murthy** is currently working as chairman, department of computer science, at faculty of science, osmania university, Hyderabad, India. His area of interest is information security. He has more than 28 years of teaching experience. He has several publications

**Mr. Mahaboob Sharief Shaik** has completed master's degree in computer applications in the year 1998 and presently working as lecturer at faculty of computing & information technology, King abdulaziz university, Jeddah, Saudi Arabia. His area of interest is network/information security, image processing and database.