Handwritten Hindi Character Recognition

Abhishek Sharma,Vikram Verma

***Abstract*—to recognize handwritten Hindi character automatically is very difficult task because of character written in different ways like curves and cursively written characters are different in various ways.OCR is the process of taking images or pictures of letters or typewritten content and change them into information that a machine can easily interpret for eg. large organizations and libraries taking manual duplicates of books, magazines and old printed material and utilizing ocr to put them into computers. There are many approaches for character recognition. In this paper we have reviewed several techniques of character recognition. Important stages of character recognition include preprocessing, segmentation, feature extraction and classification. Many classification techniques have been surveyed in this paper like neural networks, support vector machine (SVM), template matching etc.**

***Index Terms*— Ocr System, Segmentation, Neural Network, Combination Classifier.**

INTRODUCTION

1.1 Recognition Process

The process of changing of scanned image into a text document consist the following steps shown in the figure1.

1.2 Stages in Design of OCR Systems

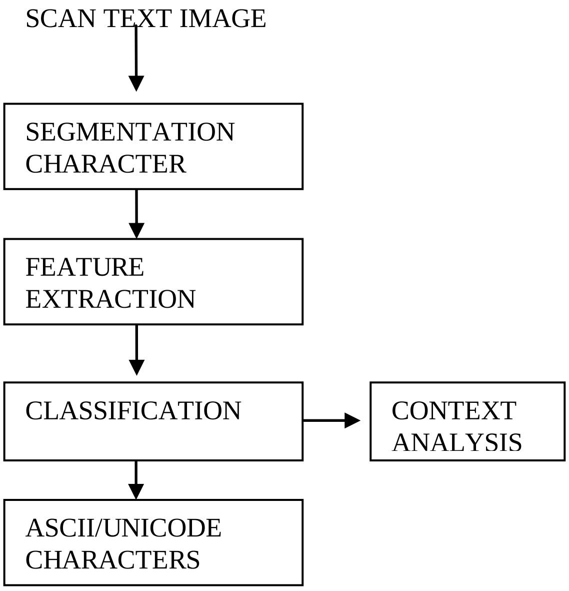

Figure1. Stages in OCR Design [4]

1.3 OCR SYSTEM

In the design of proposed OCR system following steps has been followed:

- Preprocessing
- Segmentation
- Feature Extraction
- Classification [4]

1.3.1 Pre-processing: - The pre-processing stage yields a clean document in the sense that maximal shape information with minimal noise and maximal compression on normalized image is obtained [2].

1.3.2Segmentation:- In the segmentation stage, an image made up of sequence of characters is break down into sub-images of discrete character. The pre-processed input image is divided into unique characters and using a labeling process assigned a number to each character [8]. Segmentation is an important part because the extent one can reach in separation of words, lines or characters directly

Manuscript received May, 2016.

***Abhishek Sharma**, M.Tech student in ECE dept. of Seth Jai Parkash Mukand Lal instt.of Engg. And Tech(Kurukshetra University,Kurukshetra).*
***Vikram Verma,**Professor in IT dept. of Seth Jai Parkash Mukand Lal Instt.of Engg. And Technology.*

affect the rate of identification of the script [2].Line, Word and Character Segmentation: -Once the blocks of text are detected, the OCR system automatically finds individual text lines, then segments the words, and then separates the characters appropriately [4].

1.3.3Feature Extraction: - After segmentation of the character, feature extraction like top and bottom ,height, width, horizontal line, vertical line detection is done [2].

1.3.4Classification:-For classification or identification back propagation algorithm is used [2]. Based on the extracted features Classification is carry out. For inceptive classification of characters; considered three features as follows:

- Histogram of projection based on pixel value
- Mean Distance
- based on spatial position of pixel Histogram of projection [4].

1.4 CLASSIFICATION OF CHARACTER

For Optical character Recognition. These are the different techniques of classification

- Neural Networks.
- Statistical Techniques.
- SVM (Support Vector Machine) algorithms.
- Combination classifier.
- Template Matching

1.4.1 Neural Network: To develop an accurate OCR system is a difficult task and requires a lot of effort. Such types of systems are often complex and can hide much of logic behind the code. To achieve the good performance and improved quality of recognition of character, in OCR applications the artificial neural network performs the important function [5]. Neural networks have been used in many different areas to solve a wide range of problems. Artificial Neural Network execution in identify extended sets of image pixel data. Offline recognition of character is used for this printed text document. It is a process of conversion of printed document or scanned page to ASCII character that a computer can identify [2].

1.4.2 Template Matching:

Another Optical Character Recognition techniques is Template matching. Template matching is the process of finding the location of a sub image which is called a template inside an image. Once a number of related templates are found their centers then are used as related points to determine the registration parameters [5].

1.4.3 Support Vector Machine Classifier:

Currently, SVM is widely used in object recognition and detection, text recognition, biometrics, speech recognition, etc. SVMs are now a day’s for a number of classification tasks among the best performers. SVM is based on binary classification, means at a time it can classify two class groups [6]. To produce a model (based on the training data) which predicts the target values of the test data is the goal of SVM [5].

1.4.4 Combination Classifier:

Decisions of different classifiers can be combined to improve recognition results. The combination can be done in various ways which depends on the types of information produce by the distinct classifiers. So Many techniques to combine multiple classifiers can be defined into three main groups according to their architecture:

1) Cascading (serial combination)
2) Hierarchical (tree-like), and
3) Parallel [5].

LITERATURE SURVEY

Richard G. Casey et al. [1996] In OCR process Character segmentation has long been a difficult area. The higher recognition rates for isolated characters and those obtained for words and connected character strings well define this fact. A good part of recent development in reading free printed and written text may be impute to more insightful managing of segmentation. Rather than to easy list sources, been developed, Segmentation methods are listed under four main headings. The "classical “approach which consists of methods that divide the input image into sub images, which are then classified. The process to break the image into Identifiable units is called "dissection." The second class of methods avoids dissection (i.e. cutting up), and segments the image either explicitly, by classification of pre defined windows, or implicitly by classification of subsets of structural features collected from the image as a whole. The third technique is a hybrid of the first two; involve dissection together with recombination rules to define possibe segments, but using classification to select from the range of acceptable segmentation possibilities offered by these sub images. Lastly, holistic approaches that stop segmentation by recognizing entire character strings as units are described.

Nisha Vasudeva et al. [2012] Expansion in Artificial Intelligence has lead to the expansions of many “smart” devices. The huge challenge in the area of image processing is to recognize documents both in handwritten and printed format. Character recognition is one of the frequently used biometric features for authentication of person as well as document. OCR (Optical Character recognition) is a kind of document image analysis where scanned digital image which contains either machine printed or handwritten script given as input into an OCR software engine and translating it into a modifiable machine readable digital text format. A Neural network is made to model the way in which the brain performs a specific task or function of interest. Each image character is made of 30×20 pixels. We have applied feature extraction technique for calculating the feature. Extracted Features from characters are guidelines of pixels with respect to their neighboring pixels. And these inputs are providing to a back propagation neural network with hidden and output layer. Back propagation Neural Network for proper recognition are used where the errors were improve through back propagation and rectified neuron values were transferred by feed-forward method in the neural network of multiple layers.

Meha Mathur et al. [2014] Hindi is a national language of India, near about 300 million people in India speak Hindi and write Devnagari script. There is problem of Hindi character recognition occurred and I propose k- means clustering based a recognition mechanism. Because of The large dataset of Hindi characters and their similar appearance makes the problem as there is no difference between the characters of texts written in Hindi as in English. K-means provides a pure degree of font independence and this is to compact the size of the training database. In this paper I propose an K-means clustering for OCR for Hindi characters. The major steps which are followed by a general OCR are preprocessing, character segmentation, classification, feature extraction and recognition. The paper propose a two masks one is for horizontal projection and second for vertical projection of gray scale image to find out & remove shirorekha of word to decompose into individual characters from the words.

Raghuraj Singh et al. [2010] Near about 300 million people in India who write Devnagari script and speak Hindi. Research in OCR (Optical Character Recognition) is popular for its application capacity in banks, post offices, defense organizations and library automation etc. Most of the OCR systems are available for European texts. In this paper, we have given a technique for OCR System for dissimilar five fonts and printed Devnagari size script using Artificial Neural Network. The recognition rate of the suggested OCR system with of Devnagari Script image document has been found to be fully high.

Neeraj Pratap et al. [2012] from last half century, the recognition of English character was studied and the results were that’s it can provide technology driven applications. But the similar technique cannot be used in case of Indian languages because of the complex nature in terms of computation and structure. Now days there are different methodologies which are rapidly growing in the area of Indian languages and character recognition. In the area of digital document processing the offices, schools and other organizations, banks, are working. Devnagari is the national language of India and generally 600 million people in India speaks this. Devnagari should be given more special consideration for analysis and document betterment due to its popularity. This paper is mainly related to the people working in the Devnagari Optical character and it provides an outline about DCR(Devnagari character recognition system). The current status of DCR is given and future research is also suggested in this paper.

Sonika Dogra et al. [2012] OCR (Optical Character Recognition) is a technique which can automatically identify the characters with an optical mechanism. OCR technology allows you the identification of printed or handwritten text documents. Main aim of this research paper is to make a recognition system which is used for the recognition of offline handwritten Hindi characters. For this suggested Diagonal feature extraction approach is used to extract features and SVM (system Support Vector Machine) is used as classifier .

Sonal P Ajmire et al.[2015] The idea of character recognition has achieved a lot of attention due to its many applications such as filling of various forms, in printed postal addressing, , multiple choice questions in certain examination and so on. This paper is an well defined study of handwritten Devanagari character recognition. It defines the organized techniques for handwritten character recognition for feature extraction. The character recognition can be of two types that are offline and online character recognition. This can be also classified as handwritten and printed character recognition. as compared to the printed character recognition. The handwritten character recognition has more applications.

J.Pradeep et al.[2011] in the paper using multilayer feed forward neural network An off-line alphabetical handwritten character recognition system is defined. Method called diagonal based feature extraction is introduced for extracting the features of the handwritten alphabets. Fifty sets of data each containing 26 alphabets written by different people, are used for training the neural network and different handwritten alphabetical characters nearly 570 are used for testing. The suggested recognition system performs well generating higher levels of recognition accuracy compared to the systems employing the usual vertical and horizontal methods of feature extraction. This system will be suitable for changing handwritten documents into structural text form and identify names written with hands.

KanakUpmanyu et al. [2014] to automatically recognize handwritten Hindi characters is a very difficult because of characters written in different methods like curves and cursively written characters are different in many ways. so, these characters are written in different sizes, orientation, dimension, thickness and format. text images which are written offline from a piece of paper are scanned optically i.e. OCR. Devanagari script has 13 vowels and 33 consonants so, an offline Hindi handwritten characters recognition system using neural network is presented in this paper, which can be used in famous and common applications like government data records, commercial forms, bank cheques , signature checking, billing process systems, post code recognition, , and passport readers. In this paper, Devanagari script characters are OCR from document images by using Gradient descent approach.

Dr.N.Rajalingam et al. [2011] Clustering is a data mining (machine learning) approach used to place data elements into related groups without having preliminary knowledge on the group definitions. This paper the authors provides an in depth explanation of execution of divisive and agglomerative clustering algorithms for many types of attributes. Database - the details of the sufferers of Tsunami in Thailand during the 2004 year, taken as the test data. Algorithms are constructed using VB (Visual programming) language and for the formation of the running time and clusters time needed of the algorithms using different linkages (agglomerative) to various types of data are taken for analysis.

Ref No	Technology Used	Parameters Included	Year	Findings
1	1. for clean fixed-pitch typewriting simple techniques based on white separations between characters are used.	Optical character recognition, character segmentation	1996	classification and segmentation have to be treated in an integrated manner to obtain high reliability in complex cases.
2	Artificial Neural Network	Character recognition ,multilayer perceptron	2012	time taken to find the charaters written with hands is also increases If the hidden nodes increases, .

3	K-Means clustering	Pre-processing , Segmentation, Feature Extraction, Classification,	2014	the problem of Hindi character recognition solves successfully with the use of K means. the horizontal and vertical gradient for edge detection using manual canny edge detector and take the vertical and horizontal projection of each character for matching with other character and their Classification.
4	Artificial Neural Network	Segmentation, Feature Extraction, Classification, ANN	2010	Input matrix of size 48X57 gives good results than other Options. With the image document of Devnagari Script The recognition rate of OCR system is quite high
5	Different approaches like HMM, neural networks	Image Classification, Segmentation, Devnagari Character Recognition	2012	For the high sureness in Classifications, character recognition and segmentation have to be treated in an combined manner to obtain more accuracy in complex cases.
6	Support vector machine	Handwritten Character Recognition, OCR, Feature Extraction, SVM.	2012	Combination of SVM classifier and diagonal feature extraction approach is best method for the recognition of handwritten characters.
7	statistical techniques	Pattern recognition, Classification, Feature Extraction.	2015	Statistical features give better result as compared to structural. Similarly Support Vector Machine is better classifier.
8	diagonal feature extraction scheme	Feature extraction, feed forward neural networks.	2011	method of diagonal of feature extraction genrate the highest recognition accuracy of 97.8 % for 54 features and 98.5% for 69 features
9	back-propagation neural network	Neural network Devanagari script, handwriting recognition, OCR, Feature extraction	2014	back-propagation network show recognition accuracy 92%
10	Hierarchical Clustering Algorithms .	Agglomerative , Divisive, Clustering, Tsunami Database, Data mining	2011	divisive algorithm works as twice as fast as that of Agglomerative algorithm. . running time get increased on an average of 6 times when the number of records get doubled.

Conclusion

Various papers have been studied for optical character recognition system techniques. From these papers we have concluded that different techniques have different way to recognize the character. In neural network printed document or scanned image converted into ASCII character that a computer can identify. Template matching is the process of finding the location of a sub image which is template.SVM is a binary based classification. Integration of SVM classifier with diagonal feature extraction method is the best approach for recognition of handwritten characters. HMM, neural networks and their integration are used as the powerful tools for the character recognition. Study paper have illustrated that the artificial neural network technique can be applied successfully to solve the devnagri OCR problem. It is also concluded that the 48x57 size input matrix gives better results than other choices. By segmentation method we can separate the touching characters and remove shrirorekha from the word. It is also concluded that k-means provide a real degree of font independence and this is to reduce the size of the training database.

A. References

1. RichardG.Casey and Eric Lecolinet,"A Survey of Methods and Strtegies in Character Segmentation", Member, IEEE, VOL. 18, NO. 7, JULY 1996.
2. Nisha Vasudeva, Hem Jyotsana Parashar and Singh Vijendra," Offline Character Recognition System Using Artificial Neural Network", International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012.
3. Meha Mathur, Anil Saroliya, " HCR Using K-Means Clustering Algorithm", Volume 3.Issue 7, July 2014.
4. Raghuraj Singh, C. S. Yadav, Prabhat Verma, Vibhash Yadav," Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network",Vol. 1, No. 1, January-June 2010, pp. 91-95.
5.Neeraj Pratap and Dr. Shwetank Arya ," A Review of Devnagari Character Recognition from Past to Future ",Volume 3, Issue 6, June 2012.
6.Sonika Dogra and ChandraPrakash,"PEHCHAAN:HINDI HANDWRITTEN CHARACTER RECOGNITION SYSTEM BASED ON SVM", Vol. 4 No. 05 May 2012.
7. Sonal P Ajmire K. G. Bagade P.LRamteke," An Analytical Study of Handwritten Devanagari Character Recognition", Volume 5, Issue 10, October-2015.
8. J.Pradeep , E.Srinivasan and S.Himavathi, " DIAGONAL BASED FEATURE EXTRACTION FOR HANDWRITTEN ALPHABETS RECOGNITION SYSTEM USING NEURAL NETWORK", Vol, 3, No 1, Feb 2011.
9. Kanak Upmanyu, Mr. Shahid Hussain and Dr. Rizwan Beg," Handwritten character recognition system with Devanagari script (SWARS)", Volume 2, Issue 9, September 2014.
10. Dr. N.Rajalingam and K.Ranjini,"Hierarchical Clustering Algorithm – A Comparative Study", Volume 19– No.3, April 2011.

.