ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

Optimized Hindi Script Recognition using OCR Feature Extraction Technique

Nisha Goyal[1], Er. Shilpa Jain[2]

Student, CSE, JCDM College of Engineering, Sirsa, India[1]

Assistant Professor, CSE, JCDM College of Engineering, Sirsa, India[2]

Abstract: Feature Extraction in any handwritten script is very important part of OCR. If right feature extraction technique is used, then it will make the recognition of individual character easy. In this we will try to use some technique that extracts the feature of individual character. Devanagari script is the popular and widespread script of India which is mostly used in India. It is used to write Hindi, Nepali and Marathi. We will try to extract feature of Hindi language. Hindi Language consists of vowels, constants and various modifiers. Proper Feature Extraction technique on Hindi character is challenging task. In this we use simple Feature Extraction technique to recognize handwritten Devanagari script.

Keywords: OCR (Optical character recognition), Feature Extraction, Zoning.

I. INTRODUCTION

Optical character recognition is the recognition of printed or handwritten text by a computer. This involves scanning of the handwritten text, which converts the paper document into an image, and then translation of the text image into character codes. In OCR implementation consists of a number of steps followed by the actual recognition. Recognition of handwritten characters has been good research area for many years because it has many applications in all fields. Devanagari script is the script for writing Hindi language.

Hindi is the official language of India. Recognition of handwritten Optical character is very difficult due to different writing styles of the different person. Due to large number of characters and presence of half characters and some confusing characters makes the recognition process even more complex. In this we take the data from many users and found that the writing style of every user is different. So, the recognition of the characters is very difficult. In this work the objective is to recognise character in Hindi by using some Feature Extraction Technique.

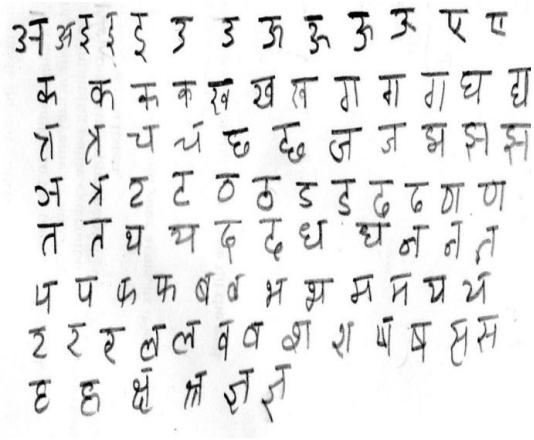

Fig 1: Handwritten character by different users

II. RELATED WORK

[1]. NARESH KUMAR GARG, DR. LAKHWINDER KAUR, DR. MANISH JINDAL" RECOGNITION OF OFFLINE HANDWRITTEN HINDI TEXT ",INTERNATIONAL JOURNAL OF IMAGE PROCESSING , 2013

In this paper, a segmentation based technique is used to recognize the text. This segmentation based text recognition is very efficient procedure. Handwritten Hindi text identification is popular section of research in the area of optical character recognition. Shape features are pulled out from the characters and put into classifier for identification. The offline handwritten text is divided into lines, and also lines are divided in words after that words divided into character concerning acknowledgment. The results achieved with the proposed attribute set by means of classifier is very challenging and demanding. [1]

[2]. Banashree N. P., and R. Vasanta" OCR for Script Identification of Hindi (Devnagari) Numerals using Feature Sub Selection by Means of End-Point with Neuro-Memetic Model ", International Journal of Computer, Control, Quantum and Information Engineering, 2007

This paper suggests a recognition format for handwritten Hindi numerals; most accepted one in Indian subcontinent. Their work emphasized over methods used in attribute pull out which means universal foundation method with the help of end-points information, which is drawn from images of inaccessible numerals. In suggested scheme data sets are put to neurometric flowchart that recognizes the regulation having most sound worth of closely 100 % & overlay follows that the regulation is none except a recognized number. Identification of Indian languages scripts is fascinating problems. There are quite a lot of phenomenon's that deal with dilemma of identification of numbers/alphabets hanging on the genre of grade take off and dissimilar way of extracting them. In OCR, an alphabet or number which will be identified can be machined printed or handwritten lettering or numerals. These aspect vectors are fed to neurometric model that has

DOI 10.17148/IJARCCE.2015.4891

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

been qualified to be acquainted with devanagari number. The prototype of network is being checked over various sets of image of numbers. Experimentation consequences show that identification rate is 92-97 % in contrast to supplementary models. [2]

[3]. Shruti Aggarwal, Dr. Naveen Hemarjani" Offline Handwritten Character Recognition with Devnagari Script", IOSR Journal of Computer Engineering (IOSR-JCE) ,2013

Here in this paper the main focus is on the identification of offline handwritten Hindi characters that can be used in frequent applications like, bill dispensation systems, bank cheques, government records, marketable forms, Signature authentication ,Postcode detection, passport readers, offline document recognition generated by the growing scientific society. In this paper, by the use of template corresponding algorithm devanagari lettering characters are OCR from document images. Handwriting recognition is the capabilities of a processor to take delivery of appreciate understandable handwritten input from sources such as photographs, touch-screens, paper documents and other devices. Written book figure may be perceived "off line" from a fragment of paper by visual scanning Devanagari script has 14 vowels and 33 consonants. Vowels occur either in remoteness or in amalgamation with consonants. Aside from vowels and consonants characters called basic characters, compound characters are there in Devanagari script, that are organized by coupling two or more basic characters. Combined to this in Devanagari script there is a exercise of having twelve forms of modifiers with each for 33 consonants, giving escalation to modified shapes which, hinges on whether the modifier is placed to the left, right, top or bottom of the character. The net upshot is that there are more than a few thousand dissimilar shapes or patterns, which makes Devanagari OCR more complicated to enlarge. [3]

[4]. Vneeta Rani, Dr. Vijay Luxmi" Character Segmentation of Handwritten Document in Devnagari Script", IJCST, 2013

In this paper, there is focal point on the character segmentation of handwritten document in Devanagari script. There are numerous languages bring into play in devanagari script i.e. Hindi, Marathi, Sanskrit & Nepali etc. They have worked merely Hindi text and these are used for varied vowels, consonants and modifiers in the Hindi language.. Segmentation is an significant part of OCR. Segmentation is a process to partition a absolute manuscript into tiny parts like subsection into the lines, the line into words & word into characters. This process has used for removing the tender characters and lines. It has also been used for eliminating the overlayered and discontinuous characters.[4]

[5]. Aditya Raj, Ranjeet Srivastava, Tushar Patnaik, Bhupendra Kumar" A Survey of Feature Extraction and Classification Techniques Used In Character Recognition for Indian Scripts ", International Journal of Engineering and Advanced Technology (IJEAT) , 2013

Feature extraction and categorization are necessary steps of character recognition process disturbing the overall accurateness of the recognition system. The Constitution of India has legendary for 22 languages as representative languages of India. Amongst these most of the recognition investigation effort has been completed for Devanagari, Gurumukhi, Telugu, and Bangla scripts etc. OCR system improvement for Indian script has many functional regions like maintaining manuscript and ancient literatures written in dissimilar Indian scripts and making digital libraries for the documents. In this paper the OCR system which they use produce the more appropriate results for the recognition.[5]

III. PROPOSED METHODOLOGY

Methodology is the systematic, theoretical analysis of the methods applied to a field of study. It comprises the theoretical analysis of the body of methods and principles associated with a branch of knowledge.

In this work the methods used are as follows:-

Optical Character recognition system

OCR System comprises of following steps:

1) **Create Image**

Firstly image is created by scanning the handwritten document and then it is converted into required format like png or jpeg etc.

2) Pre-processing

The scanned image may now contain some form of noise which is removed here. In this work the format wanted is png so after scanning, the jpeg image is converted into png format with the help of image converter. Binarization of the image and segmentation includes line, word and character.

3) Binarization

The image scanned and pre-processed is binarized here means image is converted in the 0 and 1 form. The place where part of image is shown or it is black is displayed as 1 and the part where image is not shown or it is white is displayed as 0.

4) Feature Extraction Technique

This is the most important step of OCR process. As based on the feature or some favourable quality of the image only recognition of handwritten character can be performed. Only based on this property the efficiency of OCR can be taken to next higher level. Some attribute extraction technique are Template matching, Deformable Templates, Zoning, Projection Histogram, Contour Profile, Moments calculation. In this work we are using Zoning feature extraction technique in which 100*100 matrix of image is created, which contains presence of pixels as one and absence of pixels as zero.

ZONING:

The image which consists of the alphabet is fragmented in many overlaying or non-overlaying regions and the density of object pixels in every region is computed. We compute the density with the help of the no of object pixels in every region and then divide it by total no of pixels.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

Density= Number of Object Pixels in Every region / Total Number. Of Pixels.

Zoning attribute is one of the most popular, easy and fast working method to recognise a character. It is accurate technique. It is computed by the density of pixels in all regions so we divide the character sample. Zoning terms is used for divide the trained characters in 100*100 segments by different perspective and store its values in forms of bits such as true or false means Zero or one such as it denotes to black and white. If there is a black part in the segmented 100*100 zone then it will save the value as 1 in matrix and if there a white part then it will store the value 0 in the 100*100 matrix.

There are some steps of Zoning:

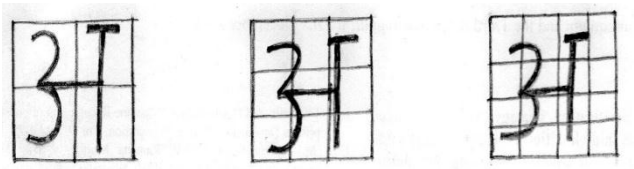

Fig 2: Character Divided by Different Grid Segment

When the input image with multiple characters will be given as input to recognize the handwritten characters, then the zoning technique will recognize the characters and then divide in segments. Then the feature will be extract of input image's character using Zoning and then Support vector machine will classify the characters according to the feature extracted.

5) Classification Technique

Classification is performed after all the stages of OCR. In this based on feature extraction technique classification is performed as to which class the character belongs to. SVM technique is used as a classification technique in this work which classifies the class of the character. Some classification phenomenon used in formerly developed OCR is Neural Network, Support Vector Machine, K-Nearest Neighbours, Bayesian classification and Decision Tree classification.

SUPPORT VECTOR MACHINE:

In machine learning, **support vector machines** (**SVMs**, also **support vector networks**[1]) are supervised learning models with associated learning algorithms that analyse data and recognize patterns, used for classification and regression analysis.

Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier.

An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. [6]

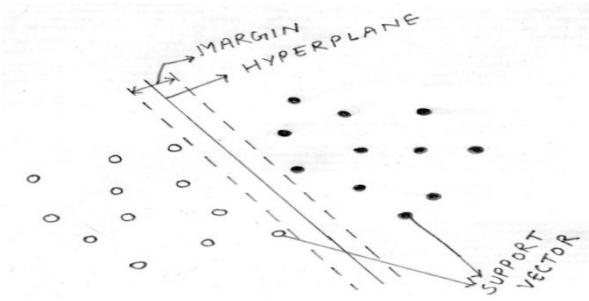

Fig 3: Separation hyper planes [6]

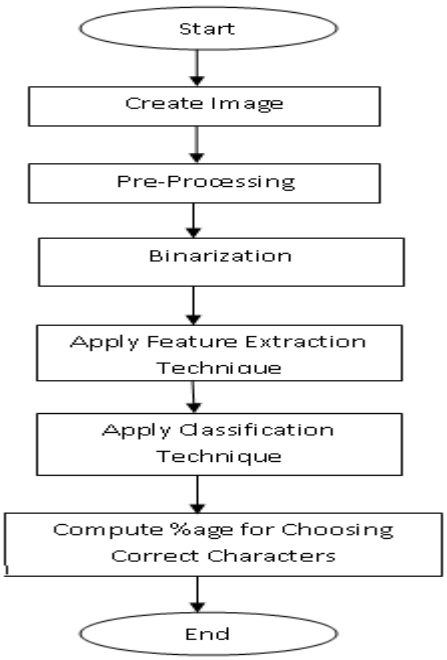

Fig 4 Steps in OCR system

Pattern recognition

If all the persons in the world can write all the letters in same way then there was no problem for the OCR to recognise any handwritten or printed document or image as the OCR would match the testing and training samples and it would be a match which is far more easy process. But the problem lies that all the human beings write in their own styles and handwritings which is a big issue for OCR. So in this OCR is given all the possible samples in which anyone can be able to write the letter suppose A , now the OCR will match the given A letter with all the samples which is a time consuming process. Recognition is done based on patterns or strokes or you can say different writing styles of every person. This OCR-A was used for the bank checks and further it has many applications. The design of the letters are properly and carefully made i.e. their design include every letter was exactly of the same withstand the bangs were anxiously arranged so that every letter could fluently be dignified from the others.

Feature detection

It is also known as feature extraction. This is far more easy process than Pattern recognition as in this there is no need to provide a many samples for recognising a single letter.

DOI 10.17148/IJARCCE.2015.4891

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

Only based on some feature or some important quality like connected components or correlation , the OCR recognises the character. It is also called as intelligent character recognition (ICR), this is to a large extent more popular and easy technique of finding characters.

OBJECTIVES

The overall goal of the work is to extract feature of Handwritten Devanagari script.

a. To study about various phases of OCR such as feature extraction, recognition and classification.
b. To Implement the Zoning Technique for Feature Extraction of Characters i.e. handwritten Devanagari script.
c. To provide the training with Multiple Training Datasets for improving the Recognition Accuracy.
d. To improve the efficiency of OCR.
e. To implement SVM as classification technique.

REQUIREMENTS

a. MATLAB Tool to Apply Character Recognition algorithm.
b. A system with Windows Vista/7 OS
c. Existing algorithm's and approach.
d. Training and testing handwritten samples.

IV. RESULTS

The testing sample provided to the OCR should be in .png format and it should be a binary image which has only two types of values 0 which means background and 1 which means foreground or you can say black and white only. The testing sample must be inscribed in such a way that it is visible easily and the background must be of white colour only. After processing of image the displayed results on the notepad file will determine whether character is recognized properly or not. The matrices of testing as well as training samples are correlated with help of correlation function in MATLAB and after that classification is performed with the help of SVM which tells the class of character to which it belongs.

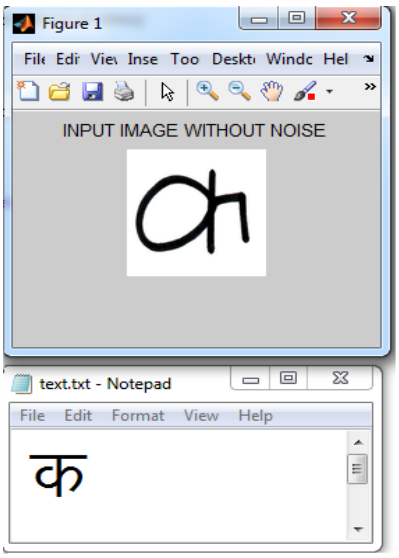

Fig 5 Testing data 1

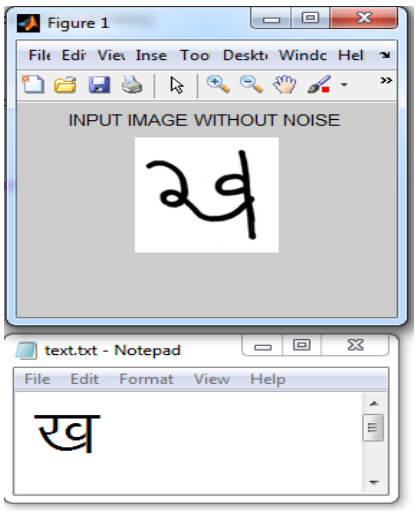

Fig 6 Testing data 2

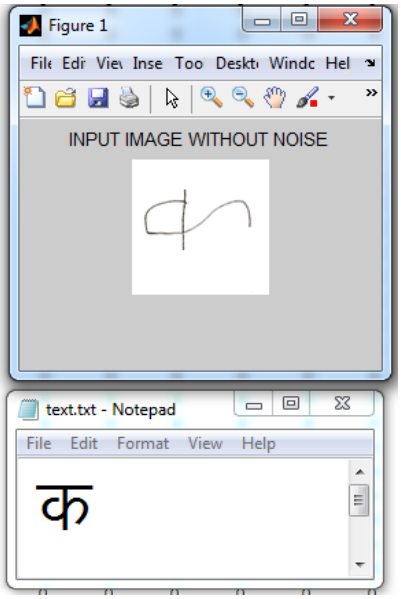

Fig 7 Testing data 3

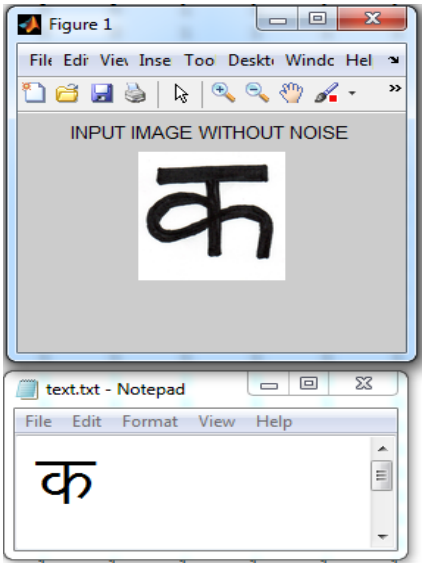

Fig 8 Testing data 4

DOI 10.17148/IJARCCE.2015.4891

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

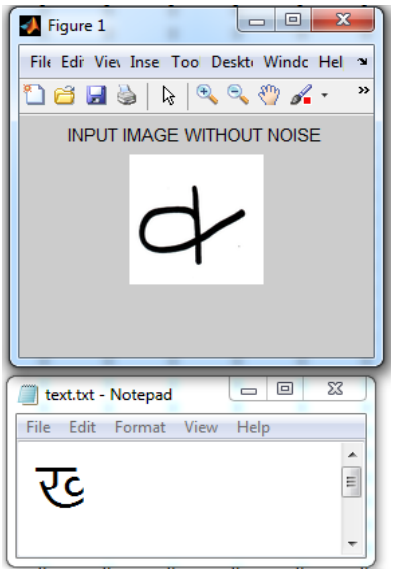

Fig 9 Testing Data 5

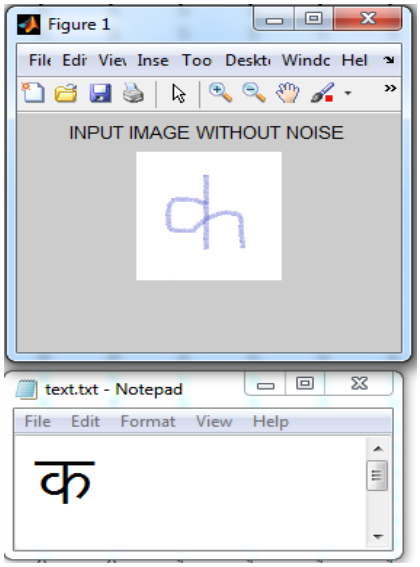

Fig 10 Testing Data 6

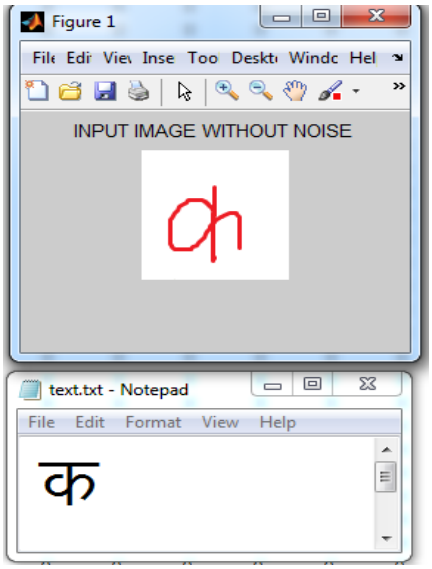

Fig 11 Testing Data 7

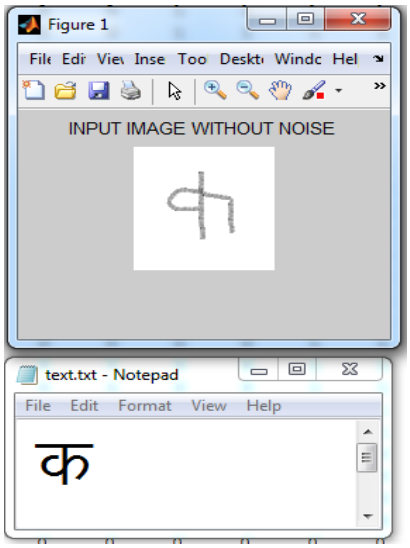

Fig 12 Testing Data 8

Half characters are not recognized by this OCR. Consonants are recognized in any form whether in pencil or marker. Like we can see below in fig 9 that half character is not recognized which is an area of future work. Fig 5 and 6 are inscribed with marker. Fig 7 is inscribed with pencil and Fig 8 is inscribed with dark marker which is recognized. Fig 10, 11, 12 are inscribed with sketch of different colour which are also recognized by this OCR.

Output

Total Testing Image	**Recognized Correctly**	**Unrecognized**
65	50	15

Efficiency

Efficiency=No. of characters recognized correctly / Total testing images* 100

So, Efficiency of OCR = 70% - 80%.

V. CONCLUSION

Our main motive is to create an OCR which can recognize handwritten Hindi character. This OCR can easily recognize all forms of character in different handwritings with different objects such as pen, pencil, marker or sketch even of different colour as well as oriented handwritten Hindi characters and it is an efficient one, but it needs more polishing as there are some problems with it. The problem or difficulty of this OCR is that it does not recognize pure form of consonants or half characters and missing part of Hindi characters, which is one of the main issues and future work must be concentrated on this. This OCR does not recognize words as it is made only for single characters and also there is a difficulty in recognition of characters which are not in black marker. We have recognized distorted vowels as well as consonants of Hindi script. The future work must be done on the problems narrated above. In this work we have developed OCR in MATLAB.

DOI 10.17148/IJARCCE.2015.4891

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

As explained above there are some shortcomings. So the future work must be focused on this modification in existing research:-

a. To recognize pure form of consonants or half characters.
b. To recognize words or combination of characters.
c. To recognize missing part of Hindi characters.

REFERENCES

[1]. Naresh Kumar Garg, Dr. Lakhwinder Kaur, Dr. Manish Jindal" Recognition of Offline Handwritten Hindi Text",International Journal of Image Processing , 2013
[2]. Banashree N. P., and R. Vasanta" OCR for Script Identification of Hindi (Devnagari) Numerals using Feature Sub Selection by Means of End-Point with Neuro-Memetic Model ", International Journal of Computer, Control, Quantum and Information Engineering, 2007
[3]. Shruti Aggarwal, Dr. Naveen Hemarjani" Offline Handwritten Character Recognition with Devnagari Script", IOSR Journal of Computer Engineering (IOSR-JCE) ,2013
[4]. Vneeta Rani, Dr. Vijay Luxmi" Character Segmentation of Handwritten Document in Devnagari Script", IJCST, 2013
[5]. Aditya Raj, Ranjeet Srivastava, Tushar Patnaik, Bhupendra Kumar" A Survey of Feature Extraction and Classification Techniques Used In Character Recognition for Indian Scripts ", International Journal of Engineering and Advanced Technology (IJEAT) , 2013
[6] https://en.wikipedia.org/wiki/Support_vector_machine