ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

A REVIEW: Optimized Hindi Script Recognition using OCR Feature Extraction Technique

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

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

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

ABSTRACT: Feature Extraction is a crucial part of any OCR. Without feature extraction technique no OCR can perform its function correctly. So it's a very necessary as well as important step in recognition of handwritten Hindi script. In this paper the main focus is to extract features of handwritten Hindi script in such a way that it yields a maximum efficient result. Hindi script is one of the famous and most spoken languages of India. Hindi Language consists of vowels, constants and various modifiers. Finding of proper and best fitting Feature Extraction technique is a bit difficult. In this work zoning Feature Extraction technique is used to recognize handwritten Hindi script and SVM is used as classification technique.

KEYWORDS: OCR(Optical characterrecognition), Feature Extraction, Zoning

I. INTRODUCTION

Optical Character Recognition, or OCR, is technology which converts the images handwritten,scanned or printed into character codes or machine encoded text. It is widely used in bank cheques, business documents, number plate recognition, historical book scanning and many others. The image captured by digital camera or any PDF is converted into a suitable form required by the machine. OCR is already being used largely in legal or business or bank work, where checks that once needed so much time can now be proficient in a few seconds. In this OCR technique the shapes of character are noted down by the OCR like the order in which segments are drawn. OCR is also called handwritten character recognition or intelligent character recognition. Image recognition is the most difficult part of OCR as there is a problem of distortion found in images which is not present in PDF files or scanned documents.

Reasons and Applications (usage) of OCR for single character which is preferred in this work
The applications of this specific OCR is to help the beginners of Hindi quickly familiarise themselves with Hindi script. It provides useful learning aid. E.g.: 'Beginner's Hindi script' by Rupert Snell and 'Teach yourself Hindi' by Simon Wightman and Rupert Snell. If and only if OCR is trained with characters then only it will have the knowledge of Hindi characters and have the capability of recognising words in further work. This OCR is used in Devanagari character picker and Devanagari keyboard. It is also used in typing Hindi characters on iPhones and iPods. Its vast application is sunosunao.com which provides training to young children of above 3 years. It is very quick and easy to use in conversion as well as recognition. It is user friendly as it uses GUI.

Working of OCR
Suppose there existsome consonant in Hindi script sayd. We know very well that not all the human beings write it in same way. Even a single person writes din different ways. So it is very difficult and challenging task to recognize characters by OCR as there is not a single way to write it.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

Fig1 LETTER d FORMS

Same is the case with printed or computerized text as there are different fonts available and sizes. The letter d is written in many different appearances by different writers as shown above.Generally speaking, there are many different ways to resolve this issue, we can do it by characters in their entirety, it is basically known as pattern recognition or we can also do it by perceiving the lines and strokes characters are made from which is called as feature detection and identifying them in that way. The other is writing data from various modes such as pencil, sketch and marker.

II. RELATED WORK

[1] DivakarYadav, Sonia Sánchez-Cuadrado and Jorge Morato"Optical Character Recognition for Hindi Language Using a Neural-network Approach", 2013

In this paper, they suggest an OCR for printed Hindi text in Devanagari` script, using Artificial Neural Network (ANN), which enhances its effectiveness. One of the chief reasons for the deprived acknowledgment rate is fault in character segmentation. Hindi is the spoken by a lot of people in India, with more than 300 million users. As there is no severance between the characters of texts printed in Hindi as here is in English, the Optical Character Recognition (OCR) systems urbanized for the Hindi language bear a very pitiable identification rate. Pre-processing, character segmentation, feature extraction, and finally, classification and recognition are the major steps which are followed by a general OCR. The occurrence of touching characters in the scanned documents supplementary set hurdles to the segmentation process, creating a serious difficulty when designing a striking character segmentation technique. The pre-processing tasks measured in the paper are adaptation of grayscaled images to binary images and image rectification. The basic symbols, bagged as the basic unit from the segmentation progression, are accepted by the neural classifier. For progress of the neural classifier, a back-propagation neural network with two hidden layers is used. The classifier is skilled and veteran for printed Hindi texts. In this effort, three characteristic extraction techniques are stated as histogram of projection based on mean distance, histogram of projection based on pixel value, and vertical zero crossing, have been used to recover the pace of detection. These feature extraction techniques are commanding adequate to pull out features of even indistinct characters/symbols. By applying the OCR the outcome of performance of approximately 90% accurate recognition rate is achieved.

[2]. PrabhanjanS, R Dinesh",Handwritten Devanagari Numeral Recognition by Fusion of Classifiers". International Journal of Signal Processing, Image Processing and Pattern Recognition, 2015

The main aim is to Recognize handwritten Devanagari numerals which has many applications in the field of postal address, document processing and so on. Due to its vast applications, many researchers are working towards development of effective and efficient handwritten numeral recognition. In this paper, we have proposed a hybrid method to recognize handwritten devanagari numerals. The proposed method uses, stacking approach to fuse the confidence scores from four different classifiers viz., Naïve Bayes (NB), Instance Based Learner (IBK), Random Forest (RF), Sequential Minimal Optimization (SMO). Also, the proposed method extracts both local and global features from the handwritten numerals.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

[3].Nitin Mishra , C. Vasantha Lakshmi, Sarika Singh" Shirorekha Chopping Integrated Tesseract OCR Engine for Enhanced Hindi Language Recognition", International Journal of Computer Applications ,2012
The Hindi language recognition accurateness is pretty squat yet for computerized script, but the joint alphabet groupings of Hindi Language are not simply distinguishable due to fractional overlapping. The projected come up to solve the hurdle, so that Hindi joint alphabets can readily be divided and identified by help of Tesseract OCR machine. The research shows an absolute methodology to develop The Hindi Language Recognition exactness. Tesseract OCR machine is the popular well-organized open foundation OCR engines presently obtainable. Recently, Tesseract 3.01 is susceptible of sensing devanagari script yet it quiet desires some augmentation to recover the presentation. This paper also presents assessment with additional Devanagari OCR engines presented on the origin of sensing efficiency, proceeding duration, change of font and length of database.

[4]. YogendraBagoriya,Nisha Sharma" FONT TYPE IDENTIFICATION OF HINDI PRINTED DOCUMENT",IJRET: International Journal of Research in Engineering and Technology,2014
This paper presents an attempt for typeset identification of 5 hindi font types such as the mentioned below MarathiVakr, Shusha05,DevanagariNew,Shusha02,Premchandkikahaniyan. Ocular font type identification is one of the significant but frequently abandoned troubles. Foremost utilization of dissimilar font styles in documents is to prominence on a few part of document in order that any person who reads can perceive them with no trouble. In a document, font type changes possibly will transpire at fussy points like titles, indexes, references, etc. They perhaps had done by choosing\ an additional typeface. Due to the use of divergent font types OCR engine is not able to distinguish the characters appropriately and accurateness of the system may diminish. It is not potential to duplicate old documents without acquaintance of font type. Most important applications of font type classification are, improving the recognition rate of an OCR engine, document facsimile, formation of new font types, Document indexing and information repossession.

III. PROPOSED METHODOLOGY

STEPS OF OCR:

1. **Pre-processing:** In pre-processing the noise is removed from the scanned image and the quality of image is improved. Binarization of image is performed by converting it in 0 or 1 form like if a part of image is present in segmented image then it shows 1 else shows 0.
2. **Feature Extraction:** The purpose of this stage is to extract the important features of samples. In image processing, Feature extraction is the most important part. It starts from initial training data based on which it builds various features which are informative, non-redundant and leads to better human interpretation. [5]

Following are some of the feature extraction techniques which have been studied for this work:-

Shadow features: In this technique the image is divided into 8 octants and for each octant there exist 2 shadows on both the perpendicular sides of that octant. So total there will be 16 shadows of all 8 octants. Shadow is the projection of image on perpendicular side.[6]

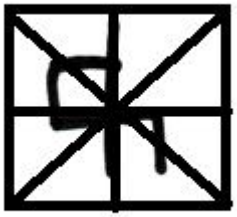

Fig 2 Shadow features

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

Chain code features: In the given binary image the contour points are found first. We consider a 3*3 window surrounding the object points of image. The contour is found and then a contour representation called chain coding takes place. Each pixel of contour is assigned a code which is different from other to indicate the direction of neighbourhood pixel that belong to contour in the given direction.[6]

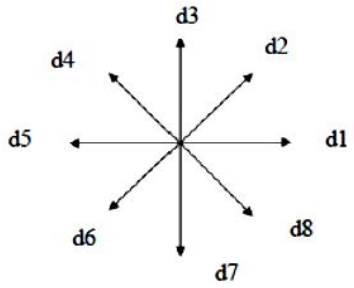

Fig 3 Chain code features

Distance Profile Features: Distance of No. of pixels from Boundary Box of characters is calculated from all the 4 directions as follows :

a) Right
b) Left
c) Bottom
d) Top

Projection Histogram Features: Calculate no. of pixels in specified direction which can be

a) Horizontal
b) Vertical
c) Left Diagonal
d) Right Diagonal

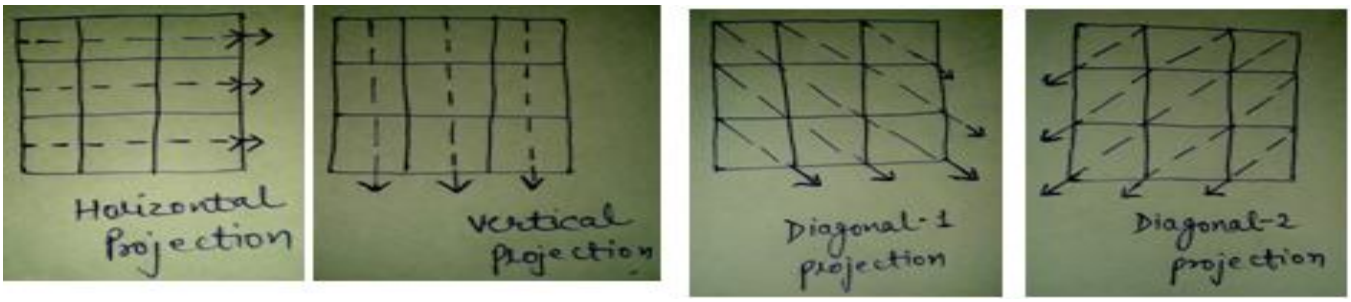

Fig 4 Projection histogram features

Zoning: In this the feature extraction technique used is zoning in which the character is divided into 100*100 zones by different perspective and store its values in forms of bits such 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. The frame containing the character is divided into several overlapping or non-overlapping zones and the densities of object pixels in each zone are calculated. Density is calculated by finding the number of object pixels in each zone and dividing it by total number of pixels.

Density= No. of Object Pixels in Each Zone/Total No. of Pixels

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

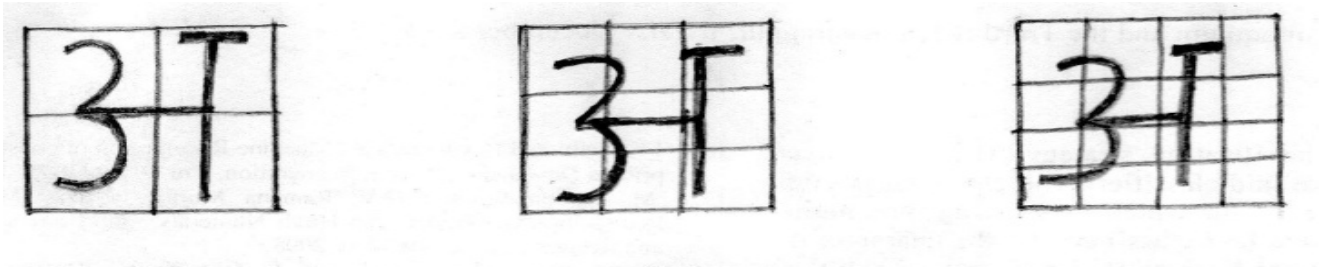

Fig 5 Character Divided by Different Grid Segment

3. **Classification:** Classification refers to categorization the process in which character or images are recognized and understood. Classification is the process which is performed after feature extraction is done. Based on feature extraction technique results classification is performed.

Following are some of the classification techniques which have been studied for this work:-

K-Nearest Neighbour (KNN): K Nearest Neighbour is one of the classification techniques which is very easy and simple to be understood, and incredibly very useful .It is a widely used technique. Initially all the training is given to the OCR and after that when the testing of samples starts then OCR finds the nearest neighbour which matches closely to the testing sample from the training given with the help of nearest neighbour algorithm. There is a minimal training phase in this but a valuable learning or can say testing process take place with the help of KNN.

Artificial Neural Network (ANN): An Artificial Neural is information dispensation standard that is considered to model the way in which the common sense carry outs a scrupulous task or purpose of interest. The network is usually put into operation by make use electronic components or is imitation in software on a digital computer. A neural network is an especially parallel distributed mainframe made up of processing units which has a usual propensity for carrying experiential acquaintance and making it obtainable for use. [1]

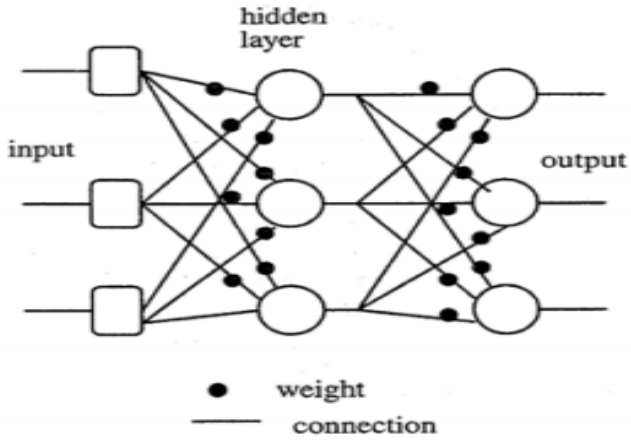

Fig 6 Neural Network Model

Neural Network (NN): Neural networks, have outstanding capability to draw meaning out of imprecise data, are an excellent solution to the classification stage of OCR. They can be used to obtain patterns and discover trends that are too hard to be noticed by either human beings or other computer or machinery techniques. Basic practice involves following steps: Training and Testing. [1]

Training: In this step training is given to network of various handwritten samples and a database is created. Here a network is taught of how to respond to a given pattern or sample.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

Testing: After the training step is complete, the testing is done. Various samples are tested based on the training provided to the network.

Support Vector Machine (SVM): In this work SVM is used as classification technique. SVM has many applications in the field of image processing, biometrics and categorization. SVM is widely used now days by scientific groups as it is very efficient and user friendly. SVM works by making a hyper plane between two classes if it is a linear classification but if it is non-linear classification which consists of more than 2 classes then hyper plane is formed between 1 to all means 1 class is in one side and rest all similar classes on other side of hyper plane. There can also be many hyper planes in case of non-linear classification. If such hyperspace exist, it is clear that it make available the most excellent disconnection margin in the middle of two classes and it is called as the maxim-margin hyperspace and such a linear distinguisher is known as the maxim margin distinguisher.[7]

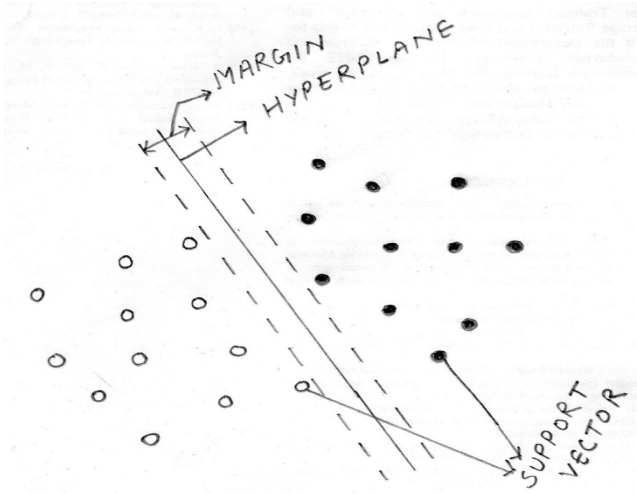

Fig 7 Separation hyper planes [7]

Fig 8 Steps in OCR system

Objectives

a. To study about various phases of OCR such as pre-processing, feature extraction and classification.
b. To Implement the Zoning Technique for Feature Extraction of handwritten Hindi Characters
c. To provide the training with Multiple Training Datasets for improving the Recognition Accuracy.
d. To find out efficiency of existing techniques and to improve it.
e. To apply SVM as classification technique.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

IV. CONCLUSION

Various papers have been studied and it is seen that work on various techniques have been done. Like there are different feature extraction techniques for eg. Chain code features, Template matching, Deformable Templates, Projection Histogram features, Contour Profile, Moments calculation, Intersection features, Shadow features, Projection features, Histogram features, Gradient features, Curvature features, Distance profile features, Zoning features and so many more. All these techniques have been carefully understood but it is found that out of all Zoning is the best for this work. So Zoning feature extraction technique is opted as the Feature extraction technique to be used in this work. Then comes the classification technique. Here also classification techniques like Neural Network (NN), K-Nearest Neighbours (KNN), Artificial Neural Network (ANN), Bayesian classification, Decision Tree classification, Projection distance (PD), Subspace method (SM), Linear discriminant function (LMD), Support vector machine (SVM), Modified quadratic discriminant function (MQDF), Mirror image learning (MIL), Euclidean distance (ED), Modified projection distance (MPD), Compound projection distance (CPD), Compound modified quadratic discriminant function (CMQDF) and many more. But in this work SVM is opted as classification technique. So the combination of Zoning + SVM is used in this work to improve the efficiency of OCR. It is tried to get a better OCR for character recognition with the help of Zoning and SVM and to solve the previous problems obtained.

REFERENCES

[1] DivakarYadav, Sonia Sánchez-Cuadrado and Jorge Morato"Optical Character Recognition for Hindi Language Using a Neural-network Approach", 2013
[2]. Prabhanjan S, R Dinesh",Handwritten Devanagari Numeral Recognition by Fusion of Classifiers". International Journal of Signal Processing, Image Processing and Pattern Recognition, 2015
[3]. Nitin Mishra , C. Vasantha Lakshmi, Sarika Singh" Shirorekha Chopping Integrated Tesseract OCR Engine for Enhanced Hindi Language Recognition", International Journal of Computer Applications ,2012
[4]. YogendraBagoriya ,Nisha Sharma" FONT TYPE IDENTIFICATION OF HINDI PRINTED DOCUMENT",IJRET: International Journal of Research in Engineering and Technology,2014
[5] https://en.m.wikipedia.org/wiki/Feature_extraction.
[6] SandhyaArora, DebotoshBhattacharjee, MitaNasipuri, Dipak Kumar Basu, MahantapasKundu"Combining Multiple feature Extraction techniques for Handwritten1Devnagari Character Recognition,2008
[7]www.intechopen.com.