International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-2, Issue-2, March-2014

Handwritten Devnagari Optical Character Recognition

Prasad Chavan, Suyog Sankpal, Akshay Sonawane, Shahid Shaikh, Prof. Anup Raut

Abstract -
Handwritten Devanagari character recognition is the ability of a computer to receive and interpret handwritten input from sources such as paper documents, photographs, touch-screens and other devices. Handwritten Devanagari Characters are more complex for recognition than corresponding English characters due to many possible variations in order, number, direction and shape of the constituent strokes. The main purpose of this project is to introduce a new method for recognition of handwritten Devanagari characters using Segmentation, Image Processing and Artificial Intelligence. The whole process of recognition includes two phases- segmentation of characters into line, word and characters and then recognition through feed-forward neural network.

Keywords- **Segmentation, OCR, Binarization, Erosion, Dilation, cropping.**

I. Introduction

Optical Character Recognition (OCR) is a program that translates scanned or printed image document into a text document. Once it is translated into text, it can be stored in ASCII or UNICODE format. Handwritten or printed character recognition is an i mportant filed of Optical Character Recognition (OCR). The recognition of handwritten text in scripts is one of the major areas of research. A lot of research is done in the past on segmentation of text. Handwritten character recognition is one of the benchmark problems in artificial intelligence research. Recognition of handwritten characters is a problem that at first seems simple, but is an extremely difficult task to program a computer to perform it.

Manuscript received March 04, 2014
Shahid Shaikh, 4th Year Undergraduate Student, Department of CSE, TCOER, Pune, India (Email-shahidsk444@gmail.com)
Suyog Sankpal, 4th Year Undergraduate Student, Department of CSE, TCOER, Pune, India (Email-suyog5835@gmail.com)
Akshay Sonawane, 4th Year Undergraduate Student, Department of CSE, TCOER, Pune, India (Email-s23aksh@gmail.com)
Prasad Chavan, 4th Year Undergraduate Student, Department of CSE,TCOER,Pune,India (Email-chavan.prasad7001@gmail.com)

Automated character recognition is of vital importance in many industries, such as banking and shipping, in TAHASIL office etc.

One of the major requirements of offline character recognition is the individual character segmentation, which involves the segmentation of individual character patterns into meaningful segments. This separation will play important role in the recognition & will increase the quality of OCR system. [1]

II. Devnagari Script

Devnagari is the script used by Sanskrit, Hindi, Marathi and Nepali. It is an alphabetic script. There is no concept of upper and lower case in Devnagari as in English. Hindi is the world's third most commonly used language after Chinese and English.

India is a multi-lingual and multi-script country comprising of eighteen official languages. One of the defining aspects of Indian script is the repertoire of sounds it has to support. There is typically a letter for each of the phonemes in Indian languages because of which the alphabet set tends to be quite large. Most of the Indian languages originated from Bramhi script. These scripts are used for two distinct major linguistic groups, Indo-European languages in the north, and Dravidian languages in the south.

Devanagari is the most popular script in India. It has 13 vowels and 34 consonants. Sometimes two or more consonants can combine and take new shapes. These new shape clusters are known as *compound characters.* These types of characters namely basic characters, compound characters and modifiers are present not only in Devanagari but also in other scripts. Hindi, the national language of India, is written in the Devanagari script. Devanagari is also used for writing Marathi, Sanskrit and Nepali. Moreover, Hindi is the third most popular language in the world.

All the characters have a horizontal line in the upper part, known as *Shirorekha*or headline. No English character has such a characteristic and so it can be taken as a distinguishable feature to extract

English from these scripts. In continuous handwriting, from left to right direction, the shirorekha of one character joins with the shirorekha of the previous or next character of the same word. In this fashion, multiple characters and modified shapes in a word appear as a single connected component joined through the common shirorekha. All the characters and modified shapes in a word appear to hang from the hypothetical shirorekha of the word. Also in

In such scripts, a text word may be partitioned into three zones. The upper zone denotes the portion above the headline; the middle zone covers the portion of basic and compound characters below the headline and the lower zone that may contain some vowel and consonant modifiers. The imaginary line separating the middle and lower zone may be called the base line.

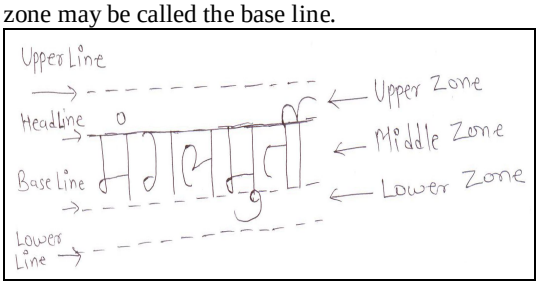

Fig 1 Three strips of a Devanagari word [1]

Three classes of core characters based on the coverage of core strip

1. Full Box Characters
2. Upper Half Box Characters
3. Lower Half Box Characters

Three classes of Full Box Characters based on presence and position of vertical bar

1. End Bar Characters
2. Mid Bar Characters
3. Non-Bar Characters

Two Classes of End Bar Characters based on the Joining Pattern of the Character with the header line

1. Characters with more than one junction point with header line.
2. Characters with only one junction point with header line. [1]

Handwritten character recognition is an important field of Optical Character Recognition (OCR). The objective of OCR is automatic reading of optically sensed document text materials to translate human readable characters to machine readable codes. A good survey about the OCR is given in Research in OCR is popular for its various application potentials in banks, library automation post-offices and defense organizations. [4]

III. Segmentation

Segmentation is a technique which partitions handwritten Devanagari words into individual characters. Since recognition heavily relies on isolated characters, segmentation is a critical step for character recognition because better is the segmentation, lesser is the ambiguity encountered in recognition of candidate characters of word pieces. Handwritten character segmentation is difficult task because of different writing style of person. The handwritten characters do not have a fixed size and shape. So they are quite different from the printed characters. In the case of printed characters, vertical bar occupies a single column whereas handwritten characters might occupy more than one column. Moreover the header line is never straight. Position of the header line in all the Devanagari word is not same. So it can be removed from all the characters at the time of preprocessing. The position of the header line in printed words of Indian script is found using the horizontal pixel projection profiles.

This technique does not work for the handwritten words in which the header line covers multiple rows instead of a single row as in the printed words. This is true of handwritten characters. A selective algorithm is developed for the identification and removal of header line and for further segmentation from the handwritten Devanagari characters. The algorithm takes care of slant in the header line as well as end bar lines. Proposed system gives promising results for printed as well as handwritten text.

Optical Character Recognition deals with the problem of recognizing optical characters. Optical recognition can be performed off-line after the writing or printing has been done, as divergent to on- line recognition where the computer recognizes the characters as they are drawn. Both hand written and printed characters may be recognized, but the feat is directly dependent upon the quality of the input documents. Optical Character Recognition (OCR) system is efficiently developed for character recognition of Non-Indian languages, as the complexity of characters is less as compare to Indian script. Efficient Indian language OCR basically depends upon the preprocessing step(charactersegmentation) forbetterrecognition of fused or conjunct characters. Therefore, overall success rate and accuracy of an Indian script OCR system depends upon on the proper segmentation of characters.

IV. Method

The following diagram illustrates the architecture of our proposed system:

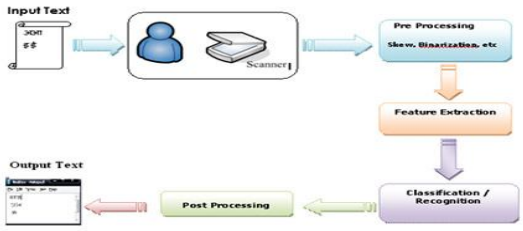

Fig 2 Architecture of Proposed System [1]

The following flowchart give us an idea about steps in our proposed system:

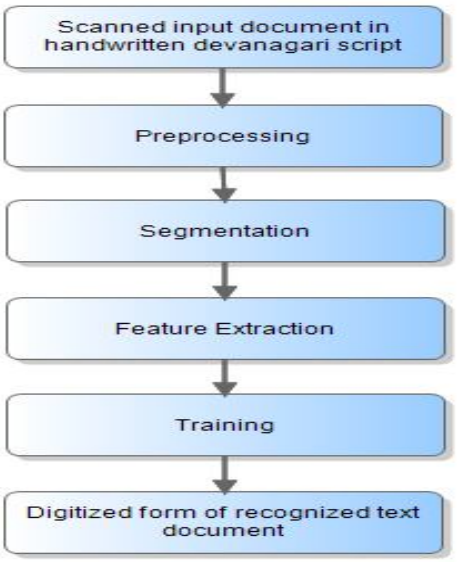

Fig 3 Flowchart for proposed system [1]

1. Scanning:

- Using Devanagariscript, write a text on A4 page
- Scan the page using scanner
- Save the scanned document image in JPG format

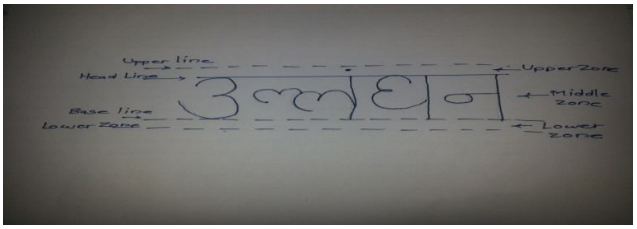

Fig 4: Handwritten fused word input

2. Preprocessing:

In pre-processing the following 3 steps are mainly involved:

- Binarization(SauvolasBinarization Algorithm)
- Skew Correction (Dilate and Thin Approach)
- Hough transform

2.1 Binarization:

In a binary image, each pixel assumes one of only two discrete values: 1 or 0. We binarize the image to separate the background and foreground. The handwriting to be recog-nized is present in the foreground. Also the foreground is represented with 1 as most functions in the programming language consider it to be so. A binary image is stored as a logical array which needs less number of storage space. By convention, this documentation uses the variable name BW to refer to binary images.

The input image is first binarized and complemented so that the background is black and foreground i.e. the words are white and are represented as 1's. Connected components are found in the binary image to isolate each word separately and segmentation process has to concentrate on a smaller regions.For binarization of input image we used SauvolasBinarization Algorithm is as follows:

2.2 Estimate Skew Angle:

- It is unavoidable during scanning or writing on plain paper
- Skew is nothing but the deviation from the x-axis
- Bounded box is followed by selected components.
- Dilate horizontally
- Apply thinning algorithm
- Apply Hough transform to get Skew present in image

2.2.1Rotation Invariant Rule Based Thinning:

- Thins symbols to their central lines means shape of the symbol is preserved.
- It uses 20 rules
- This is iterative method
- Repeated until no further changes occur
- On various conditions it is decided to delete the pixel or not.

Few Thinning Rules:

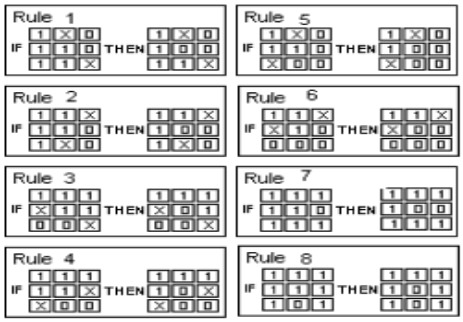

Fig 5: Thinning rules [1]

2.3 Hough Transform:

- Preferred to find lines
- Here each point votes for every line it could be on. The lines with most votes win
- Equation of line is y=mx+c
- Now check this line with the x-axis and calculate the skew present in the image. [1]

3.Segmentation:

The proposed system deals with segmentation of modifiers in upper zone called top modifiers; segmentation of modifiers in lower zone called lower modifiers and fused characters. Thesystem executes the following steps on scanned text.

1. Segmentation of header line
2. Segmentation of the top modifiers
3. Segmentation of the bottom modifiers
4. Segmentation of fused characters

3.1. Segmentation of Header Line:

Header line is the most prominent part of the word image that glues all the characters in the image. Once the header line is separated, the word image gets divided into two parts. One comprising of the top strip and the other comprising of middle and the bottom strip. The top strip contains top modifiers, middle stream contains the characters some of which may befused and the bottom strip contains the lower modifiers. For header line segmentation the proposed system deploys the morphological operations of image processing .For segmentingthe header line the proposed system carries out the following steps:

Step 1 : Erosion
Step 2 : Dilation
Step 3 : Cropping
Step 4 : Object recognition
Step 5 : Header line detection
Step 6 : Accuracy correction

(i) Erosion:

The input word image is eroded with the following structuring element to obtain the processed image. The processed image obtained is sharper. The structuring element used to carry out the erosion is as follows.

1 1 1 1
1 1 1 1

Fig 6: Original image [2]

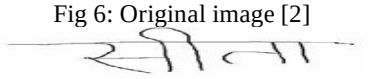

Fig. 7: Eroded image [2]

(ii) Dilation:

The structuring element used to carry out the dilation is chosen in such a way that it is sensitive to the shape of header line in the word image. Processed image obtained as a result of the dilation contains the components that correspond to horizontal line. The structuring element designed for carrying out the dilation is as follows.

1 1 1 1

(iii) Object Detection:

In this step the proposed system makes use of Contour Tracing algorithm which takes a point on the contour of an object. It then returns all the pixels in its neighbourhood of that pixel which comprises of the object. Multiple objects that are present in the dilated image are thus detected and pixels comprising these objects are stored.

(iv)Header line Detection

From the set of objects which are detected in step 4, the object having max size is detected, that object corresponds to the header line.

3.3.2. Segmenting the Top Modifiers

For segmentation of Top Modifiers the proposed system makes use of the Moore Neighbor tracing algorithm. The algorithm is as follows:

Input to the algorithm:
Object T,
containing a connected component P of black cells.
Output from the algorithm:
A sequence B (b1, b2, bk) of boundary pixels.
N (a) be the neighborhood of pixel a.
p denotes the current boundary pixel.
c denotes the current pixel under consideration.
Begin
Set B to be empty.
From bottom to top and left to right scan the cells of T until a black pixel, s, of P is found.
Insert s in B.
Set the current boundary point p to s.
Move back to the pixel from which s was entered.
Set c to be the next clockwise pixel in N (p).
While c not equal to s do

If c is black
insert c in B
set p=c
backtrack (move the current pixel c to the pixel from which p was entered) else
advance the current pixel c to the next clockwise pixel in N(p) is end While
End

	P1	P2	P3	
	P8	P	P4	
	P7	P6	P5	

Fig. 8: Path traced around the current boundary pixel[2]

3.3.3. Segmenting lower modifiers from characters

For segmenting the lower modifier, the point where the lower modifier touches the character is determined by the system. Once the segmentation point is determined, the character image is cropped below the segmentation point to separate the lower modifier from the character.

Determining the segmentation point

For finding the segmentation point the proposed system uses Highlighted Point Detection, Vertical Line Detection and Bottom modifier identifier algorithms.

(a). Highlighted Point Detection Algorithm

This algorithm takes the character image as input carries out erosion to get the eroded image. Structuring element used for carrying out the erosion is of dimension 8X1 .The eroded character image is further dilated to get the dilated image. The structuring element used tocarry out dilation is of dimension 8X4.The dilated image is further dilated; the structuring element used for carrying out the dilation is of dimension 2X2. Next, the object detection algorithm is used to extract the various object from the image .One of the object that may corresponds to the segmentation point is returned by the procedure.

(b). Vertical Line Detection Algorithm

This algorithm takes the character image as input, carries out erosion to get the eroded image. Structuring element used for carrying out the erosion is of dimension 1 X 8.The eroded element is further dilated to get the dilated image. The structuring element used for carrying out the dilation is of dimension 4X8.The dilated image is further eroded, the structuring element used for carrying out the erosion is of dimension 4X4.Next, the object detection algorithm is used to extract the various objects from the image. One of the objects that may correspond to the segmentation point is returned by the procedure.

(c). Bottom Modifier identifier

Highlighted Point Detection algorithm as well as Vertical Line Detection algorithm return the point, from which one of the point is the final segmentation point. For detecting the coordinates of final segmentation point, both the points are examined; the point where y coordinates maximizes the final segmentation point.

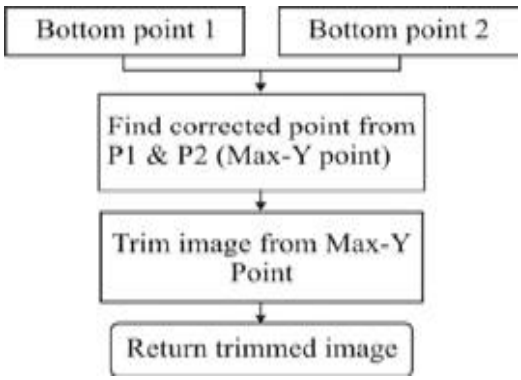

Fig. 9: Flow chart for Bottom modifier identifier [2]

(d). Separating the lower modifier

Once the segmentation point is determined the character image is trimmed from the segmentation point to separate the lower modifier from the character.

3.3.4**Segmentation of fused characters**

(i). Extraction of left consonant

Proposed system carries out segmentation of fused character in to left and right consonant .Left consonant is the half consonant and right consonant is the full consonant .The proposed system scans the character image vertically till a column with one pixel thick intensity is found. Keeping y constant the proposed system checks the consecutive columns till column with more than 3 pixels is found. System stores the start column and end column position. If the column width calculated as width between start and end column is more than 65% of the average width, the portion between the two columns is considered as the left consonant of the composite character. If the width between start and the end column is less than 65% of the average width the process is repeated till next three pixel thick columns is found.

(ii). Extraction of right consonant

From the end column determined in the previous step to the end of the character image, entire part is enclosed in a rectangular box. The width of the average width character enclosed in rectangular box is categorized as the right consonant. [2]

4. Feature Extraction:(Pixel density):

1.Takea character image from segmentation
2.Cropthe image & resize it
3. Represent image into 5 by 7 grids of Boolean values to represent its unique value.

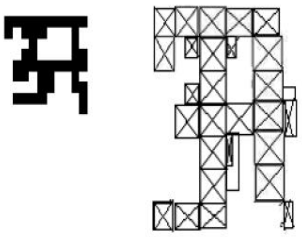

Fig 10: Feature extraction [1]

V. Training

Training stops when any of these conditions occurs:

1. The maximum number of epochs (repetitions) is reached.
2. Performance is minimized to the goal [1]

VI. Conclusion

This paper presents a system for handwritten devnagri character recognition for Devnagri script. A huge character dataset is collected from various writers and used for database creation for neural network training and testing. The recognition of characters is done using multistage multi-feature hybrid recognition scheme.

ACKNOWLEDGEMENT

Authors are thankful to the Department of Computer Engg, Trinity College of Engineering, Pune, Maharashtra, India, for providing the necessary facilities for carrying out this work. Authors gratefully acknowledge the support given by University Of Pune, Pune, Maharashtra, India, for carrying out this research work. Authors also grateful to Prof. Anup Raut, Trinity College of Engineering, Pune, Maharashtra for their time to time guidance and support towards carrying out the research work. Authors are also thankful to Mrs.Sonali Pathare and Ms.Sampada Pingale for influencing us to take forth this research concept and work towards its implementaion and also the anonymous reviewers for their valuable suggestions towards improving the paper.

REFERENCES

[1] N.ARICA, F.T.Y. VURAL, "AN OVERVIEW OF CHARACTER RECOGNITION FOCUSED ON OFFLINE HANDWRITING", IEEE TRANS ON SYSTEM ,MAN,CYBERNATICS-PARTC, VOL 31,N0.2(2001)

[2] OVIND TRIER, ANIL JAIN AND TORFINN TAXT," A FEATURE EXTRACTION METHODS FOR CHARACTER RECOGNITION-A SURVEY", PATTERN RECOGNITION, VOL 29, NO-4, AND PP 641-662, 1996.

[3] U.PAL AND B.B. CHAUDHURI," AN IMPROVED DOCUMENT SKEW ANGLE ESTIMATION TECHNIQUES", PATTERN RECOGNITION LETTERS 17:899-904, 1996.[3] B.B. CHAUDHRURI AND U.PAL, "A COMPLETE PRINTED OCR", PATTERN RECOGNITION,(5):531-549, 1998.

[4] REJEAN PLAMONDON AND SARGUR N. SRIHARI, "ON-LINE AND OFF-LINE HANDWRITTEN RECOGNITION" A COMPREHENSIVE SURVEY", IEEE PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL 22, NO. 1, JANUARY 2000.

[5] U. PAL, T. WAKABAYASHI, F. KIMURA, "COMPARATIVE STUDY OF DEVNAGARI HANDWRITTEN CHARACTER RECOGNITION USING DIFFERENT FEATURE AND CLASSIFIERS", 10TH INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION 2009.

[6] ZHIYI ZHANG, LIANWEN JIN, KAI DING, XUE GAO,"CHARACTERSIFT: A NOVEL FEATURE FOR OFFLINE HANDWRITTEN CHINESE CHARACTER RECOGNITION" 10TH INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION, 2009

[7] T.V.ASWIN AND P S SASTRY, "A FONT AND SIZE-INDEPENDENT OCR SYSTEM FOR PRINTED KANNADA DOCUMENTS USING SUPPORT VECTOR MACHINES", SADHANA VOL.27.PART I, PP.35-58, FEBRUARY 2002.

[8] T.V.ASWIN, "A FONT INDEPENDENT OCR FOR PRINTED KANNADA USING SVM", MASTER THESIS, INDIAN INSTITUTE OF SCIENCE, BANGALORE, 2000.

[9] VEENA BANSAL AND R. M. K. SINHA, "A DEVANAGARI OCR AND A BRIEF OVERVIEW OF OCR RESEARCH FOR INDIAN SCRIPTS", PROCEEDINGS OF STRANS01, IIT KANPUR 2001

Prasad Chavan, 4th Year Undergraduate Student, Department of CSE, TCOER, Pune, India
Email-chavan.prasad7001@gmail.com

Shahid Shaikh, 4th Year Undergraduate Student, Department of CSE, TCOER, Pune, India
Email-shahidsk444@gmail.com

Akshay Sonawane, 4th Year Undergraduate Student, Department of CSE, TCOER, Pune, India
Email-s23aksh@gmail.com

Suyog Sankpal, 4th Year Undergraduate Student, Department of CSE, TCOER, Pune, India
Email-suyog5835@gmail.com