Handwritten Devanagari Character Recognition

Akhil Deshmukh, Rahul Meshram, Sachin Kendre, Kunal Shah
Department of Computer Engineering
Sinhgad Institute of Technology (SIT)
Lonavala
University of Pune, India

***Abstract*—Recognition of Devanagari character consists of Image correction, segmentation and character recognition. Image correction digitizes the input characters making it available for further processing. Principle component analysis is used to discover the hidden and unclear part and segmentation separates individual characters to identify each character. The most crucial part of any character recognition system is the process of segmentation as characters are recognized individually. The result of recognition is dependent on the accuracy of segmentation. For extraction and recognition we used Eigen space method which uses Gerschgorin's theorem for comparison. Handwritten Devanagari script is nowadays a popular topic for researchers as less work is done on this topic. Handwritten Devanagari characters are difficult to recognize due to the presence of header line and various modifiers. Recognition of fused characters is also a major concern for researchers as fused character is treated as a single character resulting in an error.**

***Key Words*—Eigen space, Eigen values, PCA, Adaptive thresholding, Character recognition**.

I.INTRODUCTION

There are many algorithms to determine the English character recognition but very less algorithms for Hindi character recognition. About 600 million people in India use Hindi language for their communication. Many researchers of India are busy developing this software so that Hindi document can be recognized by computer. The upper line is called Headline or "shirorekha" which runs thorough out the script. Each word may consist of upper and lower modifier or "maatra" which makes it difficult for the developer to develop an algorithm with good efficiency. The characters which are half joint with the other consonants are called conjuncts. Handwritten scripts do not have strict rules and hence irregularities, writing style, skew and fluctuating line style appears. Hence segmentation becomes difficult. Devanagri character recognition is the recent topic for researchers due to its limited scope for implementation and digitizing the rare handwritten documents. Unlike English characters, Devanagri characters are very difficult to separate and identify individually. For the identification of such words we are proposing a method that recognizes handwritten Hindi letters.

CHARACTERISTICS OF DEVANAGARI SCRIPT

1. Preprocessing
2. Segmentation
3. Feature extraction
4. Recognition
5. Post processing

Most of the characters in Devanagari scripts are made by curves, holes, and also strokes. Devanagari has 11 vowels and 34 consonants. Besides these consonants, there are set of vowels modifiers called MATRA, pure consonant also called half-characters. The horizontal line is called as SHIROREKHA**.** By using these line top-modifiers, middle zone and lower modifiers are separated.

Consonants + Vowels modifiers =Complete Character

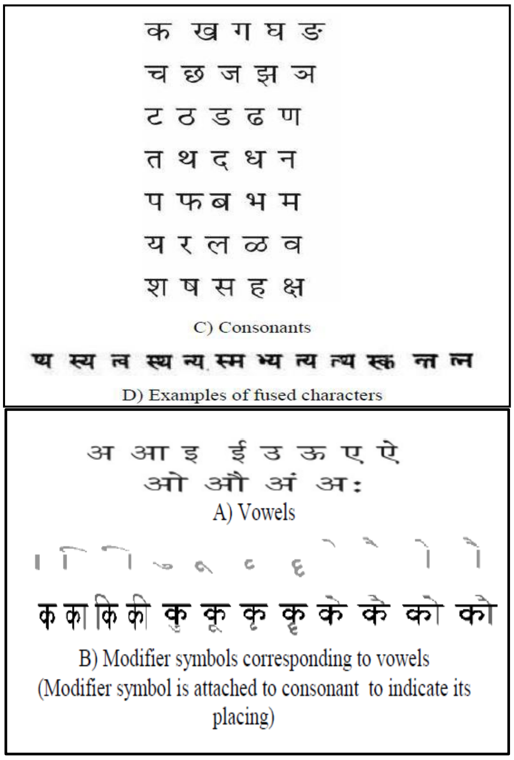

क ख ग घ ङ
च छ ज झ ञ
ट ठ ड ढ ण
त थ द ध न
प फ ब भ म
य र ल ळ व
श ष स ह क्ष
C) Consonants
D) Examples of fused characters
अ आ इ ई उ ऊ ए ऐ
ओ औ अं अः
A) Vowels
क का कि की कु कू कृ कॄ के कै को कौ
B) Modifier symbols corresponding to vowels
(Modifier symbol is attached to consonant to indicate its placing)

Fig 1: Vowels, Modifiers, Consonants, Fused Characters

II. LITERATURE SURVEY

Method 1: "Novel Approach to segmentation of Handwritten Devnagri word"

According to [1] Ms.Vandana and Latish Malik an attempt to segment the handwritten Devanagari words is made. The proposed system carries out segmentation in hierarchical order. The system deploys the morphological operations of image processing for segmentation. Neighbourhood tracing algorithm is used for finding the segmented objects in the specific zones that correspond to constituent symbols of the Devanagari script. 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.

Method 2:"Segmentation of Touching and fused Devanagari characters"

In this paper [2] Ms.Veena Bansal and R.M.K Sinha focuses on recognition of fused characters using pixel intensity. Left consonant is the half consonant and right consonant is the full consonant. Thus keeping y constant the proposed system checks the consecutive columns till column with more than three pixel is found. If the column width calculated as width between start and end column is more than 65 percent of average width.

Method 3:"Skewness and Nearest Neighbour Approach For Historical Documents Classification"

In this paper [3] A.S Kavitha and P.Shivkumar proposes namely Skewness Based Approach (SA) and Nearest Neighbour Approach (NNA). The SA explores the fact that skewness between the components in the Indus document image with respect to x-axis is higher than skewness between the components in English and South Indian documents. The NNA identifies the presence or absence of modifiers which are common in South Indian document images and are not present in English document images to study the straightness and cursiveness of the components for classification.

Method 4:"A New method for line segmentation of Handwritten Hindi Text"

In this paper [4] a new method for Line Segmentation of Handwritten Hindi text is discussed. The method is based on header line detection, base line detection and contour(A line drawn on a map connecting points of equal height) following technique. No pre-processing like skew correction, thinning or noise removal has been done on the data.

III.DRAWBACKS OF EXISTING SYSTEM

1.Touching Characters are not recognized as they are considered as one word.

2.Words not touching the header line are not taken into consideration.

3.Characters such as (bha, ksha, tha, sha, ba, dha, shra) are excluded from existing system

IV.PROPOSED METHOD

We have divided our method in three steps: Image correction, segmentation and recognition

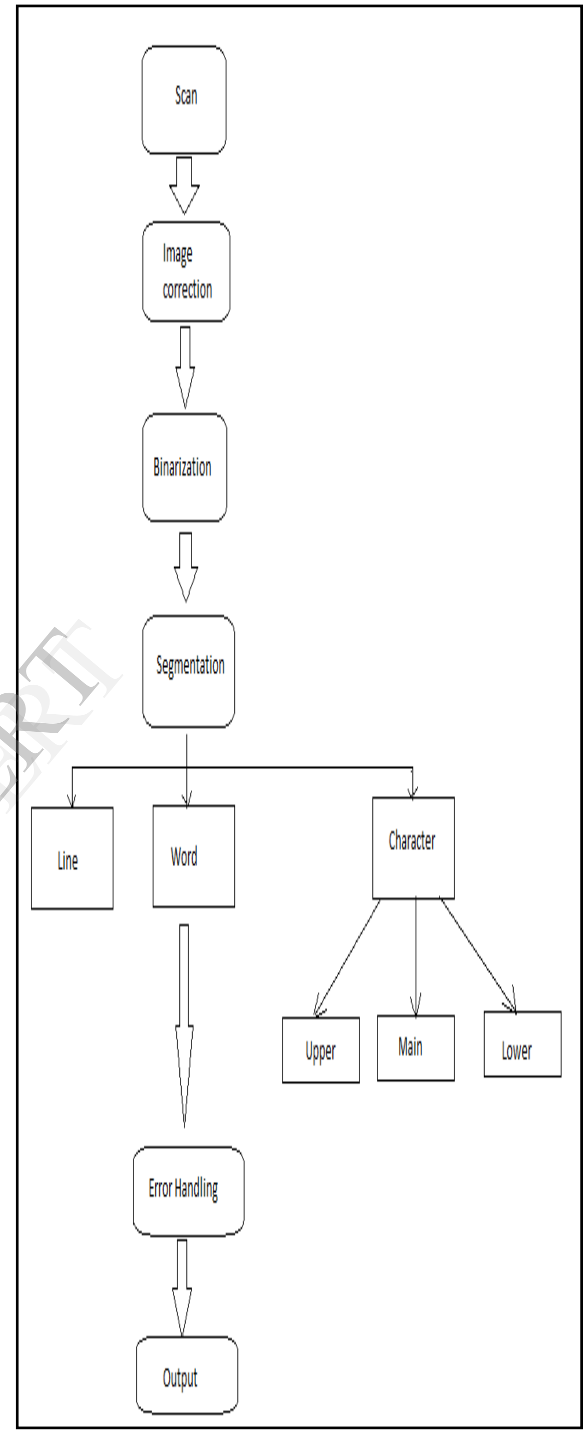


Fig 2:Flowchart

Step 1: Image correction:

Principle component analysis (PCA) is used to discover the hidden part and to obtain the clear information of the complex data in the image. It consists of Noise reduction, Gray scaling, Binarization.

- Noise reduction: Filtering is used to remove noise.

- Gray scale: In gray scaling image, RBG image is converted in to colorless image.

- Binarization: Adaptive thresholding method is used to binarize the image. Mean value is considered and the value below mean value is considered as black and above is considered as white. It is represented in the matrix form as black is considered as 1 and white as 0.

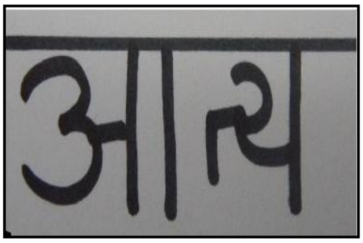


Fig 3:Original Image

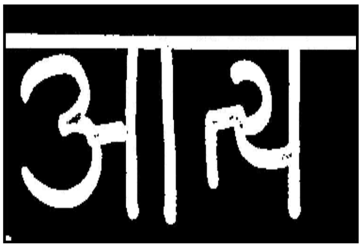


Fig 4:Binarized Image

Step 2: Segmentation:

There is segmentation on each step i.e Line-segmentation , Word-segmentation , Character-segmentation and fused word-segmentation for recognition . Character-segmentation includes further segmentation i.e Main character , upper modifier and lower modifier.

1.It finds the starting pixel of the header line traversing vertically.

2. After finding the starting point of the header line, it measures the breadth of the header line until it finds the white pixel traversing downward.

3. It removes the header line after finding the end point of the header line traversing downward until it finds the black pixel.

4.After finding the black pixel it removes the remaining portion. In this way height of the character is known.

5.Now we are left with only basic characters in an image. Those basic characters are segmented into equal single characters.

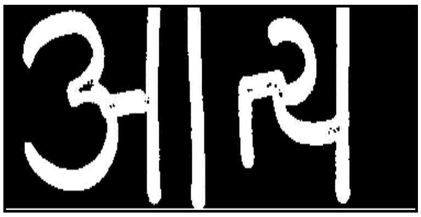


Fig 5:Removal of Header line

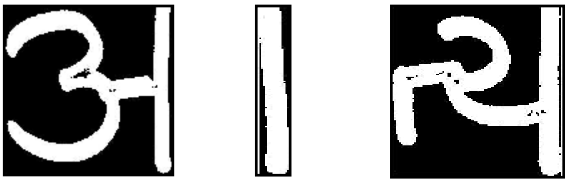


Fig 6:Extracted Letters

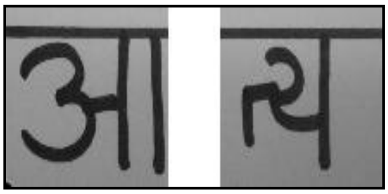


Fig 7:Segmented Images

Step 3: Extraction and Recognition:

We have stored 5 different samples per character with which the comparison will be done. In this method we used Eigen space algorithm using the concept of Gerschgorin's theorem which is usually used for face recognition in the image processing realm.

1.For each segmented characters finding the connected component and considering that for comparison with the stored image.

2. Eigen values are the values on the basis of which comparison is done of the images. Comparison is done with the help of distance measured between x and y coordinates.

3. It will match and compare the Eigen values of both the stored image and input image and based on which it will find the best and closest match and consider it as output.

4. This is how the basic character extraction is done.

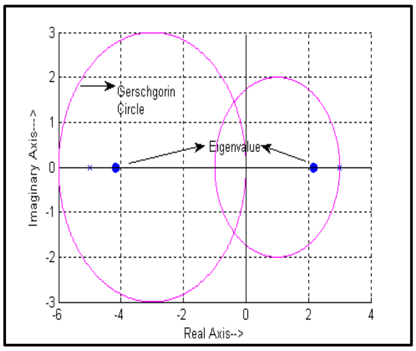
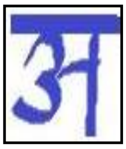


Fig 8: Gerschgorin circles and eigenvalues



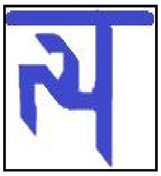


Fig 9:Standard Images for comparisons

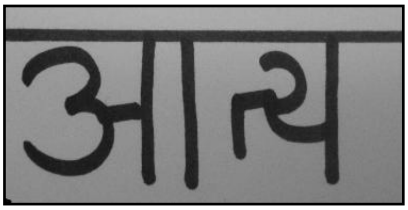


Fig 10:How I/O looks

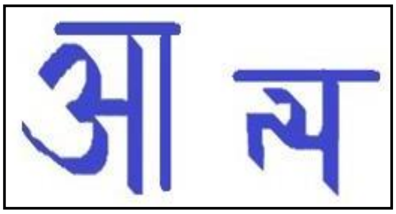


Fig 11:Final Results

Step 4: Modifiers recognition:

For each character there are stored images of the characters with all the modifiers consider the image after the binarization with removal of header line and segmented as described above without removing the modifiers.

1.For each basic character, modifiers are checked above and below the line.

2. Thinning of the image is done in order to obtain the pixel count.

3.Modifier on header line of the character are recognized after thinning of the image where horizontal line is traced above the header line and according to the count of the pixel the modifiers are decided.

4.Similarly for the lower modifiers, horizontal and vertical line is traced and based on the count of the pixel the modifiers are decided on the character.

5.In this way the result is displayed.

IV. CONSTRAINTS

1. Each character should be straight enough

2. Each character should be written in a separate equal space with all its modifiers

3. A word must have one end point i.e. the size of the upper limit and lower limit should be same.

4. Every character must touch the header line at the upper limit

5. There must be some distance between the starting pixels of header line and starting pixel of the first letter

V.APPLICATIONS

1.It can be used for automatic reading and processing of forms, old degraded documents, banks cheques.
e.g. Offices , Museums

2.It will be helpful for physically Handicapped people for the processing of the documents.

3.It can be used as a language converter.

VI.CONCLUSION

A new approach of Eigen space method which uses the concept of Gerschgorin's theorems in order to recognize and extract the characters. This algorithm is tested for different characters with different modifiers and it has shown the best results as compare to other character recognition techniques. Thus the proposed algorithm for extracting the characters from image using Eigen values is very efficient in recognition of characters.

TABLE I. Literature Survey

Sr. No	**Method**	**Feature**	**Data**	**Accuracy**
1	Miss Vandana[1]	Segmentation Of handwritten characters	124	56%
2	Veena Bansal[2]	Segmentation of fused characters	40	66%
3	A. S. Kavitha [3]	Component Angle, Bounding Box	600	83%
4	Naresh Kumar[4]	Header line,Base line	200	84%
5	Saiprakash Palakollu [5]	Segmentation Approach	200	73%
6	Aditi Goyal[6]	S.V.M Technique and HOG technique	278	80%
7	Prof.Kunal Shah, Prof. Badgujar[7]	DHCR for Ancient Documents	120	72%
8	Prof. Kunal Shah, Prof.Jaideep Singh[8]	Segmentation of Devnagari Documents	180	65%

VII.REFERENCES

1. Miss Vandana M.Ladwani ,Mrs Latish Malik "Novel Approach to Segmentation of Handwritten Devanagari Word", Third International Conference on Emerging Trends in Engineering and Technology.
2. Veena Bansal and R. M. K. Sinha "Segmentation of fused characters and touching words"
3. S. Kavitha, P. Shivakumara, G. Hemantha Kumar "Skewness and Nearest Neighbour based Approach for Historical Document Classification", 2013 International Conference on Communication Systems and Network Technologies
4. Naresh Garg ,Lakhwinder Kaur ,M.K Jindal "A New Method for Line Segmentation of Handwritten Hindi Text ", 2010 Seventh International Conference on Information Technology
5. Saiprakash Palakollu ,Renu Dhir ,Rajneesh Rani "Handwritten Hindi Text Segmentation Techniques for Lines and Characters " ,Proceedings of the World Congress on Engineering and Computer Science 2012 Vol I WCECS 2012, October 24-26, 2012, San Francisco, USA
6. Aditi Goyal ,Kartikay Khandelwal ,Piyush Keshri" Optical Character Recognition for Handwritten Hindi Characters" , ,Stanford University , CS229 Machine Learning
7. Kunal Shah, Prof. D. D. Badgujar, "DHCR for Ancient Documents: A review" – IEEE (ICT2013)
8. Kunal Shah, Jaideep Singh "A new approach for segmentation of Devanagari Documents"- IJCA.
9. R.M.K. Sinha and Veena Bansal, "Segmentation Of Touching And Fused Devanagari Characters", Technical Report TRCS-95-232, I.I.T. Kanpur, India
10. U. Bhattacharya and B. B. Chaudhuri, "Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals," IEEE Trans.Pattern Anal. Mach. Intell., vol. 31, no. 3, pp. 444– 457,Mar. 2009.
11. Vilas H.Gaidhane, Yogesh V. Hote, Vijandersingh "A new approach for estimation of eigan values of images", IJCA Volume 26– No.9, July 2011