ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 2, February 2018

Segmentation of Hindi Handwritten Words and Personality Prediction Using Word Spacing and Margin

Prof. Kalpana Malpe, Manjiri Wankhade

Dept. of Computer Science and Engineering, Gurunanak Institute of Engineering and Technology, Nagpur, India

Dept. of Computer Science and Engineering, Gurunanak Institute of Engineering and Technology, Nagpur, India

ABSTRACT: English Character Recognition approaches have been studied greatly in the last few years and its progress and success rate is quite high. But for regional languages these are still emerging and their success rate is moderate. There are millions of people who speak Hindi and use Devnagari script for writing. As digital documentation in Devnagari script is gaining popularity. Segmentation is one of the major stages of character recognition. This paper mainly deals with the new methods for line segmentation and character segmentation of characters of Handwritten Hindi text and shows word spacing and margin results. The text is segmented into lines, lines into words and then from characters. Algorithm is finding the header lines and base lines by estimating the average line height and based on it. This algorithm works efficiently on characters for different text images.

KEYWORDS: Devnagari script, Character Segmentation.

I. INTRODUCTION

From the past few decades handwritten character recognition is a frontier area of research on handwritten documents. No complete hand written text recognition system is available in Indian scenario and it is difficult due to large character set of Indian languages and the presence of vowel modifiers and compound characters in Indian script. Segmentation is one the most important process that decides the success of character recognition technique.

It is used to dissolve of an image of a sequence of characters into sub images of individual symbols by segmenting lines and words. Segmentation is to separate line, word and character from the image.

Background: In Recent years there is much work has been done in automated human behavior prediction system. It appears that there are many research studies exploiting various techniques blended with automated human behavior prediction system/ Graphology. If the graphology test is still done manually it takes a long time considering the aspects reviewed in graphology very much. Besides, accuracy of handwriting analysis depends on how skilled the analyst is. Development in image processing and pattern recognition lead to analyzing of handwriting can be done automatically. So it can be used by society at large. Handwriting is an image, so that recognition can be done through the stages of conversion of images into numerical vector, image processing for quality improvement, followed by feature extraction and pattern recognition.

India is a multi-lingual and multi-script country comprising of eighteen official languages. Because there is typically a letter for each of the phonemes in Indian languages, the alphabet set tends to be quite large. Hindi, the national language of India, is written in the Devnagari script. Devnagari is also used for writing Marathi, Sanskrit and Nepali. Moreover, Hindi is the third most popular language in the world. It is spoken by more than 500 million people in the world. Devnagari has 11 vowels and 33 consonants. They are called basic characters. Vowels can be written as independent letters, or by using a variety of diacritical marks which are written above, below, before or after the consonant they belong to. When vowels are written in this way they are known as modifiers and the characters so

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 2, February 2018

formed are called conjuncts. Sometimes two or more consonants can combine and take new shapes. These new shaped clusters are known as compound characters.
These types of basic characters, compound characters and modifiers are present not only in Devnagari but also in other scripts.

Characteristics of Hindi Script
Devnagari is most well-known script to write Hindi as well as Marathi, Sanskrit, Sindhi, and Nepali language with minor modifications. The alphabets of Devanagari script contain 33 consonants and 14 vowels. There is no any concept of lower or upper case in Hindi language. In Devnagari script, a text word may be divided into three zones. Upper zone represents the part above the headline, also middle zone covers the part of basic and compound characters under the headline, and lower zone may contain where some vowel and consonant modifiers can reside. A long number of characters (basic as well as compound) there exists a horizontal line at the upper part called as "shirorekha" or headline in Hindi.

II. RELATED WORK

1. Shubhra Saxena, P. C. Gupta, "**A Novel Approach of Handwritten Devanagari Character Recognition through Feed Forward Back Propagation Neural Network**".
Proposed Work:
Handwritten character recognition performs an important role in the modern world. It can solve more complicated problems and makes human's job easier. The present paper represents a novel approach in recognizing handwritten Devanagari character through feed forward back propagation neural network.

2. Snehal S.Patwardhan, R.R. Deshmukh, "**A Review on Offline Handwritten Recognition of Devnagari Script**".
Proposed Work:
Character Recognition is a process by which a computer identifies letters, numbers, or symbols and turns them into a digital form that a computer can use and it is an active field of research today. It consists of Pattern Recognition and Image Processing. Character Recognition is generally categorized into Optical Character Recognition (OCR) and Handwritten Character Recognition (HCR).
3. Neha Sahu, R. K. Rathy, Indu Kashyap, "**Survey and Analysis of Devnagari Character Recognition Techniques using Neural Networks**".
Proposed Work:
Research in Optical Character Recognition (OCR) is very important particularly with an eye on its applications in banks, post offices, defense organizations, library automation, etc. Devnagari Optical Character Recognition needs more consideration as it is national language and there is less development in this field due to complexity in the script.
4. Saiprakash Palakollu, Renu Dhir, Rajneesh Rani, "**Handwritten Hindi Text Segmentation Techniques for Lines and Characters**".
Proposed Work:
This paper mainly deals with the new approach for line segmentation and character segmentation of overlapping characters of Handwritten Hindi text. The text is segmented into lines, lines into words and then from lines words header lines are detected and converted as straight lines. Each word is divided into three parts upper modifier, consonant and lower, so that character segmentation becomes easy.
5. Naresh Kumar Garg, Lakhwinder Kaur, M. K. Jindal, "**Segmentation of Handwritten Hindi Text**".
Proposed Work:
The main purpose of this paper is to provide the new segmentation technique based on structure approach for Handwritten Hindi text.
6. Vikas J Dongre, Vijay H Mankar, **"Devnagari Document Segmentation Using Histogram Approach".**
Proposed Work:
A simple histogram based approach to segment Devnagari documents is proposed in this paper.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 2, February 2018

Various challenges in segmentation of Devnagari script are also discussed.

7. Partha Pratim Roy, Umapada Pal, Josep Lladós, **"Morphology Based Handwritten Line Segmentation Using Foreground and Background Information".**

Proposed Work:

In this paper we propose a method mostly based on morphological operation and run-length smearing algorithm (RLSA) to segment individual text lines from unconstrained handwritten document images.

8. Ratnashil N Khobragade, Dr. Nitin A. Koli, Mahendra S Makesar, "**A Survey on Recognition of Devnagari Script**".

Proposed Work:

In proposed system, we work on a set of preprocessing, segmentation, feature extraction, classification and matching techniques, which perform very important role in the recognition of characters. Feature extraction provides us techniques with the help of which we can identify characters uniquely and with high degree of accuracy. So many techniques have been proposed for pre-processing, feature extraction, learning/classification, and post-processing.

III. EXISTING SYSTEM APPROACH

The existing system is work on segmentation of lines and then characters. The segmented characters are then stores in result variable. First it Separate the lines and then it Separate the characters from the input image. This procedure is repeated till end of file. I.e. the existing system worked only on segmentation of word.

Disadvantages of Existing System:

Existing system work only on segmentation of words, so existing system not give word spacing and margin results.

IV. PROPOSED SYSTEM ARCHITECTURE

This system is works for Segmentation of words and shows word spacing and margin results. The segmented words are then stores in result. It segments the words from the input image. This procedure is repeated till end of file.

1. Read Input image
2. Convert image to gray scale image
3. Convert image to black and white i.e. threshold.
4. Invert image to change foreground pixels as background pixel and vice versa
5. Apply bounding box on segmented words.
6. Display the related result i.e. shows spacing and margin of words to predict the personality.

Classification:

Classification is based on following point

1. Classification using spacing between line structures.
2. Classification using spacing between word structures
3. Classification using page margin structure.

Advantages of Proposed System:

1. System takes minimum time to segmentation.
2. It also helps you to identify the hidden talents and aptitudes, thus, helping you to choose a right career path for yourself.
3. Handwriting analysis plays an important role in motivating employees to get better performance. It also helps a manager to understand the needs of his employees and work on them. It can revise teamwork by helping the manager to understand individual & team strengths and weaknesses.
4. Proposed approach work on segmentation of words, also proposed system give word spacing and margin results.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 2, February 2018

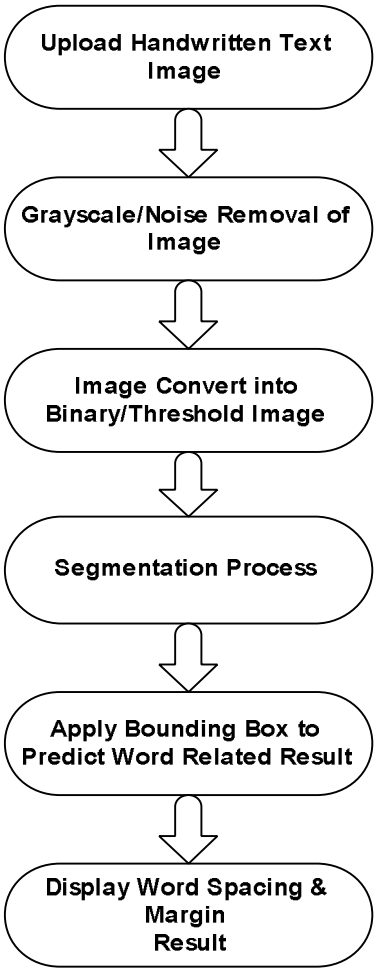

Fig1. Proposed System Architecture

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 2, February 2018

V. MATHEMATICAL MODEL

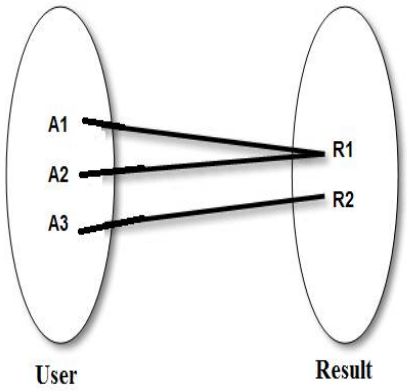

Fig2. Venn diagram

Where,
A1: Data provided by the user. Eg: Handwritten Image.
A2: Data provided by user. Eg: Handwritten Image.
R1: Resultant output provided by the input data.
A3: Wrong or incorrect data submitted.
R2: Error occurred.

SET THEORY:

S={s, e, X, Y,Φ}

Where,

s = Start of the program.

i. Log in.
ii. Upload Handwritten Image.
iii. Pre-processing.
iv. Image Thresholding.
v. Image Morphology.
vi. Word Segmentation
vii. Margin and Spacing Result.
viii. Predict personality.

e = End of the program.
Resultant output provided by the input handwritten image.

X = Input of the program.
Input should be Image file which is should be handwritten image.

Y = Output of the program.
Handwritten image will be uploading. Then the further processing will be done and finally appropriate result will provided.

$X, Y \in U$

Let U be the Set of System.
U= {Client, I, S, P}

Where, Client, I, S, P are the elements of the set.

Client= User

I= Handwritten Image

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 2, February 2018

S=Word Segmentation.
P=Personality prediction using word segmentation result.

SPACE COMPLEXITY:

The space complexity depends on presentation and visualization of discovered patterns. More the storage of data more is the space complexity.

TIME COMPLEXITY

Check No. of images available in the input image dataset= n
If (n>1) then retrieving of information can be time consuming.
So the time complexity of this algorithm is $O(n^n)$.

Φ = Failures and Success conditions.
Failures:

- Huge dataset can lead to more time consumption to get the information.
- Hardware failure.
- Software failure.

Success:

- Search the required information from available in Datasets.
- User gets result very fast according to their needs.

So the above mathematical model is NP-Complete.

VI. ALGORITHM

Algorithm for Segmentation of Lines and words

We make following assumptions about the data:
1. The minimum height of character consonant in a line is eight pixels. Average line height is 30 pixels.
2. The skew in a text is not more than the height of a character consonant.

The algorithm for line segmentation has following steps:

Step 1: Initially, rough estimate of the header lines in whole of the text are made by the formula

pcol (i) > 15 & pcol (i) > pcol (i: i+8) & pcol (i: i+8)>0
pcol (i) > floor (width (i/7))

Where,
pcol (i): No. of pixel in row i

Width (i): width of line i i.e. difference between last pixel and first pixel position of line i.

Step 2: After finding first header line, we skip 8 rows (equal to minimum height of consonant) to find the next header line.

Step 3: From $(i+8)^{th}$ row to $(i+22)^{th}$ row, we find the m^{th} row with minimum of pixels.

Step 4: We skip the rows up to m^{th} row and go to step 1 to find the next header line.
After finding the header lines, the most challenging task is to find the base line. For finding the base line following procedure is followed:

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 2, February 2018

Step 1: Two consecutive rough header lines are taken.
Step 2: The line is again divided into two equal halves (stripes).
Step 3: The rows with minimum of pixels are taken as base lines separately for each half.
Step 4: Then the lines are separated between header lines and base lines separately for each half.
Step 5: Then two separate lines are joined to get the actual text line.

Word Segmentation

After lines are segmented from text, words are segmented from lines by vertical projection profile. For each column of the line the number of black pixels is counted and the columns with zero black pixels are used as delimiters for word separation. To distinguish the character separation from the word separation, we have selected the delimiter as at least three continuous columns with zero black pixels for word separation.

Classification Algorithm:
Support Vector Machine (SVM) Algorithm:
Support vector machine is supervised learning algorithm which can be used for both regression and classification.
In this algorithm, system plots each data item as point in n-dimensional space with the value of a particular coordinate.
After that classification by finding the hyper-plane that differentiate the two classes very well.
Support Vectors are simply the co-ordinates of individual observation.
Support vectors are simply the best segregates the two classes.
SVM is one of the best known methods in pattern classification and image classification. It is designed to separate of a set of training images two different classes,
(x1, y1), (x2, y2), ..., (xn, yn) where xi in Rd, d-dimensional feature space, and yi in {-1,+1}, the class label, with i=1..n [1]. SVM builds the optimal separating hyper planes based on a kernel function (K). All images, of which feature vector lies on one side of the hyper plane, are belong to class -1 and the others are belong to class +1.

VII. RESULT ANALYSIS

Proposed work segmentation of lines and then segmentation of characters. First it Separate the lines and then it separate the characters from the input image. This procedure is repeated for all character images in image file. Then the segmented characters are stores in a result variable which are use to create the database for training and recognition of characters.

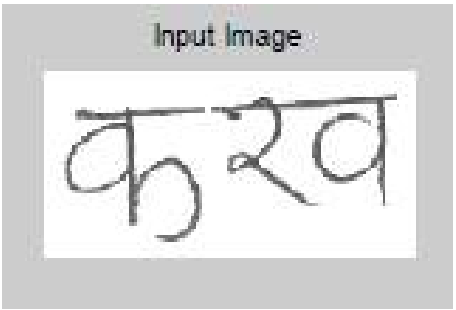

Fig3. Input Image

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 2, February 2018

Fig4. Grayscale Image

Fig5. Binary/Threshold Image

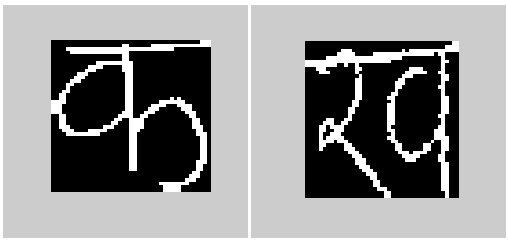

Fig6. Segmented Characters

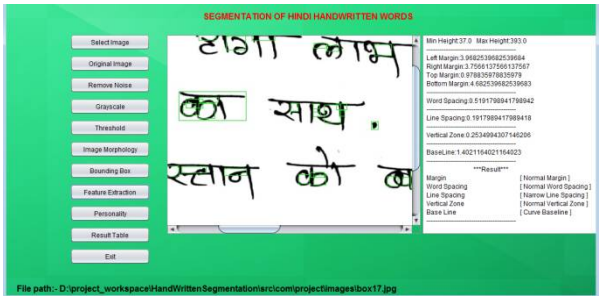

Fig7. Result

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 2, February 2018

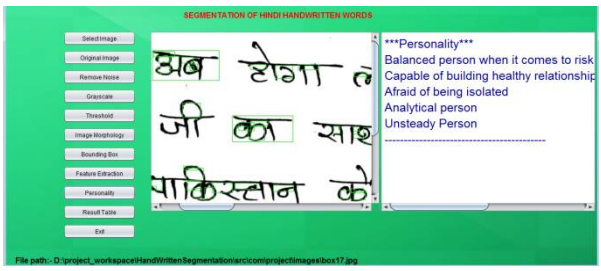

Fig8. Result

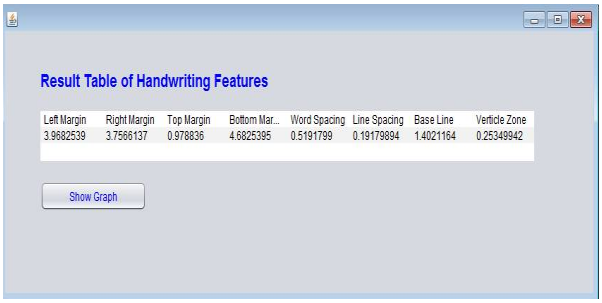

Fig9. Result Table

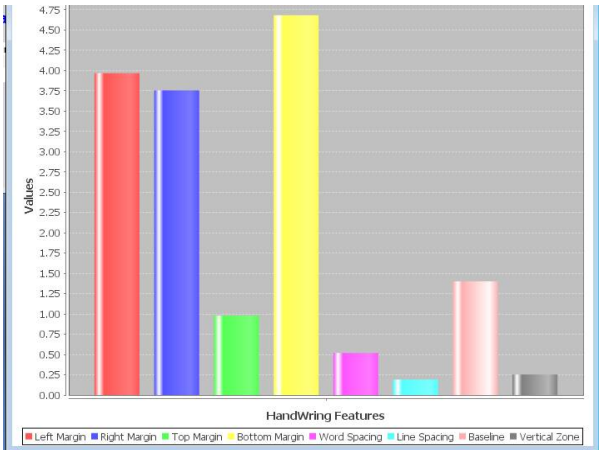

Fig10. Result Graph Using Features

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 2, February 2018

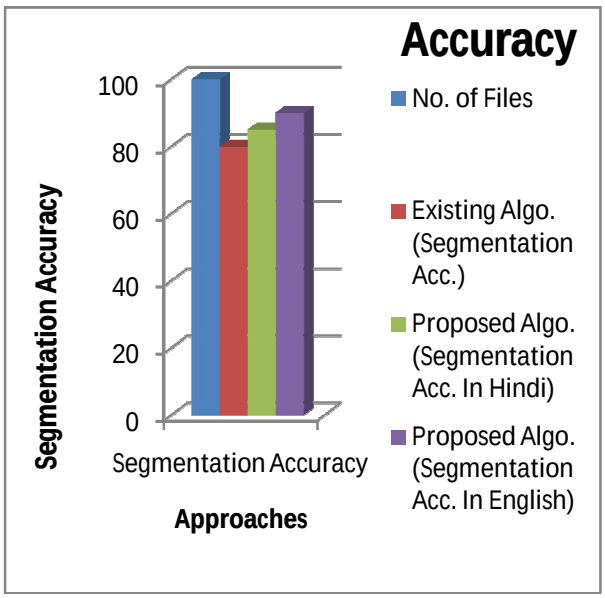

Fig11. Accuracy Graph

Fig. 11 shows the segmentation accuracy of existing approach and proposed approach using Hindi handwritten and also English handwritten. So our work gives the better accuracy as compared to existing work.

Table: Accuracy table

	No. of Files	Existing Algo. (Segmentation Acc.)	Proposed Algo. (Segmentation Acc. In Hindi)	Proposed Algo. (Segmentation Acc. In English)
Segmentation Accuracy	100	80%	85%	90%

VIII. CONCLUSION

In this work, we have presented a primary work for segmentation of words of Hindi Manuscript i.e. handwritten Hindi script. We obtained successful segmentation result in word segmentation. It is problematic to identify exact connecting points for segmentation of upper and lower modifier, separating anuswara (.), full stop (.) from noise is critical as both feature the same. After segmentation we have to show the word spacing and word margin related result to predict the personality.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 2, February 2018

REFERENCES

[1] Shubhra Saxena, P. C. Gupta, "**A Novel Approach of Handwritten Devanagari Character Recognition through Feed Forward Back Propagation Neural Network**".
[2] Snehal S.Patwardhan, R.R. Deshmukh, "**A Review on Offline Handwritten Recognition of Devnagari Script**".
[3] Neha Sahu, R. K. Rathy, Indu Kashyap, "**Survey and Analysis of Devnagari Character Recognition Techniques using Neural Networks**".
[4] Saiprakash Palakollu, Renu Dhir, Rajneesh Rani, "**Handwritten Hindi Text Segmentation Techniques for Lines and Characters**".
[5] Naresh Kumar Garg, Lakhwinder Kaur, M. K. Jindal, "**Segmentation of Handwritten Hindi Text**".
[6] Vikas J Dongre, Vijay H Mankar, **"Devnagari Document Segmentation Using Histogram Approach".**
[7] Partha Pratim Roy, Umapada Pal, Josep Lladós, **"Morphology Based Handwritten Line Segmentation Using Foreground and Background Information".**
[8] Ratnashil N Khobragade, Dr. Nitin A. Koli, Mahendra S Makesar, "**A Survey on Recognition of Devnagari Script**".