Handwritten Marathi Compound Character Segmentation with Morphological Operation

Mrs.Snehal S. Golait[1], Dr.L.G. Malik[2], Prof.A.Thomas [3]

[1]*Research Scholar ,Department of Computer Science and Engineering, G.H.Raisoni College of Engineering,Nagpur,*
[2]*Former Professor, Department of Computer Science and Engineering, G.H.Raisoni College of Engineering,Nagpur,*
[3]*Head of Department, Department of Computer Science and Engineering, G. H. Raisoni College of Engineering, Nagpur*

Abstract *–Segmentation phase plays vital role in any handwritten script Identification system. Aside from the large variation of individual's handwriting, many researchers found difficulty to separate characters from the captured text document Image. The key factor of selection of segmentation algorithm is used to improve efficiency of character segmentation as well as good feature extraction. There are so many features of Marathi Script like large character set, complex shape, modifier in that one of the feature is compound character. Segmentation of such type characters is very difficult due to their complex structure. This paper proposed novel technique for separation of handwritten Marathi compound characters. The first step in the segmentation process to segment the line of text document, word from the line and at the last character of the word. For separating characters from compound character our aim is to first find termination points and bifurcation points of the characters. We proposed a novel algorithm minutiae detection algorithm which is used to find termination and bifurcation points in the given image.*

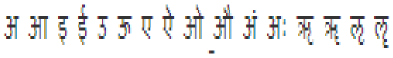
Keywords-Segmentation, Morphology, Minutiae, Compound character

I- INTRODUCTION

Segmentation partitioned an image into its constituent regions or objects. That is, it partitions an image into different regions that are meant to correlate strongly with objects or features of interest in the image. The segmentation process is not the easiest task, main goal of segmentation is to simplify change the representation of an image into meaningful and easier to recognize. Image segmentation is basically used to locate objects and boundaries in images. More precisely, image segmentation is the process of allocating a label to every pixel in an image such that pixels with the same label share certain characteristics.

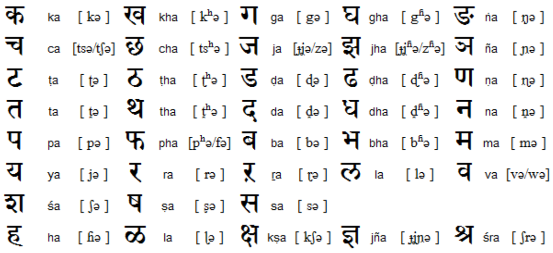
In optical character recognition, a proper segmentation of characters is required before individual characters are recognized. An OCR has a wide variety of Commercial and physical applications. It can be used for postal automation, institutional repository, in the health care system, in CAPTCHA, automatic reading, processing of the forms, old degraded documents, bank cheques etc. It can prove as an aid for visually handicapped persons. There are so many scripts and languages in India, but very less work is done in recognition of handwritten Indian scripts.

Handwritten character recognition for Indian scripts is quite a challenging task for the researchers. This is due to the various characteristics of these scripts like their large character set, complex shape, presence of modifiers and similarity between characters. Marathi is the language spoken by the native people of Maharashtra. Marathi belongs to the group of Indo-Aryan languages which are a part of the largest group of Indo-European languages, all of which can be traced back to a common root. It is the 4th most spoken language in India and 15th most spoken language in the world. [1] Marathi script consists of 16 vowels and 36 consonants, making 52 alphabets. Marathi is written from left to right. It has no upper and lower case characters. Every character has a horizontal bar at the top called as the header line. The header line joints the characters in a word. The vowels, consonants and modifiers in Marathi language shown in figure 1, 2 and 3.



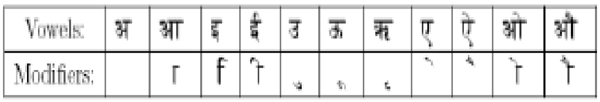
अ आ इ ई उ ऊ ए ऐ ओ औ अं अः ऋ ॠ ऌ ॡ

Figure 1: Vowels In Marathi Script



क	ka	[kə]	ख	kha	[kʰə]	ग	ga	[gə]	घ	gha	[gʱə]	ङ	ṅa	[ŋə]
च	ca	[tsə/tʃə]	छ	cha	[tsʰə]	ज	ja	[ɟjə/zə]	झ	jha	[ɟjʱə/zʱə]	ञ	ña	[ɲə]
ट	ṭa	[ʈə]	ठ	ṭha	[ʈʰə]	ड	ḍa	[ɖə]	ढ	ḍha	[ɖʱə]	ण	ṇa	[ɳə]
त	ta	[t̪ə]	थ	tha	[t̪ʰə]	द	da	[d̪ə]	ध	dha	[d̪ʱə]	न	na	[n̪ə]
प	pa	[pə]	फ	pha	[pʰə/fə]	ब	ba	[bə]	भ	bha	[bʱə]	म	ma	[mə]
य	ya	[jə]	र	ra	[rə]	ऱ	ṛa	[ɽə]	ल	la	[lə]	व	va	[və/wə]
श	śa	[ʃə]	ष	ṣa	[ʂə]	स	sa	[sə]						
ह	ha	[ɦə]	ळ	la	[ɭə]	क्ष	kṣa	[kʃə]	ज्ञ	jña	[ɟjɲə]	श्र	śra	[ʃrə]

Figure 2: Consonants In Marathi Script



Vowels:	अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ
Modifiers:		ा	ि	ी	ु	ू	ृ	े	ै	ो	ौ

Figure 3: Modifiers In Marathi Script

Marathi also has a complex system of compound characters in which two or more consonants are joined forming a new special symbol. Compound characters in Marathi script occur more frequently in the script as compared to other languages derived from Devanagari. The occurrence of compound characters in Marathi is found to be about 15 to 20% whereas in other scripts of Devanagari and Bangla script, it is just 10 to 15% [1]. Compound can be formed by joining one or more consonants together. Different joining patterns for Marathi character as shown in Figure 4.

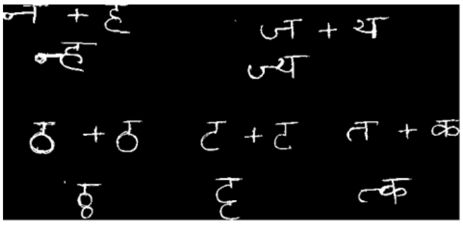


Figure 4: Joining Patterns of Handwritten Marathi Compound Characters

The various patterns for forming Marathi compound character is shown in figure 4. Compound character is formed by first truncating the side bar of a character and joined it to the left hand side character. Such patterns for joining is more typical in Marathi script. Another way of forming compound character is just by tie the character one aboveanother.

Segmentation is a technique which subdivides handwritten text into individual characters. Since recognition heavily relies on isolated characters, segmentation is a difficult phase for character recognition because better is the segmentation, lesser is the ambiguity encountered in recognition of candidate characters of word pieces.[7]

This paper gives a novel approach for segmenting compound character for handwritten Marathi Script.

II- RELATED WORK

Devnagari is the most widely used script in India. Sanskrit, Nepali, Hindi and Marathi are the devnagri script used by more than 400 million people. Unconstrained Devnagari writing is more complex than English language due to the possible variations in the shape, number and direction of the constituent strokes. Devnagari script has 50 characters which can be written as individual symbols in a word. Devnagari Character recognition is complicated process due to presence of multiple conjuncts, loops, lower and upper modifiers and the number of disconnected and multistroke characters, in a word where all characters are connected through Shirorekha. OCR is further complicated by compound characters that make character separation and identification is very difficult.

OCR work on printed Devnagari Script started in early 1970's. Sinha and Mahabala published presented a syntactic pattern analysis system with an embedded picture language for the recognition of handwritten and machine printed Devnagari characters [1]. Veena Bansal described number of knowledge sources to recognize the Devanagari character in her doctoral Thesis. She proposed work with the use of a hybrid approach for classification of characters and symbols. She obtained an overall performance of 93% accuracy at the character level. The first OCR system was developed for machine printed Devanagari character by Pal and Chaudhuri as well as by Patil. They worked on detection of headline, also worked on an approach for dividing text document such as word into three zones like lower zone ,upper zone and middle zone.They are getting the recognition accuracy up to 96% .

First research report on handwritten Devnagari characters was published in 1977. At present researchers have started to work on handwritten Devnagari characters and few research reports are published recently. Hanmandlu and Murthy proposed a Fuzzy model based recognition of handwritten Hindi numerals and characters and they obtained 92.67% accuracy for Handwritten Devnagari numerals and 90.65% accuracy for

Handwritten Devnagari characters. Bajaj et al employed three different kinds of features, namely, the density features, moment features and descriptive component features for classification of Devnagari Numerals. They proposed multi-classifier connectionist architecture for increasing the recognition reliability and they obtained 89.6% accuracy for handwritten Devnagari numerals.Segmentation approach is to recognize handwritten Devanagari word proposed by Shaw. With the knowledge of the Shirorekha , a word input image is separated to pseudo characters.Dr. Latesh Malik proposed techniques for word isolation, segmentation and recognition.She obtained 95% accuracy[4]. Shubair Abdulla proposed novel segmentation algorithm to recognize handwritten Arabic characters with Rotational Invariant Segment features. Segmentation algorithm achieved 95.66% accuracy for segmentation of word for Arabic handwritten Script [12]. Sushama Shelke worked on handwritten Marathi Compound Character Recognition using Structural feature extraction technique wavelet transform obtained 94.22 % accuracy.Mr. Dipak V. Koshti, Mrs. Sharvari Govilkar proposed method for segmentation of touching characters in Handwritten Marathi Text. They used joint point algorithm for segmenting touching characters. Sirisha Badhika proposed multilevel Segmentation algorithm using cognitive approach. Sharad Gupta and Abdul Momin proposed a novel algorithm to segment the fused and merged characters. As per related research no one using the minutiae technique to segmenting character. This paper discussed how the concept of minutiae is used for segmenting Marathi character from the handwritten Marathi compound character.

III- PROPOSED APPROACH

The proposed system consists of following stages of OCR which includes preprocessing steps and recognition step. The preprocessing steps Shown in Figure 5.

Image Enhancement

This phase includes the scanning of text document, the document which is scanned as color or grey image is converted into binary image. At the time of scanning, if document is scanned as black and white then no conversion is needed. After converting normal image into binary image, the noise reduction has to be done, for removing the small dots that were added at the time of scanning.

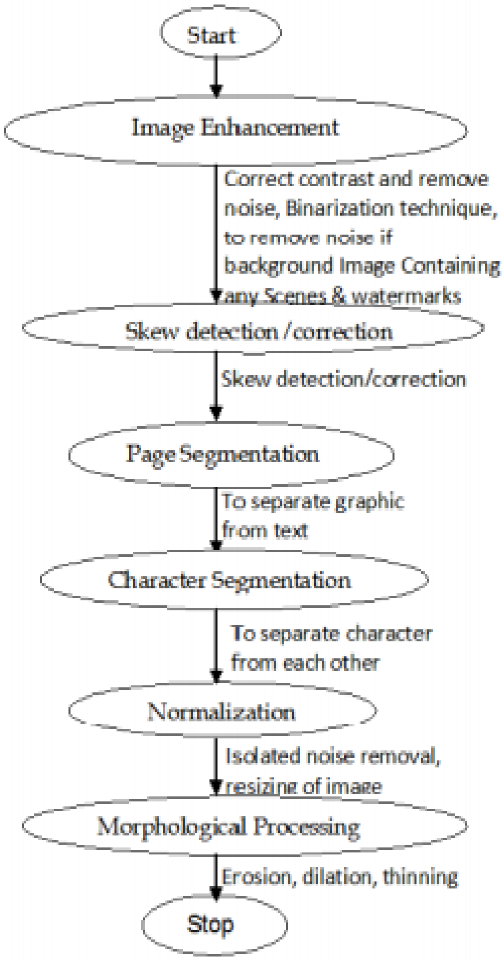

Start
Image Enhancement
Correct contrast and remove noise, Binarization technique, to remove noise if background Image Containing any Scenes & watermarks
Skew detection / correction
Skew detection/correction
Page Segmentation
To separate graphic from text
Character Segmentation
To separate character from each other
Normalization
Isolated noise removal, resizing of image
Morphological Processing
Erosion, dilation, thinning
Stop

Figure 5: Flowchart for proposed approach

Skew Correction

At the time of scanning or writing something on paper, some amount of skew is introduced with respect to the horizontal line. Document skew is nothing but the angle introduced while scanning the text document. This skew angle is, the angle made by Shirorekha with the horizontal line. There are several methods to calculate the angle and correct the skew. The skew is corrected by rotating the skew angle with a horizontal line.

Line Segmentation

The first step of the segmentation process is segmenting the text region into lines, also called as line segmentation. Before line segmentation first we have to locate the position of the text in a scanned document. For this check all the pixels on each scan line. If the pixel intensity value of each scan line is one, then store that scan line number. The process continues till we get no black pixels. Note the dimension of the text line will be found from stored scan line positions.

Word Segmentation

Word segmentation is an easier task as compared to line segmentation and character segmentation. The space between two words is generally more than two or three pixels. Word segmentation is done by the projection based method. For word segmentation uses the following algorithm.

Proposed algorithm for Identifying Compound characters

Method1:

1. Find the width of all Characters.
2.Calculate the average width of a character.

```
If  Cw > CAvgW  then
Character is Compound Character
```

Proposed Segmentation approach

For Segmenting the compound character our aim is to find the termination points and bifurcation points.

1. Apply minutiae detection algorithm to find termination and bifurcation points.
2. If(pixel having only one neighbor)
   The point is termination point.
3. If(Pixel having three neighbors)
   The point is bifurcation points.

The pseudo code for finding the termination and bifurcation point is as follows.

Pseudo Code for finding termination and bifurcation Points:

```
[pbif,pterm,img_out] applyMinutae(logical(current_char_thin));
num_bif = length(find(pbif));
num_term = length(find(pterm));
% Find the maximum number of discontinuous characters
max_discon = length(find(current_char_thin(:))) / length(current_char_thin(:));
% Find the factors which we are using for joint character detection
 factor1 = num_term/num_bif;
 factor2 = max_discon;
% Show the character and print the factor
Imshow (current_char_thin);
title(sprintf('T:%d,B:%d,Factor:%0.08f, disconnectivity:%0.04f',num_term,num_bif,num_term/num_bif,max_discon));
```

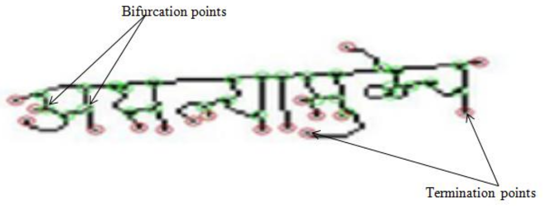
Character Segmentation

With the help of factor1 , factor2 and threshold value we have to segment the character from compound character. The pseudo code for character segmentation is as follows

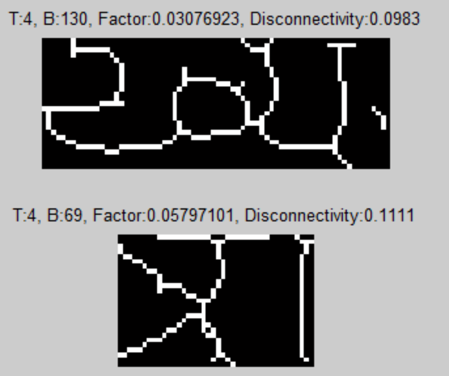
Pseudo Code for Character Segmentation

```
% Apply thresholding to find the joint characters
if(factor1 < 0.03 && factor1 > 0 && factor2 > 0 && factor2 > 0.08)
% Split the characters
size_index = size(current_char_thin,2);
left_char = current_char_thin(:,1:round(size_index/2));
right_char = current_char_thin(:,round(size_index/2):end);
```

IV- EXPERIMENTAL RESULTS

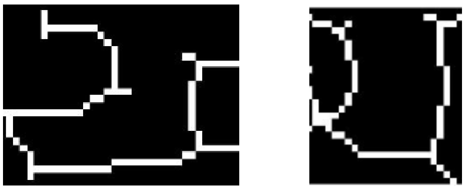


Output of Segmentation Algorithm



T:4, B:130, Factor:0.03076923, Disconnectivity:0.0983

T:4, B:69, Factor:0.05797101, Disconnectivity:0.1111



Output of Character segmentation

V- CONCLUSION

In this paper, we proposed algorithm for segmentation of handwritten Marathi Compound Character using morphological operation. From the experimental results it is observed that minutiae algorithm is successful in finding termination and bifurcation points with 98% accuracy and for character segmentationgot 96% accuracy.

VI- ACKNOWLEDGMENT

This paper is made possible through the help and valuable support of my guide Dr.Latesh Malik. Finally, I sincerely thanks to my parents and God.

REFERENCES

[1]Sushama Shelke, Shaila Apte, "A Multistage Handwritten Marathi Compound Character Recognition Scheme using Neural Networks and Wavelet Features " International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 4, No. 1, March 2011.

[2] Miss Vandana M. Ladwani, Dr.latesh Malik, "Novel Approach to Segmentation of Handwritten Devnagari Word", 978-0-7695-4246-1/10 $26.00 © 2010 IEEE DOI 10.1109/ICETET.2010.143.

[3]U.K.S. Jayarathna, G.E.M.D.C. Bandara," A Junction Based Segmentation Algorithm for Offline Handwritten Connected Character Segmentation", International Conference on Computational Intelligence for Modelling Control and Automation,and International Conference on Intelligent Agents,Web Technologies and Internet Commerce (CIMCA-IAWTIC'06).

[4] Ambadas B. Shinde, yogesh H. Dandawate, " Shirorekha Extraction in Character Segmentation for Printed Devanagri Text In Document Image Processing", 2014 Annual IEEE India Conference (INDICON).

[5]Roli Bansal, Priti Sehgal & Punam Bedi," Effective Morphological Extraction of True Fingerprint Minutiae based on the Hit or Miss Transform", International Journal of Biometrics and Bioinformatics(IJBB), Volume (4) : Issue (2).

[6] Dhaval Salvi, Jun Zhou, Jarrell Waggoner, and Song Wang," Handwritten Text Segmentation using Average Longest Path Algorithm", 978-1-4673-50542-95/132/$31.00 ©20132 IEEE.

[7] R.G. Casey et.al. "A Survey of Methods and Strategies in Character Segmentation", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 18,pp 690-706, 1996.

[8] Veena Bansal and R.M.K. Sinha. "Segmentation of touching and Fused Devnagari characters, ". Pattern recognition, vol. 35: 875-893, 2002.

[9] Dr. Latesh Malik," A Graph Based Approach for Handwritten Devnagari Word Recognition", 2012 Fifth International Conference on Emerging Trends in Engineering and Technology.

[10] Ms. Aarti Desai, Dr. Latesh Malik," A Modified Approach to Thinning of Devanagri Characters", 978-1-4244-8679-3/11/$26.00 ©2011 IEEE.

[11] K.B.M.R. Batuwita, G.E.M.D.C. Bandara," Meaningful Segmentation of Offline Individual Handwritten Numeric Characters", 2006 IEEE International Conference on Fuzzy Systems ,Vancouver, BC, Canada July 16-21, 2006.

[12] Shubair Abdulla, Amer Al-Nassiri , Rosalina Abdul Salam , " Off-line Arabic Handwritten Word segmentation Using Rotational Invariant Segments Features" , International Arab Journal of Information Technology , Vol. 5, No. 2,April 2008.