A Survey for Segmentation Techniques for Handwritten Devnagari Text Document

[1]Mrs. Snehal S. Golait

Assistant Professor
Priyadarshini College of engineering,Nagpur
snehal.golait@gmail.com

Abstract *— Segmentation is one of the basic functions of handwritten Script Identification. Aside from the large variation of different handwriting its very difficult to segment the character from word. This paper give the short review on segmentation methods for handwritten character recognition. The aim is to provide an appreciation for the range of techniques that have been developed rather than to simply list sources. Various types of segmentation methods proposed for handwritten script identification include histogram based approach, average longest path approach, morphological approach, Junction based approach.*

Keywords — *Histogram, Morphology, Statistical, End points, loops, junctions.*

I. INTRODUCTION

Segmentation is the process of partitioning a digital image into multiple segments. The goal of segmentation is to simplify change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics.

In optical character recognition (OCR), a perfect segmentation of characters is required before individual characters are recognized. An OCR has variety of commercial and physical applications. It can be used for automatic reading and processing of the forms, old degraded documents, bank cheques. It can prove as an aid for visually handicapped persons. There are many scripts and languages in India but not much research is done for recognition of handwritten Indian characters. Segmentation is a technique which partitions handwritten text 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. Segmentation is broadly classified into four units.

1. Classical approach consists of methods that partition the input image into sub images, which are then classified
2. The operation of attempting to decompose the image into classifiable units is called dissection
3. The third strategy is a hybrid of the first two, employing dissection together with recombination rules to define potential segments.
4. Holistic approaches that avoid segmentation by recognizing entire character strings as units are described.

This paper gives short review on segmentation technique used in handwritten devnagari character recognition. 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, presence of compound characters 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 larger of group of Indo-European languages, all of which can be traced back to a common root. Among the Indo-Aryan languages, Marathi is the southern-most language. All of the Indo-Aryan languages originated from Sanskrit. Three Prakrit languages, simpler in structure, emerged from Sanskrit. These were Saurseni, Magadhi and Maharashtri. Marathi is said to be a descendent of Maharashtri which was the Prakrit spoken by people residing in the region of Maharashtra. The script currently used in Marathi is called 'Balbodh' which is a modified version of Devnagari script. Earlier, another script called 'Modi'

was in use till the time of the Peshwas(18th century). This script was introduced by Hemadpanta, a minister in the court of the Yadava kings of Devgiri (13th century). This script looked more like today's draviDian scripts and offered the advantage of greater writing speed because the letters could be joined together. Today only the Devanagari script is used which is easier to read. Marathi script derived from Devanagari, is an official language of Maharashtra. It is the 4th most spoken language in India and 15th most spoken language in the world. 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 line at the top called as the header line. The header line joins the characters in a word. The vowels, consonants and modifiers in Marathi language shown in fig 1, 2 and 3.

अ आ इ ई उ ऊ ए ऐ ओ औ अं अः ऋ ॠ ऌ ॡ

Fig 1- Vowel in Marathi Script

क	ka	[kə]	ख	kha	[k^hə]	ग	ga	[gə]	घ	gha	[$g^ɦ$ə]	ङ	ṅa	[ŋə]
च	ca	[tsə/tʃə]	छ	cha	[ts^hə]	ज	ja	[ʤə/zə]	झ	jha	[$ʤ^ɦ$ə/$z^ɦ$ə]	ञ	ña	[ɲə]
ट	ta	[ʈə]	ठ	ṭha	[$ʈ^h$ə]	ड	da	[ɖə]	ढ	ḍha	[$ɖ^ɦ$ə]	ण	ṇa	[ɳə]
त	ta	[t̪ə]	थ	tha	[$t̪^h$ə]	द	da	[d̪ə]	ध	dha	[$d̪^ɦ$ə]	न	na	[nə]
प	pa	[pə]	फ	pha	[p^hə/fə]	ब	ba	[bə]	भ	bha	[$b^ɦ$ə]	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	ksa	[kʃə]	ज्ञ	jña	[ʤ̃ɲə]	श्र	śra	[ʃrə]

Fig 2-Consonants in Marathi Script

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

Fig 3- Modifiers in Marathi Script

Marathi also has a complex system of compound characters in which two or more consonants are combined 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 11 to 12% whereas in other scripts of Devanagari and Bangla script, it is just 5 to 7% [1]. The compound characters exhibit following features: the consonants in the compound character are not joined in an arbitrary manner but the combination of some specific characters is done in order to give a meaningful combination. The compound character can have two or more characters joined together in various ways as shown in Fig 4. One way of forming compound character is by removing the vertical line of a character and then joining it to the other on its left hand side. This type of joining is more common. Another way of connection of characters in the compound character is by just joining the characters side by side or one above the other.

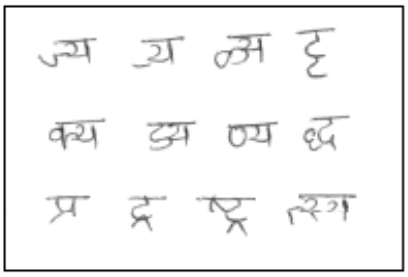


Fig 4- Samples of Marathi compound character.

Segmentation Methods
Histogram Based Segmentation

The first step of segmentation process is segmenting the text region into lines, also called as line segmentation. After line segmentation perform the word segmentation. Word segmentation is easier than line segmentation and character segmentation. Space between two words is generally more than three pixels. Words are segmented by the projection based method. For recognizing the character we need to separate the character from word. One of the method for character segmentation is histogram approach shown in fig 5. In Line based script segmentation method all black pixels on every row are computed horizontally. In this method separating individual line in a script document image based on the peak of the horizontal Histogram[2]. For the word segmentation construct the Vertical Histogram for the image in that count the white pixel in each column. Using the Histogram, find the columns containing no white pixel. Replace all such columns by 1. Invert the image to make empty rows as 0 and text words will have original pixels. Mark the Bounding Box for word. Copy the pixels in the Bounding Box and save in separate file.

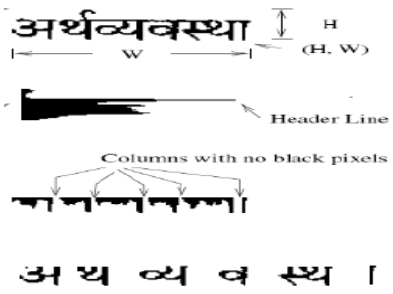


Fig 5- Histogram Based Segmentation

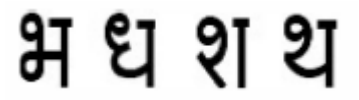
Average Longest Path Segmentation

The proposed method is used for handwritten text segmentation, which does not make any limiting assumptions on the character size and the number of characters in a word. Specifically, the proposed method finds the text segmentation with the maximum average likeliness for the resulting characters. For this purpose, use a graph model that describes the possible locations for segmenting neighboring characters and develop an average longest path algorithm to identify the globally optimal segmentation. The algorithm is proposed as

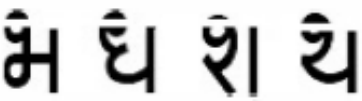
1. Construct the candidate segmentation boundaries
2. Construct a directed graph where each vertex represents a candidate segmentation boundary, including the left and right image border, where each edge represents the text segment between two candidate segmentation boundaries.
3. Weigh the graph edges by the character likeliness derived from a character recognition algorithm.
4. Find the average longest path between the leftmost vertex (left image border) and rightmost vertex (right image border) in the graph.
5. Take the candidate segmentation boundaries, corresponding to the vertices along the identified average longest path, as the final segmentation boundaries for text segmentation.

Morphological based Segmentation

In proposed morphological operation,[4]scanned document image is once again scanned horizontally i.e. the horizontal projection of document image is taken to plot horizontal projection profile. The existence of Shirorekha is found by the large projection in the graph, and then appropriate operations are done to extract/remove Shirorekha. It is found that sometimes horizontal projection profile method is not suitable. To overcome this drawback proposed new method, without using projection profile information. The proposed morphological operations is used to locate and extract the Shirorekha. In this method first dilation of image is takes place, after inverting the image followed by addition of inverted image to the original image. In Character segmentation a word is separated into characters. To do the character segmentation for Devanagari text, we need to remove the Shirorekha. By removing Shirorekha able to separate the characters and plot bounding boxes for the separated Devanagari characters. For recognizing the characters, need to locate the Shirorekha first. At the initial stage of Devanagari character recognition many researchers suggested to use horizontal projection profile method. In this horizontal projection profile method, the projection of the binary image is taken horizontally, as shown in fig 6a) and 6b). In this projection we are getting a large projection (projection peak along x axis) along the Shirorekha place. Form this profile able to locate the Shirorekha, but the difficulty with this method is that, if multiple letters look like them as shown in fig 6a) will get second largest projection in the horizontal projection profile which will lead us to create confusion in locating Shirorekha line. The existence of Shirorekha is found by the horizontally large projection in the graph, then appropriate operations are done to remove Shirorekha. One of the drawback of this method is that, some part of Devanagari characters is removed which may leads towards improper segmentation



(a)



(b)

Fig 6 - Output of Horizontal Projection

The proposed method for locating and removing Shirorekha line is:

1. Scan the document.
2. Morphological Dilate operation on image
3. Invert the image
4. Add inverted image with original image
5. Convert this image to binary image
6. Character segmentation

Dilated Image: The words are dilated to allow formation of bounding boxes only around words. Morphology is a broad set of image processing operations that process images based on shapes. Morphological operations apply a structuring element to an input image, creating an output image of the same size. In a morphological operation, the value of each pixel in the output image is based on a comparison of the corresponding pixel in the input image with its neighbors. By choosing the size and shape of the neighborhood, can construct a morphological operation that is sensitive to specific shapes in the input image.

The most basic morphological operations are dilation and erosion. Dilation adds pixels to the boundaries of objects in an image, while erosion removes pixels on object boundaries. The number of pixels added or

removed from the objects in an image depends on the size and shape of the structuring element used to process the image. To dilate the image, pass the Image BW and the structuring element SE to the Imdilate function. Dilation adds a rank of 1's to all sides of the foreground object.

Junction based segmentation

When considering the connected character segmentation, it is almost impossible to segment connected character skeleton in to a set of meaningful segments[1]. This is due to the possibility of having different connections between two-digit character strings. Therefore the segmentation stage of the proposed work consists of two phases namely, the initial segmentation phase and the total segmentation phase. In the initial segmentation, the input character image undergoes Junction Point identification and Junction Based Segmentation.

A Junction Point is a pixel point in the Correlation area, having three or more neighboring pixels. It is assumed here that the skeleton is in one pixel thickness. After the identification of the all junction points (Fig 7,) in the correlation area, each is used to segment the connected character skeleton in to initial segments.

Fig 7- Connected Devnagari character

As Shown in fig 7 green circles are junction points and red circles are end points. For segmentation of connected characters used template matching algorithm.

Algorithm

1.Find the connected character

2. Find the terminating points.

3. Go to each set of terminating points

4.Perform template matching with each entry obtained after termination contour.

CONCLUSION

In this paper there are various methods of segmentation was discussed. In Histogram approach from peak of histogram will separate individual line in a script. In average longest path segmentation will find the maximum average for finding the boundaries for text segmentation. Morphological based segmentation are using some morphological operations such as dilation and Erosion etc. and Junction based approach is used for finding the junction of character .

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 Intelligen 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] Neha Sahu,"DEVANAGIRI DOCUMENT SEGMENTATION USING HISTOGRAM BASED APPROACH", International Journal of Electronics, Electrical and Computational System IJEECS ISSN 2348 117X Volume 3, Issue 3 May 2014.

[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.