SEGMENTATION OF OVERLAPPING GURUMUKHI HANDWRITTEN CHARACTERS

Arwinder Kaur[1], Ashok Kumar Bathla[2]

[1]Student of Computer Engineering, [2] Assistant Professor of Computer Engineering,
Yadavindra College of Engineering, Talwandi Sabo, Punjab (India)

ABSTRACT

Optical Character Recognition is the process of automatic recognition of characters from optical scanned images or digitized pages of text. It is changing them into the information that can be easily interpreted by machine. Segmentation is the most crucial step for recognition of characters. Segmentation is the process used to segment the text into beneficial segments for recognition. Recognition of printed fonts is easy but recognition of handwritten scripts is still difficult. It is because of variations in writing style of writers and other abnormalities like presence of touching, broken and overlapping characters. Now days, Handwritten character recognition is gaining importance. Punjabi is one of the most widely spoken languages. This paper shows a strategy for the segmentation of overlapping characters in Gurumukhi handwritten script on Punjabi language. It is based on the distance metric and distance transform of the neighboring pixels for overlapping characters.

Index Term: Gurumukhi, overlapping, Punjabi, recognition, segmentation

I. INTRODUCTION

Character segmentation is very important step in recognition process, accurate segmentation leads to minimum errors in the proceeding stages. Segmentation in case of hand written text is not easy because of variation in writing styles of different writers and complex structure of Gurumukhi script characters. Uneven header line presents a big problem which is further elevated by presence of touching, overlapping and broken characters. Much research has been done on this issue in the languages like English but much proficiency is needed in some Indian scripts like Gurumukhi. [1] shows basic segmentation strategies- classical, holistic, recognition based and hybrid approach. Holistic approach avoids much segmentation by recognizing entire character string as unit. This paper mainly provides a review of basic principles and methods used for segmentation. [2] gives an iterative algorithm for segmentation of Gurumukhi text. It uses horizontal profile and vertical profile projection methods. They considered three categories for segmentation of words as- Basic segmentation, under segmentation and over segmentation and results in 96.22% of efficiency. [3] shows an algorithm for segmenting isolated and touching characters using Water Reservoir principle for Gurumukhi script. The projection profile methods have also been used for identifying and segmenting the touching characters. The accuracy 93.5% has been resulted for different documents using this technique. [4] works on Gurumukhi handwritten documents for segmenting skewed, broken, touching and overlapping characters. Different handwritten samples are tested for each type of challenging characters and overall accuracy of 73.6% has been obtained. [5] proposes an efficient algorithm for the segmentation of horizontally overlapping lines for printed text in Gurumukhi Script. Horizontal Projection Profile has been used for detecting and removing header line. Further it is used to find the average height of lines for segmenting overlapping characters; and upper and lower modifiers. [6] proposes a method based on Vertical Projection Profile to segment the characters from word and extract the base characters by identifying the empty spaces between them. This algorithm is implemented on bank cheques and concludes 97% of efficiency for isolated characters only. [7] has given algorithm for segmenting lines, words and characters in Hindi language. Projection profile approaches have been used for segmenting characters and leads to 79.12% accuracy for simple characters; this algorithm doesn't give good results for half and touching characters**.** [8] describes the watershed transformation as a particular method based upon the regions approach to the segmentation of an image. The complete transformation incorporates a pre-processing and post-processing stage dealing with problems such as edge ambiguity. Watershed Transform has its application to gray scale, textual and binary images. [9] has used a technique which consists of Difference In Strength map, K-means and watershed segmentation method to segment the image and detect the edges. The edge maps obtained have no broken lines on entire image and the final edge detection result is one closed boundary as per actual region in the image. [10] discusses the segmentation of touching and broken characters in handwritten Gurumukhi word. The segmentation technique described here is based on neighboring pixels for broken characters and touching characters. This End detection algorithm has achieved the accuracy of 95%**.** [11] shows a technique for segmenting touching characters in upper zone of printed Gurumukhi script. This technique uses the structural properties of the Gurumukhi script. Concavity and convexity of the characters has been analyzed and top profile projections have been used to segment the touching characters in upper zone. Recognition up to 91% has been achieved for segmenting the touching characters in upper zone. [12] shows a new strategy for the segmentation of overlapping characters and conjuncts, in Devanagari script on Hindi language. The algorithm is focused around Cluster Detection technique. This technique segments the middle region of the word accurately. It gave the accuracy of 94.5% for various input characters. [13] defines an algorithm to segment text lines in Gurumukhi handwritten script. This algorithm solves the problem of overlapping of lines and connected components. It is based on inserting a gap between beneficial segments of lines. The obtained results of line segmentation prove that the algorithm is good for overlapped text lines. Still the lines with the broken parts in upper modifiers and lower modifiers are not correctly segmented.

II. FEATURES OF THE GURUMUKHI SCRIPT

Gurumukhi script is the basic script to write Punjabi language. It has also been utilized to write other languages such as Sanskrit. Initially, Gurumukhi character set consisted of 32 consonants and 3 vowel bearers. Later on six additional consonants were added making 41 consonants. Some characters modify the consonants, these are known as half characters or sub joined characters. They are present at the feet of characters. There are ten vowel modifiers, also known as dependent vowels. There are ten independent vowels. There are three auxiliary signs bindi, tippi and addak. Writing style is from left to right and there is no concept of upper or lowercase characters. We can partition Words of Gurumukhi script are divided into three zones horizontally; lower zone, middle zone and upper zone. Middle zone contains consonants. The upper and lower zones may contain parts of vowel modifiers. A large number of characters contain a horizontal line at the upper part of the middle zone.

This line is called the headline. Generally the row with maximum number of black pixels represents the headline. The handwritten script may contain abnormalities like touching characters, overlapping characters, broken characters or skewed characters. The proposed work is on overlapping characters. These are represented in the figures below.

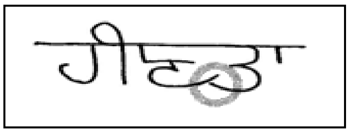


Fig.1 Two Middle Zone Characters Overlap with Each other.

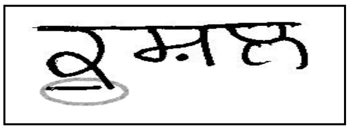


Fig. 2 Lower Zone Character Overlaps with a Middle Zone Character

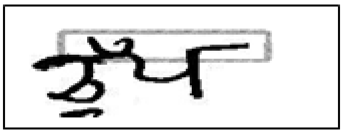


Fig.3 Middle Zone Character Overlaps with Upper Zone Character

III. PROPOSED TECHNIQUE

In the proposed system watershed technique along with horizontal profile projection technique and vertical profile projection technique has been implemented to segment the overlapping handwritten characters from the text document. Firstly, binarization of image has been done using global thresholding function which is followed by the removal of head line. Then further analysis of text is performed using horizontal projection and vertical projection, which is followed by the segmentation using watershed approach.

3.1 Horizontal Projection

The horizontal projection counts the total number of black pixels in each horizontal row. For a binary image of size X* Y where Y is the height of the image and X is the width of image, we define the horizontal projection as HP (j), j=1, 2 …H. With the help of Horizontal Profile projection technique Lines are too extracted from a given paragraph for further use. The horizontal projection profile is represented as a histogram of the number of black pixels along the horizontal axis.

3.2 Vertical Projection

The vertical projection operates on the column and finds the total number of black pixels with respect to column. For a binary image of size X*Y where Y is the height of the image and X is the width of the image, the vertical projection has been defined as VP (k), k=1, 2 ...W. With the help of Vertical Projection technique words from a segmented line can be extracted and character can be segmented from the extracted word. The vertical projection profile is represented as a histogram of the number of black pixels along the vertical axis.

3.3 Watershed Technique

Watershed transform is a powerful tool for solving image segmentation problems. Here we consider image as a surface. Because the regions in the image are characterized by small variations in gray levels having small gradient values, thus, watershed segmentation is applied to image gradient. A watershed of a greyscale image can be considered similar to the catchment basin of a height map. We find local minima which are used to find the limits of the adjacent areas. These limits define watershed lines. The aim of the watershed transform is to search for regions of high intensity gradients that divide neighboured local minima. The watershed transform finds watershed ridge lines in an image by treating it as a surface. High elevations are represented by light pixels and dark pixels represent low elevations. The elements labeled 0 do not belong to a definite watershed region. The elements which are labeled 1 belong to the first watershed region, the elements labeled 2 belong to the second watershed region, and so on. Watershed uses 8-connected neighborhoods for 2-D inputs.

3.4 Steps of Proposed Algorithm

1. Scan the handwritten text document of Gurumukhi script, set the threshold value of scanned input and binarize the given image.
2. Calculate the occurrence of black pixels starting from first row horizontally and find the row with maximum number of black pixels and threat it as header line, remove it by converting black pixels to white.
3. Using vertical profile projection technique finds the gap between the two characters by comparing its neighbour pixels along horizontal direction to check whether the character is broken or not. If gap is less than or equal to the expected value then consider it as same character.
4. Calculate the distance of every pixel to its nearest non zero valued pixel i.e. for ith pixel to next non-zero (i+1) and previous (i-1). Every single valued pixel has a distance transform value of 0.
5. Find the distance transform of image which is in the form of matrix. The elements of matrix are integer values.
6. The transform returns the ridge lines showing the different regions of the word.
7. Form the different ridge lines we get different characters.

3.5 Database Collection

The database is collected in the form of words written by different writers from various backgrounds. Database also includes data of different size and resolution. After that written text is scanned and all the steps are implemented on this scanned text. 200 words are tested independently including isolated, overlapping and skewed characters. No pre-processing except noise removal is performed on input data. We have collected data from different writers such as to test various hand writings.

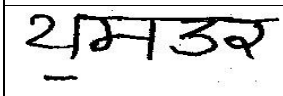
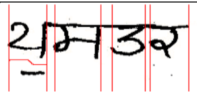
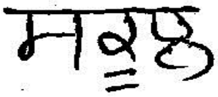
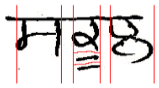
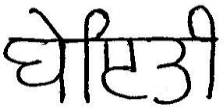
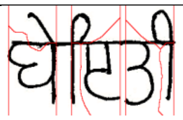
IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

For testing the proposed algorithm, various handwritten words of Punjabi language have been considered. Different structural properties are taken care of to maintain the originality of the characters. The collected database is provided to the system, so as to identify whether the proposed method will give efficient results or not. The proposed algorithm gives efficient results for isolated, overlapping and skewed characters. Here are a few results for different input characters:

4.1 Isolated Characters

The first type of data input taken consists of isolated or disconnected characters, which mean the words holding the characters with legitimate spaces in between. The segmentation of these words is relatively less difficult than other words. We have achieved the accuracy of 96% in this case. The table 1 below represents the result for isolated characters:

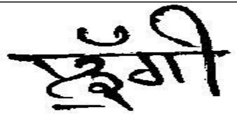
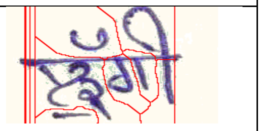
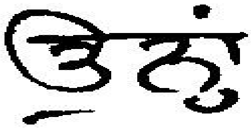
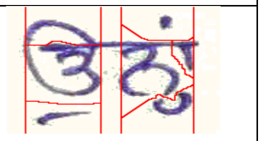
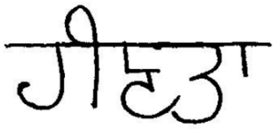
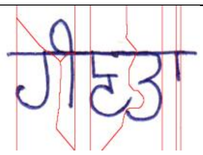
TABLE 1 Results for Isolated Characters

Input	Output
	
	
	

4.2 Overlapping Characters

We say that the characters are overlapping when one character comes in the zone of another character, which means that we cannot separate the two characters with the straight line. Proposed algorithm mainly deals with this sort of character segmentation. It can be the overlapping of two middle zone characters, a middle zone character and a lower zone character or a middle zone character and an upper zone character and; two upper zone characters. Very little work has been done so far related to overlapping characters. We have achieved the accuracy of 93%. We have considered the overlapping characters of all the types. The table 2 above represents the results for overlapping characters.

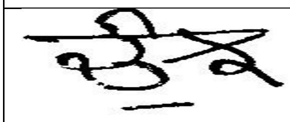
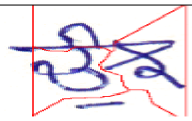
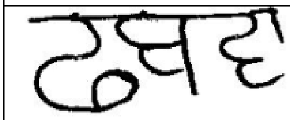
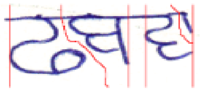
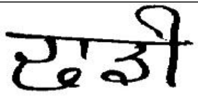
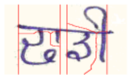
TABLE 2 Results for Overlapping Characters

Input	Output
	
	
	

4.3 Skewed Characters

Skewness is the measure of asymmetry of the character. Character may be skewed vertically or horizontally. It decreases the efficiency of recognition process. It has been solved in the case of headline within certain threshold limit i.e. horizontal skewness is restricted and maximum cases of vertical skewness are dealt with. Within the specified extent of skewness we have achieved the accuracy of 96%. The table 3 represents the results for the sekwed characters.

TABLE 3 Results for the Skewed Characters

Input	Output
	
	
	

Overall result for all the categories of words is represented in the table 4 given below. Accuracy of 95% has been achieved.

TABLE 4. Overall Result for the Different Category of Words

Type of input words with	**Total no of words**	**Correctly Segmented Words**	**Accuracy**
Isolated characters.	50	48	96%
Overlapped characters.	100	93	93%
Skewed characters	50	48	96%
Overall segmentation	200	188	95%

TABLE 5 Comparison of Existing Techniques

Ref.	Technique used	Type of input	Accuracy
Kumar et. al	Water reservoir technique [3]	Isolated and touching Gurumukhi characters	93.5%
Mangla et. al	Analysis of neighbouring pixels [10]	Touching and broken characters in Gurumukhi	95%
Thakral et. al	Cluster detection tech. [12]	Hindi touching, overlapping and conjucts	94.5%
Sharma et. al	Horizontal and vertical projection profile[15]	Simple Gurumukhi text	96.2%
Kumar et. al	Variable size window [16]	Isolated characters	90%

V. CONCLUSION AND FUTURE SCOPE

In this paper, we have proposed a method for the segmentation of overlapping characters and skewed characters of Gurumukhi script. We have used horizontal profile projection, vertical profile projection and watershed technique in our proposed work. We have tested our system on 200 words written by different writers, which includes isolated, overlapping and skewed characters. We have achieved the accuracy of 96% in the case of isolated characters, 93 % in the case of overlapping characters and 96% in the case of skewed characters within specified threshold. Overall accuracy of 95% has been achieved. The proposed system can further be extended to include touching and broken characters, so that all the segmentation problems can be tackled by single algorithm. It can also be implemented on different scripts.

REFERENCES

[1]. R. G. Casey, and E. Lecolinet, “A Survey of Methods and Strategies in Character Segmentation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, No.17, July 1996.

[2]. D. V. Sharma, and G.S. Lehal, “An Iterative Algorithm for Segmentation of Isolated Handwritten Words in Gurumukhi Script”, IEEE 18th International Conference on Pattern Recognition, 2006.

[3]. M. Kumar, M.K. Jindal, and R.K. Sharma, "Segmentation of Isolated and Touching Characters in Offline Handwritten Gurumukhi Script Recognition", International Journal Information Technology and Computer Science, PP: 58-63, Feb, 2014.

[4]. G. Bansal, and D. V. Sharma, "Isolated Handwritten Words Segmentation Techniques in Gurumukhi Script", International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 24, PP: 104-111, 2010.

[5]. M. K. Jindal, R. K. Sharma, and G. S. Lehal, "Segmentation of Horizontally Overlapping Lines in Printed Gurumukhi Script", IEEE, Vol.8, No.6, pp. 226–229, 2006.

A. S. Ramteke, and M. E. Rane, "Offline Handwritten Devanagari Script Segmentation", International Journal of Scientific & Technology Research, Vol. 1, Issue 4, PP: 142-145, May, 2012.

[6]. N. K. Garg, L. Kaur, and M.K. Jindal, "Segmentation of Handwritten Hindi Text", International Journal of Computer Applications, Vol. 1, No. 4, PP: 19-22, 2010.

A. Kaur, and Aayushi, "Image Segmentation Using Watershed Transform", International Journal of Soft Computing and Engineering, Vol. 4, No. 1, pp. 5-8, March 2014.

[7]. N. Salman, "Image Segmentation Based on Watershed and Edge Detection Techniques", the International Arab Journal of Information Technology, Vol. 3, No. 2, pp. 104-110, April 2006

[8]. P. Mangla, and H. Kaur., "An End Detection Algorithm for segmentation of Broken and Touching characters in Handwritten Gurumukhi Word", IEEE International Conference on Reliability, Infocom Technologies and Optimization, pp. 1-4, 2014.

[9]. G.S lehal, R. K. Sharma, and M. K. Jindal, "A Segmentation of Touching Characters in Upper Zone in Printed Gurumukhi Script, in Compute, Bangalore, Karnataka, India, Jan 2009.

[10]. B. Thakral, and M. Kumar, "Devanagari Handwritten Text Segmentation for Overlapping and Conjunct Characters: A Proficient Technique", IEEE International Conference on Reliability, Infocom Technologies and Optimization, ICRITO, pp. 1-4, 2014.

[11]. N. Modi and K. Jindal, "Text Line Detection and Segmentation in Handwritten Gurumukhi Scripts", IJARCSSE, Vol. 3, pp. 1075-1080, 2013.

[12]. R. Kumar, and A. Singh, "Detection and Segmentation of Lines and Words in Gurumukhi Handwritten Text", IEEE 2nd International Advance Computing Conference, pp. 353-356, 2010.

[13]. D. V. Sharma, and P. Jhajj, "Comparison of Feature Extraction Methods for Recognition of Isolated Handwritten Characters in Gurumukhi Script", Springer-Verlag Berlin, pp. 110–116, 2011.

[14]. M. Kumar, M. K. Jindal , R.K.Sharma, "segmentation of Isolated and Touching Characters in Offline Handwritten Gurumukhi Script Recognition." in IJ Information technology and computer science , 2014.