Segmentation of Isolated and Touching characters in Handwritten Gurumukhi Word using Clustering approach

Akashdeep Kaur
M.Tech Student
GZS PTU Campus, Bathinda
akashbrar702@yahoo.com

Dr.Shaveta Rani
(Associate Professor)
GZS PTU Campus, Bathinda
garg_shavy@yahoo.co.in

Dr. Paramjeet Singh
(Associate Professor)
GZS PTU Campus, Bathinda
param2009@yahoo.com

***Abstract*—**Segmentation is one of the important steps of character recognition system. It is an important step because inaccurate segmentation of characters will cause errors in the recognition stage. In optical character Recognition (OCR) system the presence of touching characters decreases the accuracy rate of character recognition. Touching of half character or full character with other full character makes the character segmentation very challenging or difficult task. In this paper, the method of segmentation for touching characters of handwritten Punjabi text that is the Gurumukhi script has been proposed. The main purpose of this paper is to provide the new segmentation technique based on clustering technique for touching characters.

Keywords— *Segmentation, Feature Extraction, Binarization, Classification,proposed work, Results*

I. INTRODUCTION

Computer is a very useful electronic machine in today's life. It is mainly used to read or to edit various documents. Some documents that are old, they are scanned to store in the computer. After scanning these documents they are stored in the form of images that could not be edited or Recognized[2]. So OCR is a useful invention that helps to read and recognize the scanned documents. The main aim of OCR is to convert the scanned documents into readable or editable format. Steps in OCR

- Pre-processing
- Segmentation
- Feature extraction
- Classification
- Recognition

Segmentation is an important pre-processing phase in the character recognition. Segmentation means to separate the various characters from one another so that they can be recognized accurately. The accuracy of character recognition depends on the segmentation. Segmentation of printed character is easy as compare to handwritten characters. The main challenge in the segmentation of handwritten language is that there are wide varieties of Styles or pen-type[2]. Presence of touching makes the segmentation of handwritten Gurumukhi characters very difficult and decreases the accuracy rate of recognition. Currently there are three methods of segmentation.

- Classical approach-the characters are identified based on their properties.
- Recognition based approach-the components of an image are matched with the various classes of alphabets.
- Holistic approach- the words are recognized as a whole instead of segmenting them into characters.

II. CHARACTER SEGMENTATION

Character segmentation is the technique used to separate the various characters from one another. There are various algorithms to segment the Gurumukhi handwritten words into character but there can be the touched characters in Gurumukhi word due to the segmentation of character becomes very tough[1]. So in this paper efforts have been made to overcome the problem of touched characters. An algorithm has been developed that will

1. Segment the isolated characters.
2. Identify the presence of touching characters.
3. Find out the break point.
4. Segment the touched characters.

Some of the examples of isolated, touched characters are shown:

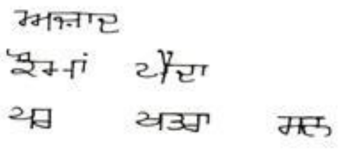


Fig.1. Handwritten Gurumukhi words containing isolated, touching characters.

To implement this algorithm the following various steps have been performed:

III. BINARIZATION

Scanning image: In this step the document is converted into scanned image with the help of image scanner.

Binarization: In this step gray scale images are converted to binary image with the help of OCR Software[3]. The images that are scanned are in the grey tone. Basically a Binarization is the process in which the grey scale images are converted into binary form means in the form of 0's and 1's.

Binarization separates the foreground (text) and background. There are various methods for binarization but the most common method for binarization is to select the proper threshold for the intensity for an image and then convert all the intensity values above the threshold to one intensity value (white) and all intensity values below the threshold to other chosen intensity (black).

IV. REMOVAL OF HEADER LINE

After removing the available noise from an image, the header line is removed header line is an important part of a Gurumukhi word that glues all the words together. So after detecting and removing the header line the characters can be segmented easily.

Steps:

- Calculate the frequency of black pixels in each row along with neighbors using horizontal profile projection technique[2].
- Find the row with highest number of black pixels and treat that row as header row.
- Replace all the 1's in that row with 0's

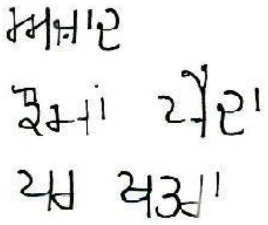


Fig. 2 Words after removal of header line

V. SEGMENT THE CHARACTERS

After the removal of header line the phase of character segment starts. Then use vertical profile projection technique parse the word column wise and Check for each ith column of the word if all the pixels are white and if so then check i-1 and i+1 number of pixels. If all three pixels are white then treat them as gap between two characters and then segment the word. After the character segmentation results are:

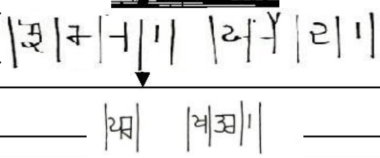


Fig.3. Words after initial

.

A. Identifictaion of touching characters

Now after the segmentation of characters, the next step is to find that whether there is any touched character or not. This is done by calculating the end of character by estimating its structural properties. In case two characters are touched in one word then assume maximum pixels and find another end of character and then break it and gets segmented.

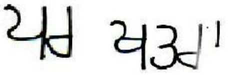


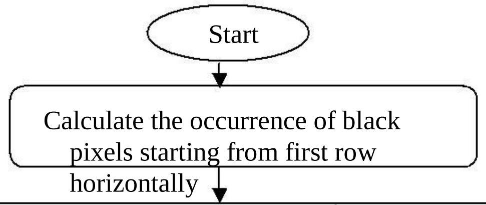
Fig4. Identification of touched characters

B. Segmentation of touching characters

Now it is determined that which character is the touched character. So now there is need to break those characters to separate them. To break the touched characters there is need to find the end of character. For this, following steps have been performed:

- Start with the mid end of character.
- Assume frequency of pixels vertically where characters touch and make a cluster is 10.
- If frequency of cluster vertically is greater than 10 it means it is touched.
- Find end of character of the previous character and segment the word from that position.

End detection Algorithm in form of flow chart that can segment isolated and touching character is shown as below:

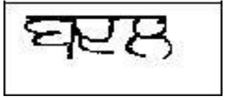

Start
Calculate the occurrence of black pixels starting from first row horizontally

segmentation

VI. RESULTS

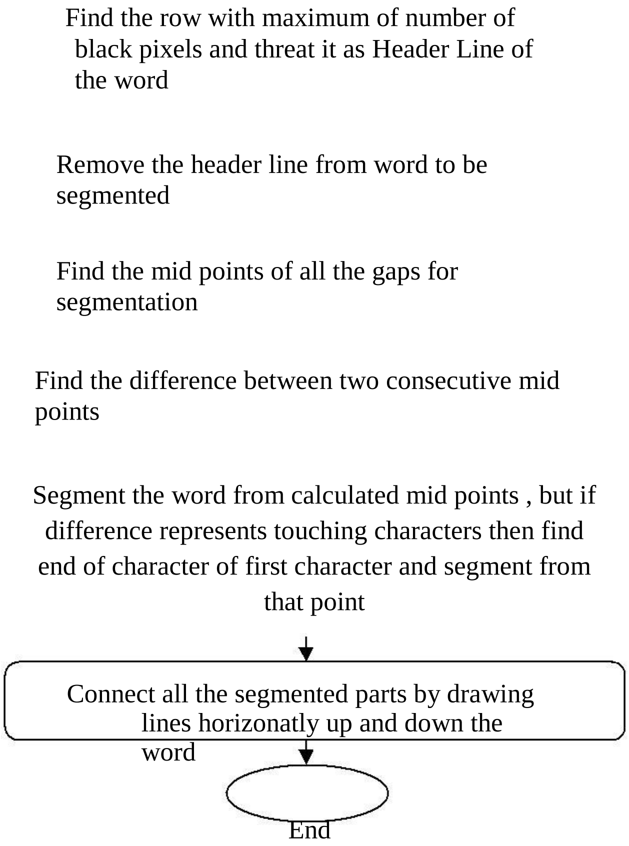
In order to detect and segment characters in scanned word of handwritten Gurumukhi script documents, neighboring pixel and end of character technique have been used. These techniques have been applied on the documents of three different categories. The category wise results of segmentation accuracy are given in following table.

Input images:



Output images:

VII DISSCUSSION AND CONCLUDING REMARKS



Find the row with maximum of number of black pixels and threat it as Header Line of the word

Remove the header line from word to be segmented

Find the mid points of all the gaps for segmentation

Find the difference between two consecutive mid points

Segment the word from calculated mid points , but if difference represents touching characters then find end of character of first character and segment from that point

Connect all the segmented parts by drawing lines horizonatly up and down the word

End

Here this algorithm has been tested on 75 handwritten words taken from different people with different handwriting. In which there was isolated and touching characters.

Phases	Words	Correctly segmented	%age
Phase1:words without any touching,broken or overlapped characters.(ISOLATED)	75	75	100%
Phase2: words with isolated, touching characters more than one character in one word (TOUCHING)	50	48	94.61%

Different phases of words showing accuracy.

In the 2nd phase the words with touching characters are handled with 94.51 accuracy and the remaining (6%) error is due to words touched with" kanna". The errors of over-segmentation were unavoidable because of the gaps in the broken characters. Any readjustment of the threshold value leads to high degree of undersegmentation in the words and therefore is not recommended.

VII. REFRENCES

[1] Munish kumar , Mk jindal , R.K.Sharma “segmentation of Isolated And Touching Characters in Offline Handwritten Gurumukhi Script Recognition.” in IJ Information technology and computer science , 2014.

[2] Simpel rani, Arbha Goyal “An efficient approach for segmentation of touching characters in handwritten hindi word”. In International conference of on Information and mathematical Sceinces, 2014 ELESVIER.

[3] Nabin Sharma, Palaiahnakote Shivakumara, Umapada Pal, Michael Blumenstein And Chew Lim Tan ,”A New Method For Character Segmentation From Multi-Oriented Video Words” In 2013 IEEE.

[4] G.S lehal, R. K. Sharma, and M. K. Jindal, “A Segmentation of Touching Characters in Upper Zone in Printed Gurmukhi Script, in Compute 2009,Jan9, 10, Bangalore, Karnataka, India.

[5] Ashwin S Ramteke, Milind E Rane,”Offline Handwritten Devanagari Script Segmentation” in 2012

[6] G.S lehal, R. K. Sharma, and M. K. Jindal, “Segmentation Problems and Solutions in Printed Degraded Gurmukhi Script”, International Journal of Signal Processing Volume 2 Number 4

[7] Sandeep N.Kamble, Prof. Megha Kamble, “ Morphological Approach for Segmentation of Scanned Handwritten Devnagari Text” in Oct-Dec, 2011, vol. 2

[8] G.S lehal, and Chandan singh, “A post-processor for Gurmukhi OCR”, in Sadhana Vol. 27, Part 1, February 2002, pp. 99–111. © Printed in India

[9] Naresh Kumar Garg, Lakhwinder Kaur & M.K. Jindal ,”The Hazards in Segmentation of Handwritten Hindi Text” International Journal of computer Applications(0975-8887) in Sep 2011, vol 29-No.2

[10] Galaxy Bansal ,Daramveer Sharma ,”isolated handwritten words segmentation techniques in gurumukhi script” 2010International Conference in computer sceince and its applications .

[11] G.S lehal and Daramveer sharma, “An Iterative Algorithm for Segmentation of Isolated Handwritten Words in Gurmukhi Script, in

[12] The 18th International Conference on Pattern Recognition (ICPR'06) 0-7695-2521-0/06 $20.00 © 2006 IEEE

[13] G.S Lehal, R. K. Sharma, And M. K. Jindal, “On Segmentation Of Touching Characters And Overlapping Lines In Degraded Printed Gurmukhi Script”, International Journal Of Image And Graphics Vol. 9, No. 3 (2009) 321–353 World Scientific Publishing Company.

[14] Naresh Kumar Garg , Lakhwinder Kaur and M.K. Jindal “Segmentation of Handwritten Hindi Text” International Journal of computer Applications(0975-8887) in 2010 , vol. 1-No. 4

[15] Vijay Kumar, Pankaj K. Sengar,” Segmentation Of Printed Text In Devanagari Script And Gurmukhi Script” International Journal Of Computer Applications (0975 – 8887) Volume 3 – No.8, June 2010.

[16] K. Wong, R. Casey and F. Wahl “Document Analysis System “, IBM j.Res . Dev., 26(6), pp.647-656, 1982.

[17] F. Hones and J. Litcher, “Layout extraction of mixed mode documents”, Machine Vision Application, vol. 7, pp. 237–246, 1994.