Volume 5, Issue 7, July 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Segmentation of Broken and Touching Characters in Handwritten Gurumukhi Word using Clustering Approach

Jatinder Kaur, Er. Navdeep Singh Sethi
Computer Science & Engineering
AIET, Faridkot, Punjab, India

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 and broken 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. Segmentation of the broken character is quite difficult because vertical profile projection technique assumes the broken parts of the characters as individual characters. In this paper, we have proposed the method of segmentation for touching and broken characters of handwritten Punjabi text that is the Gurumukhi script. The main purpose of this paper is to provide the new segmentation technique based on neighboring pixels for broken characters and increase the accuracy for touching characters.

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

I. INTRODUCTION

Segmentation is one of the testing and crucial fields in OCR. It is an operation that tries to deteriorate an image of arrangement of characters into sub images of individual symbols. It is one of the decision procedures in OCR. In text report image examination, the significant step is extraction of text lines from documents, and afterward the text lines are divided into words and characters.

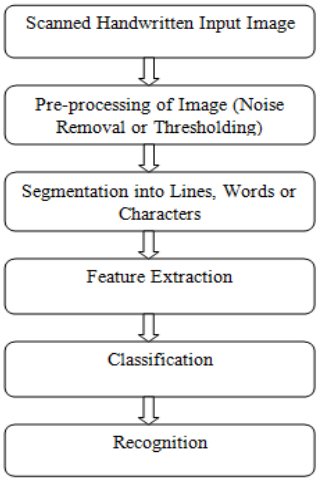


Fig. 1: Optical Character Recognition System

Categorization of OCR is focused around two primary criteria: Data Acquisition process and Type of Text composed. Data Acquisition process incorporates online and offline information, where online information speaks to handwriting that is recorded with a digitizer, as a period gathering of pen headings and offline speaks to the hard duplicate of handwriting scanned by optical scanner or camera. Text Type incorporates Machine printed and Handwritten content, Machine printed content holds the materials, for example, books, daily papers, magazines, archives, and different composition units in the feature or still image. Machine printed characters are uniform in stature, width, and pitch expecting the same textual style and size are utilized. Handwritten incorporates the content composed by distinctive writers by their hands. There are loads of varieties in handwriting of diverse users.

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 as well as broken characters in Gurumukhi word due to the segmentation of character becomes very tough[1]. So in this paper we have tried to overcome the problem of touched and broken characters. We have developed an algorithm that will

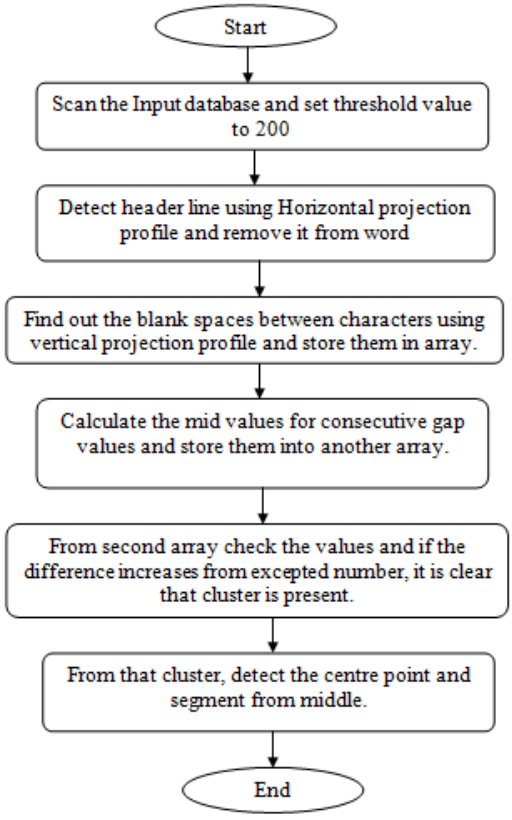
1 Segment the isolated characters.
2 Identify the presence of touching characters.
3 Find out the break point.
4 Segment the touched characters.
5 Identify the broken characters.
6 Segment the broken characters.

III. METHODOLOGY

Implementation steps:

Here is step by step procedure for segmentation of challenging characters in a word:

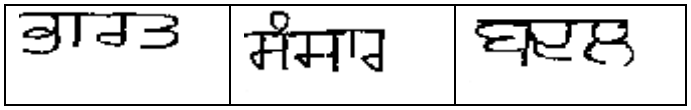
- Scan each row horizontally, by HPP and detect the header-line, i.e. the line with maximum number of black pixels and remove the header-line.
- After that traverse every pixel, check its next and previous pixel to determine if any character is broken or not.
- Detect all the gaps, i.e. empty spaces between characters and store them into array.
- Calculate the mid value for all consecutive gap values from the array and store them into another array.
- Assumption for the system is that the width of each character should be less than 25 pixels, if the difference between these mid values increases the expected number; then it is considered that cluster is present.
- From that heap of pixels (cluster), calculate the mid again and segment the character from that point by drawing vertical lines in between.



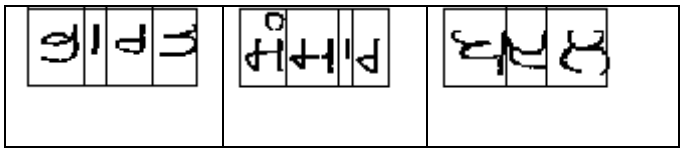
IV. RESULTS

In order to detect and segment characters in scanned word of handwritten Gurumukhi script documents, we have used neighboring pixel and end of character technique. These techniques have been applied on the documents of three different categories. The category wise results of segmentation accuracy are given in Table 1.

Input images:



Output images:



COMPARISION OF VARIOUS TECHNIQUES FOR CHARACTER SEGMENTATION FOR GURUMUKHI SCRIPT

TECHNIQUE USED	OVERALL ACCURACY
IMPROVED VARIABLE SIZED WINDOWS CONCEPT	90%
WATER RESERVIOR METHOD	87%
PROFILE PROJECTION TECHNIQUES	82%
PROPOSED SYSTEM	96%

V. DISSCUSSION AND CONCLUDING REMARKS

Here we have tested this algorithm on 150 handwritten words taken from different people with different handwriting. In which there was isolated, broken and touching characters.

Table 1: Different phases of words showing accuracy.

Phases	**Words**	**Correctly segmented**	**%age**
Phase1:words without any touching,broken or overlapped characters.(ISOLATED)	**50**	**50**	**100%**
Phase2: words with isolated, touching characters more than one character in one word(TOUCHING)	**50**	**48**	**96%**
Phase 3: words with isolated, broken in one word(BROKEN)	**50**	**47**	**94%**

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". In the third phase the words with broken characters are handled and of these 46 (94% of 50) words are properly segmented and the remaining (6%) error was primarily because of overlapping characters with broken characters. 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 under-segmentation in the words and therefore is not recommended.

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.

