Volume 5, Issue 3, March 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Review Paper on Character Segmentation in Handwritten Text Document Written in Gurumukhi Script

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

Abstract: Character Segmentation plays very important role to process or store these documents electronically. If character segmentation gives the accurate results only then we can recognize the characters accurately and hence we can convert them into machine readable format. In this paper we present the review on various techniques for character segmentation for Gurumukhi Script. Accuracy of this phase plays a vital role in overall OCR process. If this phase generates good results only then next phases can accurately work.

Keywords: Character segmentation, OCR, Gurumukhi Script.

I. INTRODUCTION

Optical character recognition (OCR) is the process of converting scanned images of machine printed or handwritten text into a computer processable format. It involves computer software designed to translate images of typewritten text into machine-printed editable text, or to translate pictures of characters into a standard encoding scheme representing them in ASCII or Unicode. If you scan a text document, you might want to use optical character recognition (OCR) software to translate image into text that you can edit. When a scanner first creates an image from page, image is stored in computer's memory as a bitmap. A bitmap is a grid of dots; one or more bits represent each dot. The job of OCR software is to translate that array of dots into text that computer can interpret as letters and numbers.

II. CHARACTER SEGMENTATION

Character segmentation is the term, which covers all types of machine recognition of characters in various application domains. The intensive research effort on the field of character segmentation was not only because of its challenge on simulation of human reading, but also, because it provides efficient applications such as the automatic processing of bulk amount of papers, transferring data into machines and web interface to paper documents . A character segmentation system can be either "online" or "offline." According to the mode of data acquisition, character segmentation methodologies are categorized into two systems as **ONLINE CHARACTER SEGMENTATION SYSTEMS and OFFLINE CHARACTER SEGMENTATION SYSTEMS:**

Online character segmentation is the process of segmenting handwriting, recorded with a digitizer, as a time sequence of pen coordinates. It captures the temporal and dynamic information of the pen trajectory. Applications of on-line character segmentation systems include small handheld devices, which call for a pen-only computer interfaces and complex multimedia systems, which use multiple input modalities including scanned documents, speech, keyboard and electronic pen. These systems are useful in social environments where speech does not provide enough privacy. Pen based computers, educational software for teaching handwriting and signature verifiers are the examples of popular tools utilizing the on-line character segmentation techniques.

Offline character segmentation is the process of converting the image of writing into bit pattern by an optically digitizing device such as optical scanner or camera. The segmentation is done on this bit pattern data for machine-printed or handwritten text. Applications of offline segmentation are large-scale data processing such as postal address reading; check sorting, office automation for text entry, automatic inspection and identification. Offline character segmentation is a very important tool for creation of the electronic libraries. Also, the wide spread use of web necessitates the utilization of offline segmentation systems for content based Internet access to paper documents.

According to the text-type, HANDWRITTEN and MACHINE -PRINTED CHARACTER SEGMENTATION SYSTEMS are two main areas of interest in character recognition field:

Machine printed text includes the materials such as books, newspapers, magazines, documents, and various writing units in the video or still image. Machine printed characters are uniform in height, width, and pitch assuming the same font and size are used. These problems for fixed- font, multi-font and omni-font character segmentation is relatively well understood and solved with little constraint.

Handwritten text can be further divided into two categories: cursive and hand printed script. Recognition of handwritten characters is a much more difficult problem. Characters are non uniform and can vary greatly in size and style. Even characters written by the same person can vary considerably. In the location of characters is not predictable, nor the spacing between them. In an unconstrained system, characters may be written anywhere on the page and may be

overlapped or disjoint. A typical segmentation system will require some sort of constraints, or added information, about the data being processed.

Vikas J Dongre, Vijay H Mankar in 2010,"A Review of Research on Devnagari Character Recognition", In this paper, recognition of handwritten character is presented. There are five steps for the recognition of character recognition: 1) Pre-processing of image 2) Segmentation of words into characters 3) Feature Extraction 4) Reorganization 5) Post processing.

Naresh Kumar Garg, Lakhwinder Kaur & M.K. Jindal in 2011,"The Hazards in Segmentation of Handwritten Hindi Text" ,OCR is used to recognize the sccaned text that can be in the form of handwritten or typed form. Segmentation is the important phase in the character recognition that can improve/decrease the aacuracy of character recognition. Segmentation of printed words is quite easy as compare to handwritten words because of the various problems that will occur in the segmentation of handwritten text. There are two types of problems that can occur in the segmentation of handwritten text: 1) The Problems that can be ignored (Like the problems due to speed of writing). 2) The Problems that can be ignored.

Ashwin S Ramteke, Milind E Rane in 2012,"Offline Handwritten Devanagari Script Segmentation", The process of Segmentation is a vital phase in the recognition of text. Devanagari is very useful Script in India. The segmentation of devanagari words is very difficult due to the presence of large character set that include consonants, vowels and modifiers. In this paper the major focus was on the segmentation of line, word and characters. Before the segmentation of an image some pre-processing of the image is done using the median filter and it aslo includes the binarization and scaling of image. After this preprocessing the segmentation is done. For the Segmentation of handwritten Devanagari script the histrogram of input image is generated that shows the space b/w the characters so from this the characters can be segmented.

Sandeep N.Kamble, Prof. Megha Kamble in 2011, " Morphological Approach for Segmentation of Scanned Handwritten Devnagari Text" This papers deals with the segmentation of modifiers and fused characters in handwritten words by segmenting the Words in hierarchical order: (a)segment the header Lines (b) segment the top modifiers (c) segment the bottom modifiers (d) segment the fused characters.

III. EXISTING TECHNIQUES

The following techniques will mainly used to segment the words into characters:

Horizontal Projection Profile (HPP): For a binary image of size H*W where H is the height of the image and W is the width of image, the horizontal projection is defined as HP (j), j=1, 2 ... H. This operation counts the total number of black pixels in each horizontal row. In this research work, HPP method is used to detect the header line from the input text image and convert it into white empty pixels.

Vertical Projection Profile (VPP): For a binary image of size H*W where H is the height and W is the width of the image, the vertical projection is being defined as VP (k), k=1, 2 ... W. This operation counts the total number of black pixels in each vertical column. Sometimes due to some irregularities the system will detect multiple header lines in single image, so VPP is used to overcome this and to extract the characters from the word.

Pixel Cluster Identification technique . This method is based on the fact that when two characters overlap or touch each other they form a cluster of pixels on the point where they touch Here cluster means the heap of pixels that increases the expected value of number of pixels. The pre assumption in this method is that a single character consists of maximum number of 20-25 pixels. So if this value increases by the expected number it is assumed that a single character is being touched or overlapped by another one. These characters can be conjuncts, touching or overlapping with one another and after that segmentation will be carried out.

IV. CONCLUSION

In this paper , various methods for character segmentation have been discussed. It is concluded that process of character segmentation have to face many problems like irregular size of characters, touching characters , overlapping characters etc. Existing techniques can not solve all these problem. Hence in future a system is required that can solve all these problems.

REFERENCES

[1] U. Pal and B.B. Chaudhuri, "Indian script character recognition: A Survey", Pattern Recognition, vol. 37, pp. 1887-1899, 2004.

[2] B. B. Chaudhuri and U. Pal, "A complete printed Bangla OCR system", Pattern Recognition, vol.31, pp.531-549, 1998.

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

[4] Likforman-Sulem, L., Zahour, A. and Taconet, B., "Text line Segmentation of Historical Documents: a Survey", International Journal on Document Analysis and Recognition, Springer, Vol. 9, Issue 2, pp.123-138, 2007.

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

[6] K. Kise, W. Iwata, and K. Matsumoto, "A computational geometric approach to text line extraction from binary document images", in Proc. IAPR Workshop Document Analysis Systems, pp. 364-375, 1998.

[7] Vikas J Dongre, Vijay H Mankar ,"A Review of Research on Devnagari Character Recognition" International Journal of computer Applications(0975-8887) in Nov 2010, vol 12-No.2

[8] 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

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

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

[11] T.K. Bhowmik , A. Roy and U.Roy "Character Segmentation for Handwritten Bangla Words Using Artificial Neural Network"

[12] Debapratim Sarkar, Raghunath Ghosh "A Bottom-up Approach of Line Segmentation from Handwritten Text"

[13] M.Hanmadlu and Puja Agrawal " Segmentation of Handwritten Hindi Text : A Structural Approch" Department of Electrical Engineering, Indian Institute of Technology, Delhi 110016

[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

