SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM

TNC Transactions on Networks and Communications

VOLUME 3, ISSUE 2
ISSN: 2054 - 7420

Segmentation of Broken and Isolated characters in Handwritten Gurumukhi Word using Neighboring pixel technique

Akashdeep Kaur, Paramjeet Singh and Shaveta Rani
Giani Zail Singh Punjab Technical University Campus, India
Akashbrar702@yahoo.com

ABSTRACT

Character Segmentation of Handwritten Documents has been an active area of research and due to its diverse applicable environment; it continues to be a challenging research topic. In this paper, the focus is on offline segmentation of handwritten documents written in Gurumukhi Script. The desire to edit scanned text document forces the researchers to think about the optical character recognition (OCR). OCR is the process of recognizing a segmented part of the scanned image as a character. OCR process consists of three major sub processes - pre processing, segmentation and then recognition. Out of these three, the segmentation process is the most important phase of the overall OCR process. In this paper, algorithm is formulated to segment the scanned document image as a character that can be isolated or broken from within the given word. According to proposed algorithm, one part is extracting line from a document other part is extracting a word from the line. Segmentation part of the algorithm extracts characters from the extracted word. To segment the characters from a word, combination of two approaches which are Horizontal Profile Project and Vertical Profile Projection is used and will formulate a new algorithm which is Neighboring Pixel algorithm for touching characters in a word written in Gurumukhi script.

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

1 Introduction

Transmission and storage of information is done not only through computers but also through paper documents. To integrate these two mediums of information flow, a solution is for computer to “read” paper documents. Machine simulation of human reading is one of the areas, which has been the subject of intensive research for the last three decades, yet it is still far from the final frontier. So, works are still going on this direction.

1.1 Natural Language Processing

Natural language processing is a field of science and linguistics concerned with the interaction between the Computers and human languages. Natural language generation systems convert information from computer databases into readable human language. The term “natural” language refers to the languages that people speak, like English and Japanese and Hindi, as opposed to artificial languages like programming languages or logic. “Natural Language processing”, programs that deal with natural language in some way or another. The study of human languages developed the concept of communicating with non-human devices.

DOI: 10.14738/tnc.32.1094
Publication Date: 15th April, 2015
URL: http://dx.doi.org/10.14738/tnc.32.1094

NLP deals with the Artificial Intelligence under the main discipline of Computer Science. The goal of NLP is to design and build software that will analyze, understand and generate languages that humans use naturally.

There are many applications of Natural Language processing developed over the years. The main applications are text-based, which involves searching for a certain topic or a keyword in a large document, translating one language to another or summarizing text for different purposes.

2 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

Offline character segmentation systems

2.1 Online 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.

2.2 Offline character segmentation systems

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.

3 Binarization

3.1 Scanning image:

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

3.2 Binarization:

In this step gray scale images are converted to binary image with the help of OCR Software [10]. 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).

4 Literature Survey

Vikas J Dongre, Vijay H Mankar[7] 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 [8] in 2011,"The Hazards in Segmentation of Handwritten Hindi Text" ,OCR is used to recognize the scanned 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 accuracy 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 cannot be ignored.

Ashwin S Ramteke, Milind E Rane [9] 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 also includes the binarization and scaling of image. After this pre-processing the segmentation is done. For the Segmentation of handwritten Devanagari script the histogram of input image is generated that shows the space b/w the characters so from this the characters can be segmented.

5 Identify the Presence of Broken Characters

Now after the segmentation of various characters the next step is to find that whether there is any broken character or not. Character can be broken due to writer's pen or page quality used. Segmentation of the broken character is quite difficult because vertical profile projection technique assumes the broken parts of the characters as individual characters and thus segmenting the word as separate character. So neighboring pixel technique is used to identify the broken character. So by concentrating on this feature, following steps are performed:

- Check the Neighboring pixels on both left and right side.
- If the black pixels are there then that represents the character is broken and not to be segmented.
- But if there are white pixels in its neighbor then these
- Pixels are treated as a gap and hence to be segmented.

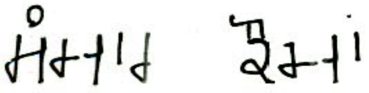


Figure 1: Identification of broken character

6 Segmentation of Broken Characters

Now after the previous steps it is determined that which character is the broken character. So now there is need to make that broken characters as one character. This is done by scanning the neighboring pixels before segmenting the word into character. For this, following steps are performed:

- 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.
- Check for the two pixels (i-1, i+1)
- If they are black, than it represents the broken character and don't segment the word from the ith pixel.

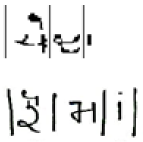
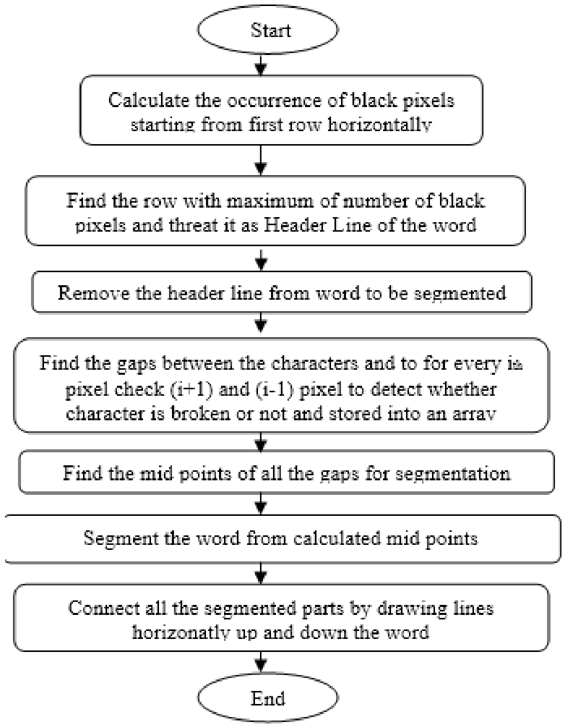


Figure 2: Segmentation of broken words

Neighboring pixel algorithm that can segment isolated, broken character is shown as below:


Start
Calculate the occurrence of black pixels starting from first row horizontally
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 gaps between the characters and to for every ith pixel check (i+1) and (i-1) pixel to detect whether character is broken or not and stored into an array
Find the mid points of all the gaps for segmentation
Segment the word from calculated mid points
Connect all the segmented parts by drawing lines horizonatly up and down the word
End

7 Results

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

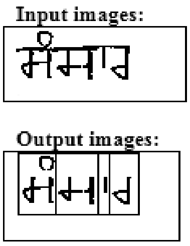


Figure 3: Input and Output images of broken words

8 Disscussion and Conclusion

Here this algorithm has been tested on 25 handwritten words taken from different people with different handwriting. In which there was isolated and broken characters.

Table 1: Different phases of words showing accuracy.

Phases	Words	Correctly segmented	%age
Phase 1:Words without an broken characters. (ISOLATED)	**75**	**75**	**100%**
Phase 2:Words with isolated, broken in one word(BROKEN)	**50**	**46**	**94%**

In the second 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] 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

[2] G.S lehal and Daramveer sharma, “An Iterative Algorithm for Segmentation of Isolated Handwritten Words in Gurmukhi Script”The 18th International Conference on Pattern Recognition (ICPR'06), IEEE 2006.

[3] G.S lehal, R. K. Sharma, and M. K. Jindal, “Segmentation Problems and Solutions in Printed Degraded Gurmukhi Script”, World Academy of Science Engineering and Technology Volume 2, 2008.

[4] Naresh Kumar Garg, Lakhwinder Kaur and M.K. Jindal “Segmentation of Handwritten Hindi Text” International Journal of computer Applications, vol. 1-No. 4,pp19-23,2010.

[5] Galaxy Bansal, Daramveer Sharma,"Isolated handwritten words segmentation techniques in gurumukhi script" International Journal of Computer Applications (0975 - 8887), Volume 1 – No. 24, 2010.

[6] 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, pp24-29 June 2010.

[7] Vikas J Dongre, Vijay H Mankar," A Review of Research on Devnagari Character Recognition" International Journal of Computer Applications (0975 – 8887), Volume 12– No.2,pp8-15 November 2010

[8] Naresh Kumar Garg, Lakhwinder Kaur & M.K. Jindal ,"The Hazards in Segmentation of Handwritten Hindi Text" International Journal of Computer Applications (0975 – 8887) ,Volume 29– No.2, September 2011

[9] Ashwin S Ramteke, Milind E Rane,"Offline Handwritten Devanagari Script Segmentation"international Journal Of Scientific & Technology Research Volume 1, Issue 4,pp142-145, MAY 2012

[10] Gazal Munjal, Ms. Neha Sahu," Study of techniques used for Devanagri Handwritten Character Recognition" International Journal of Research in Engineering and Sciences(IJRES), Vol 1, Issue 2, pp.34-40, 2013.

[11] Simpel rani, Arbha Goyal ,"An efficient approach for segmentation of touching characters in handwritten hindi word"International conference oon Information and mathematical Sceinces 2013, 2014 ELESVIER.

[12] Munish kumar , Mk jindal , R.K.Sharma, "segmentation of Isolated And Touching Characters in Offline Handwritten Gurumukhi Script Recognition" I.J. Information Technology and Computer Science, 2014, 02,58-63.