A Review on Segmentation of Touching and Broken Characters for Handwritten Gurmukhi Script

Karamjeet Kaur
M.Tech Research Scholar,
Computer Science Section,
Yadavindra College of Engineering,
Talwandi Sabo, Bathinda, Punjab.

Ashok Kumar Bathla
Assistant Professor,
Computer Engineering Dept.,
Yadavindra College of Engineering,
Talwandi Sabo, Bathinda, Punjab.

ABSTRACT
Character Segmentation of Handwritten Documents has been an interested area of research and its applicable environment becomes it a challenging research topic. The desire to edit the scanned document leads to develop the idea of optical character recognition. Segmentation plays very important role in optical character recognition system. The incorrect segmentation is just like a garbage in and garbage out. Segmentation of the broken character is quite difficult because vertical profile projection technique assumes the broken parts of the characters as individual characters. Existing methods focuses only upon the single touching characters. But our main focus is to design a robust method for the segmentation of broken and multiple touching characters. Existing systems focus only on the segmentation of fixed sized characters. But we develop the size independent algorithm which works on variable size characters. Thus, in the proposed method we develop the algorithm which works on the segmentation of broken, multiple touching characters of independent size including the three zones of the handwritten gurmukhi script. The main challenges like as variation in handwriting style etc. make the segmentation to difficult.

General Terms
Optical Character Recognition

Keywords
OCR, Pre-Processing, Segmentation, Recognition, Gurmukhi Script, Broken Characters, Touching characters.

1. INTRODUCTION
Optical character recognition (OCR) is the process which convert the scanned images of machine printed or handwritten text into a computer processable format. If you scan a text document, then the concept of optical character recognition (OCR) software is used to translate image into text that you can edit. Firstly scanner creates an image from page; image is stored as a bitmap in computer's memory. A bitmap is a grid of dots; one or more bits represent each dot. The motive of OCR software is to translate that array of dots into text that computer can interpret as letters and numbers. Steps in OCR i.e. Pre-processing, Segmentation, Feature extraction, Classification, Recognition are shown in the figure 1. 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.

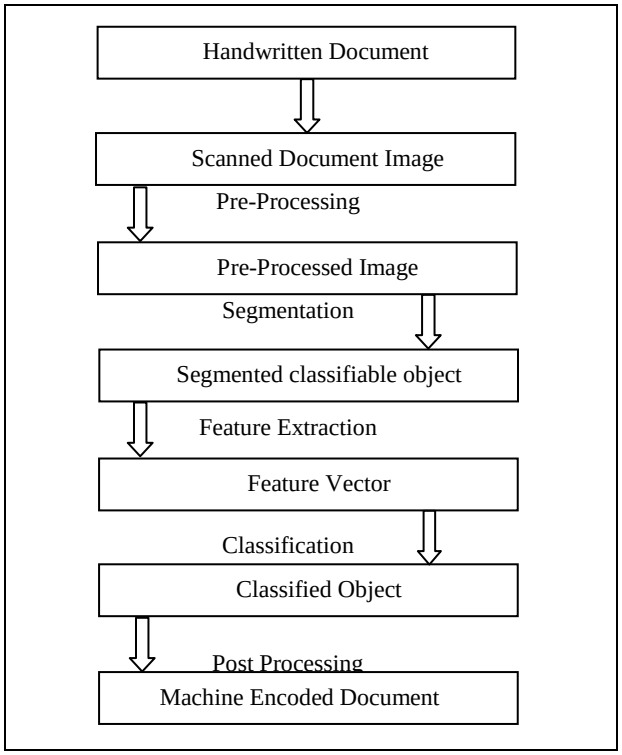

Fig 1: Optical Character Recognition Architecture

1.1 Characterstics of Gurmukhi Script
Gurmukhi is derived from the combination of two words "Guru" and "Mukh". Gurmukhi means to record the sayings from the mukh of the Gurus, i.e. from the Guru's mukh. Guru Angad Dev Ji originates this srcipt. Gurmukhi script alphabet consists of 41 consonants, 12 vowels and 3 half characters. Gurmukhi script is written in left-to-right. The zones in gurmukhi script are shown in figure 2 given below:

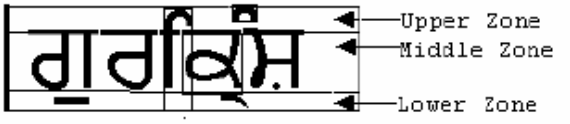

Fig 2: Three Zones of Gurmukhi Script Word

1.2 Character Segmentation

The intensive research effort on the field of character segmentation was not only because of its challenge on simulation of human being reading, but also, because it provides efficient applications such as the automatic processing of huge amount of papers, converting data into machines and web interface to paper documents . A character segmentation system can be either "online" or "offline."

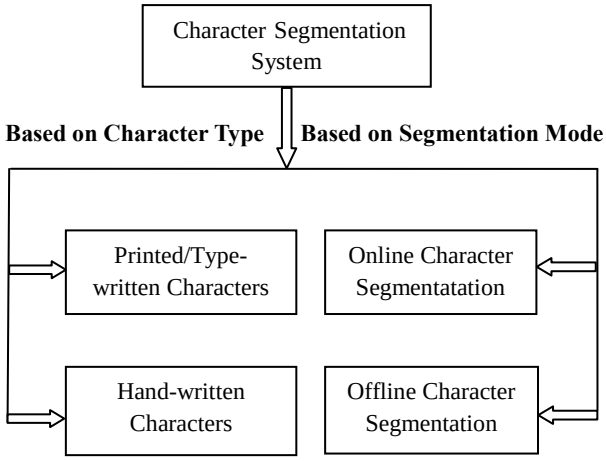

Fig 3: Character Segmentation System

Online character segmentation is the process of segmenting handwriting, used a digitizer, as a time sequence of pen coordinates. It captures the dynamic information of the pen trajectory.

Offline character segmentation is the process of converting the image of writing into bit pattern by an optically digitizing device such as optical scanner. The segmentation is carried out on this bit pattern data for machine-printed or handwritten text. Offline character segmentation is a very important tool for creation of the electronic libraries.

1.3 Segmentation Systems

Segmentation is carried out mainly on machine printed text and the handwritten text:

Machine printed text includes the machine printed material materials such as newspapers, magazines, documents, books and various writing units in the video and image. The height, width and pitch are uniform of the machine printed characters assuming the same font and size are used. Various problems related to these are solved with little constraint.

Handwritten text can be further divided into two categories: Cursive and Hand printed script. The size of the handwritten characters is not uniform. Because of non uniformity the segmentration handwritten characters is a much more difficult. Characters can vary greatly in size and style.

2. RELATED WORK

Bansal et al. (2010) [1]: This paper elaborates the segmentation of various irregular text words written in Gurumukhi script. This paper deals with the segmentation of words containing skewed, broken, irregular head line, touching and overlapped characters. Some of the new techniques like counter tracing methods are used along with horizontal and vertical projections.

Kumar et al. (2014) [3]: This paper presents the segmentation of handwritten Gurumukhi characters is carried out defining the whole process for segmentation including digitization process and pre-processed techniques. Water Reservoir method is applied for identification and segmentation of touching characters.

Kumar et al. (2010) [4]: This paper proposed a technique in which the segmentation of the scanned document image is done. In which the whole image is consider as a one large window. The large window is split into less large windows as giving lines and once the lines are recognized then each window consisting of a line is used to recognize a word that is present in a line and at the end character is recognized. This paper uses the concept of variable sized window.

Mangla et al. (2014) [7]: This paper proposed the method of segmentation for touching and broken characters of handwritten Punjabi text that is the Gurumukhi script. This paper provides the new segmentation technique based on neighboring pixels for broken characters and increase the accuracy for touching characters. In the proposed system a new technique is developed to segment the touching characters which named as End detection algorithm.

Sharma et al. (2006) [9]: This paper proposed technique that segments the words by using header line, aspect ratio and vertical and horizontal projection profiles. The overall successful segmentation achieved through the procedure is 96.22%.

Thakral et al. (2014) [10]: This paper shows a new strategy for the segmentation of conjuncts, and the characters that are overlapped characters in Devanagari script. The proposed algorithm aaplied Cluster Detection technique and gives 95% correctness for segmenting touching, conjunct characters and 88% effectiveness for the characters that overlapped. The given technique segments the middle region of the word accurately; this can be further extended on upper and lower modifiers.

Mehta et al. (2014) [8]: This paper develops a hybrid classification scheme is employed based on Horizontal Profile Projection and Vertical Profile Projection techniques and neighbouring pixels method was used to segment the broken characters. This paper discussed only segmentation of the skewed, simple and broken characters in Gurmukhi script.

Naunita et al. (2011) [11]: This paper applied straight segmentation method for segmentation of touching characters in handwritten Gurumukhi words. This method does not work properly for the characters that are broken and overlapped, but gives the better results for touching characters. The hybrid approach i.e. a combination of horizontal and vertical projection profiles techniques together have been applied on all the Gurumukhi script documents for obtaining the results.

3. EXISTING TECHNIQUES

The various techniques will mainly used to segment the words into characters are given below:

Horizontal Projection Profile (HPP): For a binary image of size H*W where H denotes the height of the image and W denotes the width of the image. The horizontal projection is represented as HP (j), j=1, 2 ... H. This operation counts the total number of black pixels in each horizontal row. In this

study, the header line is detected from the input text image by using HPP technique and converts it into white empty pixels.

Vertical Projection Profile (VPP): For a binary image of size H*W where H denotes the height and W denotes the width of the image, the vertical projection is being defined as VP (k), k=1, 2 ... W. This method counts the total number of black pixels in each vertical column. Sometimes multiple header lines due to some irregularities detected 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 at that point they touch. Cluster represents the heap of pixels that increases the expected value of pixels. The assumption in this method is that a single character consists of maximum number of 20-25 pixels. When this value increases the expected number it is assumed that a single character is being touched or overlapped by another character. These characters can be conjuncts, touching and overlapping with one another and after that segmentation will be carried out [15].

Water Reserviour Method: Water reserviour method is used to solve the touching characters problem. If poured water from top and bottom of the character, the cavity regions of the characters are known as reservoirs. This method cannot use for the broken or overlapping characters [4].

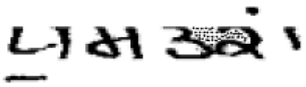

Fig. 4: A reservoir obtained from water flow from top is marked by dots [4].

End Detection Algorithm: End detection algorithm is used 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 [9].

4. CHARACTER SEGMENTATION PROBLEMS

There are various problems that can occur in character segmentation because all characters are of varied size & shapes in handwritten document. Problems are:

A. Problem of broken characters
B. Problem of touching characters
C. Problem of overlapped characters
D. Problem of Skewed characters

A. Problem of Broken Character: Broken character problem may arise due to improper writing of element e.g. some times while writing, the pen stops working properly in between the words or words do not scanned properly.

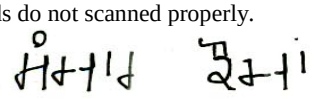

Fig. 5: Broken Characters

C. Problem of Touching Character: This problem also arises due to different writing styles. While writing, if one character touches other character then it will becomes difficult to recognize.

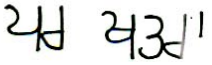

Fig. 6: Touching Characters

B. Problem of Overlapped Character: This problem arises due to different writing styles of different people. In this problem one character is written above on the other characters by mistake. As such, vertical projection of these characters will also be overlapping with each other. The problem of overlapping character is shown in Fig 5.

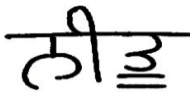

Fig. 7: Overlapping Character

D. Problem of Skewed Character: In handwritten text, the characters are skewed to left or right. Because of this skewness, the projections of two or more characters or symbols overlap with each other. These types of skewness of word are shown in Figure 6.

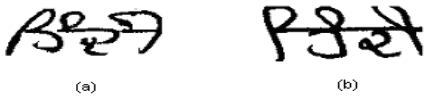

Fig. 8: Word Skewness (a) Left skewed and (b) Right skewed

5. CONCLUSION

In this paper, various methods for character segmentation like horizontal projection profile, vertical projection profile, pixel cluster identification technique, end detection algorithm and water reserviour method have been briefly discussed. Thus process of character segmentation has to face many problems like irregular size of characters, multiple touching characters, broken characters etc. It is concluded that segmentation is the necessary step for proper recognition.

Table 1. Comparative study of existing work on different characters:

Ref.	Technique used	Type of input	Accuracy
Dharam Veer et. al [9]	Horizontal and Vertical Projection Profile	Simple Gurumukhi text	96.22%
Parika et. al [7]	End Detection Algorithm	Isolated, Broken and Touching characters in Gurmukhi	95%
Munish Kuma et. al [4]	Water Reservoir Principle	Isolated and Touching characters in Gurmukhi	93.5%
Binny et. al [10]	Cluster Detection Method	Hindi touching, overlapping & conjunct characters	94.5%
Bharti et. al [8]	Horizontal and Vertical Projection Profile	Broken Characters in Gurmukhi Script	93%

6. FUTRE WORK

Existing techniques cannot solve these entire problems such as characters touching more than one character, broken characters and the character segmentation for the independent size. Hence in a future, there required a system i that can solve all these problems.

7. REFERENCES

[1] Bansal G., Sharma D., "Isolated Handwritten Words Segmentation Techniques in Gurmukhi Script", International Journal of Computer Applications, Vol. 1, No. 24, pp. 104-111, 2010.

[2] Garg N.K., Kaur L., Jindal M.K. "The segmentation of half characters in Handwritten Hindi Text", Springer-Verlag Berlin Heidelberg, pp. 48-53, 2011.

[3] Kumar D., Koshti, Govilkar S., "Segmentation of Touching Characters in Handwritten Devanagri Script", International Journal of Computer Science and its Applications, Vol. 2, Issue 2, pp. 83-87.

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

[5] Kumar R., Singh A., "Algorithm to Detect and Segment Gurmukhi Handwritten Text into Lines, Words and Characters", IACSIT International Journal of Engineering and Technology, Vol.3, No.4, 2011.

[6] Kumar R., Singh A., "Detection and Segmentation of Lines and Words in Gurmukhi Handwritten Text" Institute of Electrical and Electronics Engineers (IEEE), pp. 353-356, 2010.

[7] Kumar R., Singh A., "Challenges in Segmentation of Text in Handwritten Gurmukhi Script" Proceedings in BAIP 2010, CCIS 70, Springer-Verlag Berlin Heidelberg, pp. 388-392, 2010

[8] Lehal G.S., Singh C "A Complete OCR System for Gurmukhi Script" Springer-Verlag Berlin Heidelberg, pp. 358-367, 2002.

[9] Mangla P., Kaur H., "An End Detection Algorithm for segmentation of Broken and Touching characters in Handwritten Gurumukhi Word", Institute of Electrical and Electronics Engineers (IEEE) , pp.1-4, 2014.

[10] Mehta B., Rani S., "Segmentation of Broken Characters of handwritten Gurmukhi Script", International Journal of Engineering Sciences, Vol. 3, pp. 95-105, 2014.

[11] Naunita, Taneja A., Chawla M., "Segmentation of Touching Characters in Handwritten Gurumukhi Script", International Journal of Engineering Sciences, Vol. 3, pp. 90-94, 2014.

[12] Rani S., Goyal A., "An efficient approach for segmentation of touching characters in handwritten hindi word", International conference of on Information and mathematical Sceinces, ELESVIER, 2014.

[13] Richard G. Casey, Lecolinet E., "A Survey of Methods and Strategies in Character Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, No. 7, July 1996.

[14] Sharma D., Lehal G.S., "An Iterative Algorithm for Segmentation of Isolated Handwritten Words in Gurumukhi Script", 18th International Conference on Pattern Recognition (ICPR'06), IEEE, 2006.

[15] Thakral B., Kumar M., "Devanagari Handwritten Text Segmentation for Overlapping and Conjunct Characters-A Proficient Technique", Institute of Electrical and Electronics Engineers (IEEE), pp.1-4, 2014.