Problems of Character Segmentation in Handwritten Text Documents in Gurumukhi Script

Er.Naunita*
M.Tech Student, Department of Computer Engineering
Yadwindra College of Engineering, Talwandi Sabo
Bathinda, Punjab

Er. Rajbhupinder Kaur**
Assistant Professor, Department of Computer Engineering
Yadwindra College of Engineering, Talwandi Sabo
Bathinda, Punjab

Abstract— **Optical Character Recognition (OCR) is a process to recognize the handwritten or printed scanned text with the help of a computer. Character segmentation is a process of dividing a word into character. In this paper, prime focus is on the problems which may occur during the character segmentation. The problems in segmentation can lead to decrease in segmentation rate. This paper is divided into 5 sections. Section 3 will be focus on the major problems that may occur during the character segmentation process.**

***Keywords*— OCR, Gurumukhi Script, Character segmentation**

I. INTRODUCTION

Character segmentation is very useful in many fields. After segmentation of the character, features can be extracted from the segmented character. Accuracy in extracting the features is highly depends upon the segmented character. If character whose features are to be extracted is not segmented properly, can't be recognized correctly by the feature extraction algorithm. Character segmentation of the handwritten text documents written in the Gurumukhi script is dependent on the writing style of individual. Segmentation of characters is easy in case of printed documents as compared to the handwritten documents.

II. CHARACTERISTICS OF GURMUKHI SCRIPT

Gurumukhi script alphabet consists of 41 consonants and 12 Vowels (Fig.2). Besides these, some characters in the form of half characters are present in the feet of characters. Writing style is from left to right. The concept of upper/lowercase characters is absent in Gurumukhi. A line of Gurumukhi script can be partitioned into three horizontal zones (Fig.1) namely, upper zone, middle zone and lower zone. The middle zone generally consists of the consonants. The horizontal line is present at upper part in Punjabi language called headline. Punjabi is written from left to write. A line of Gurumukhi script can be partitioned into three horizontal zones, namely, upper zone, middle zone and lower zone as shown following:

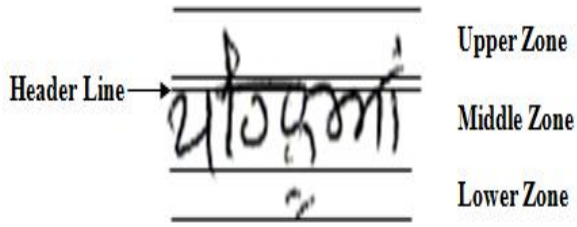

Fig1: Three zones of Gurumukhi word

The header line is the most visible and the header line we can obtain the middle zone part of a word. We have separated the header line we segment the upper modifiers and lower modifiers of a word. After then we can check the touching and overlapped character or not. If it is a touching character then we segment that word. The above proposed method suffered from various problems of broken characters, problem of

overlapping characters, problem of touching characters, problem of skewed characters, problem of irregular intensity with the character, problem of detecting the header line.

Fig2: Character Set of Gurumukhi Script

III. LITERATURE REVIEW

A good research about problems of segmentation according to three different zones is given in [1]. The algorithms on segmentation of touching characters and overlapping line in printed Gurumukhi script are referenced as [2]. In this paper the new methods and strategies are surveyed for character segmentation [3]. The main objective of this paper is to find the different character segmentation problems which may occur during handwritten Gurumukhi script [4]. Author shows a proposed algorithm to detect and segment Gurumukhi text into line, word and character. They apply this algorithm and show the different-different results [5]. This paper deals with a text line detection and segmentation of Gurumukhi handwritten text document and removal of problems occur in line segmentation [6]. This paper deals with segmentation of touching characters in handwritten devnagri script; it shows an appropriate algorithm for this problem. [7]. this paper presents an approach to text line extraction in handwritten document images which combines local and global techniques. They propose a graph-based technique to detect touching and proximity errors that are common with handwritten text lines. [8].

IV. CHARACTER SEGMENTATION PROBLEMS

There are various problems that can occur in character segmentation because all characters are of varied size & shapes in handwritten document. The problems in character segmentation can be divided into a variety of categories:

A. Problem of broken characters
B. Problem of overlapped characters
C. Problem of Touching 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. There are two examples showing broken character. This leads to the formation of broken character Image is as shown below:-

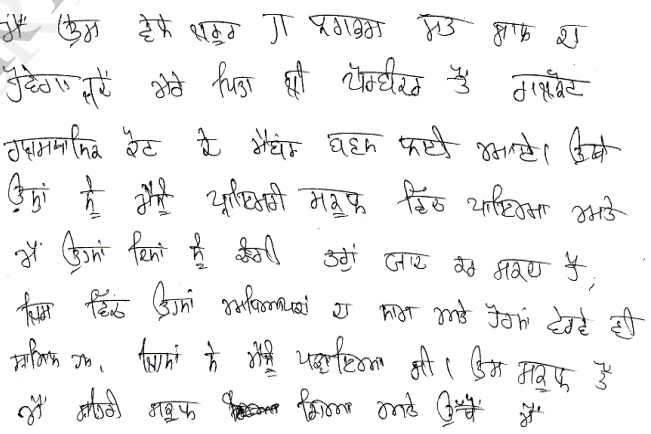

Fig 3: part of database

Fig4: Broken character

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. This is called as overlapping of character.

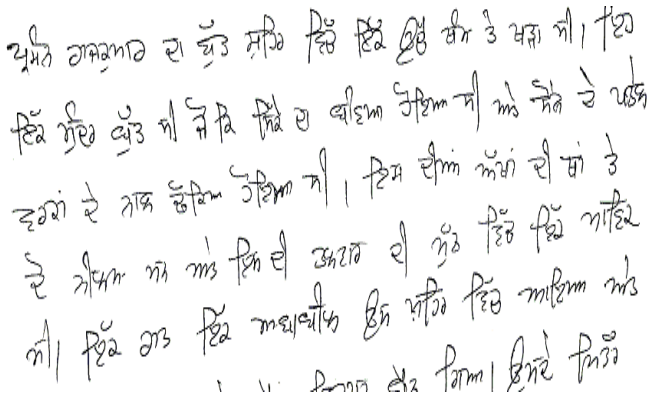

Fig 5: part of database

Fig 6: Overlapped character

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. The following example shows this problem:

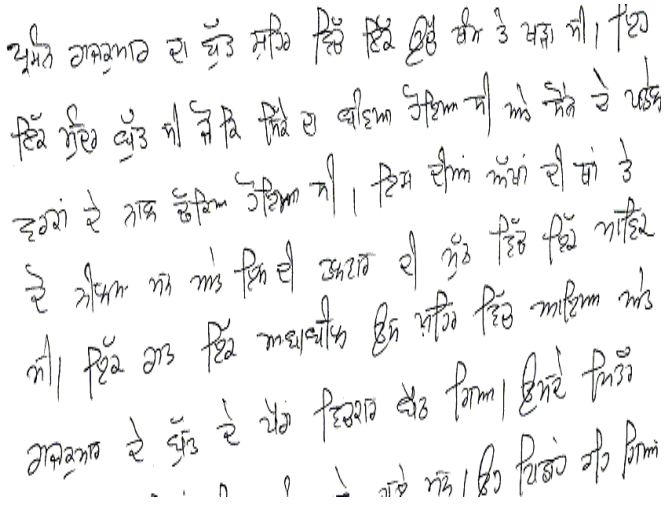

Fig 7: Part of database

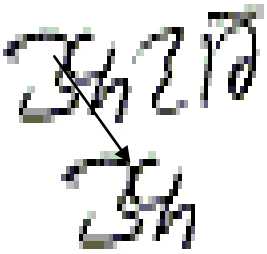

Fig 8: Touching character

D. PROBLEM OF SKEWED CHARACTER

In this problem, words in a line are not written straight but the word inclined either left-skewed or right-skewed which causes difficulty during segmentation.

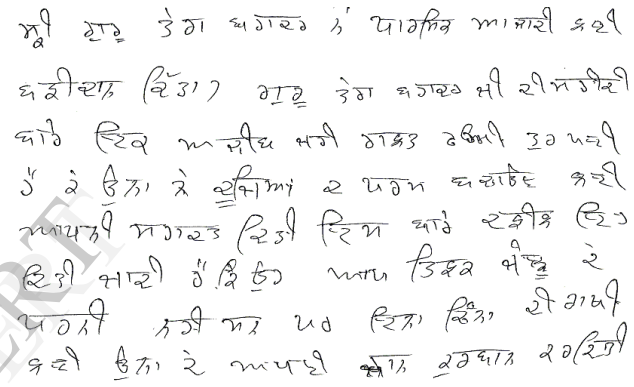

Fig 9: Part of database (In this database characters are both left and right skewed)

V. DISCUSSION

The problems explained above are very useful for complete segmentation of handwritten Gurumukhi text. Some problems can be removed if writer uses the better material and write patiently. To solve the problems related with writer's natural handwriting efficient algorithms are to be designed to segment the text.

VI. REFERENCE

[1] Garg Naresh kumar, kaur Lakwinder and jindal M.K. 2011. The hazards in segmentation of handwritten Hindi text. In international journal of computer applications (0975-8887) volume 29-No.2.

[2] Manish kumar jindal, Gurpreet singh lehal, Rajendra kumar Sharma 2009. 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.

[3] *Richard G. Casey, and Eric Lecolinet.* a survey of methods and strategies in character segmentation.

[4] K. Sharma Rajiv, S. Dhiman Amardeep Challenges in Segmentation of Text in Handwritten Gurmukhi Script. Information Processing and Management. Communications in Computer and Information Science Volume 70, 2010, pp 388-392

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

[6] Namisha Modi, Khushneet Jindal, Text Line detection and Segmentation in Handwritten Gurumukhi Scripts. International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 5, May 2013.

[7] Mr. Dipak V. Koshti,Mrs. Sharvari Govilkar. Segmentation of Touching Characters in Handwritten Devanagari Script. UACEE International Journal of Computer Science and its Applications Volume 2: Issue 2

[8] Jayant Kumar, Le Kang, David Doermann, Wael Abd-Almageed Segmentation of Handwritten Textlines in Presence of Touching Components.