A Survey of Problems of Overlapped Handwritten Characters in Recognition process for Gurmukhi Script

Arwinder Kaur[1], Ashok Kumar Bathla[2]
[1]M. Tech. Student, CE Dept.,
[2]Assistant Professor, CE Dept.,
Yadavindra College of Engineering, Talwandi Sabo, Punjab, India.

Abstract- **Gurmukhi is used primarily for writing Punjabi language. In India Gurmukhi script is mostly used for documentation purpose. Gurmukhi character recognition has been widely studied in the last few years. Much research work is present related to the printed script but very little work has been done in handwritten Gurmukhi character recognition. Gurmukhi character set has the elements which have complicated structures. People use so many different writing styles in handwritten documentation. Character recognition of handwritten text In Gurmukhi is laborious process mainly because of the structural features of the script and the various writing styles. When we try to segment the characters from word difficulty arises because of overlapped, connected and fused characters. They need to be focused. In the segmentation process, we cannot overlook the problem of overlapped characters as it is the very serious one. This paper discusses problems that arise at different stages of OCR process because of overlapped characters in Gurmukhi script.**

Index Terms- **Gurmukhi script; character segmentation; overlapping characters; touching characters; horizontally overlapping lines.**

I INTRODUCTION

As an aspect of Optical Character Recognition, handwritten character recognition is much difficult task in comparison to machine printed character recognition. Handwritten character recognition is a very useful area as it has applications in offices, banks, libraries, postal services, and many other fields. In a country like India so many ancient documents are presented in languages like Sanskrit, Punjabi, etc., which are based on the Gurmukhi character set. The study of these documents will be eased by the use of OCR technology [1]. Optical character recognition in the area of Gurmukhi should be given special attention so that efficient retrieval and analysis of ancient and modern Gurmukhi literature can be done effectively. Mainly three methods in use for segmentation are

Classical approach

Recognition based approach Holistic approach

They work well for the simple printed text. But there are many irregularities present in handwritten Gurmukhi scripts. Some common irregularities in writing the Gurmukhi script have been reported many times. The various problems likely to be faced in the machine recognition process pertaining to the words of Gurmukhi script are also detailed [2]. The techniques used for printed text cannot be always used for handwritten text due to variation in writing styles or pen types.

The problem of segmenting the so-called overlapping characters to as correct as possible is discussed many times. Stroke level evaluation studied in “On-line Chinese Character Recognition System for Overlapping Samples”. Overlapping samples consist of consecutive strokes. Logically, each “gap” between two strokes is a segmentation position candidate (SPC). The concept of imaginary stroke is defined as those virtual pen moving trajectories between two consecutive strokes. It’s a straight line from the end point of a stroke to the start point of the next stroke [3].

Further the paper is divided into following sections: section II discusses the OCR Process steps in detail, section III discusses challenging issues in OCR due to overlapped characters in Gurmukhi, section IV discusses related research, and section V, the end section concludes the paper.

II OCR STEPS

A. Image acquisition process

An image after scanning is acquired as an input image by the recognition system. Generally the images used are in black and white form in any format such as JPEG, BMT, BMP, etc. This image is acquired is then used for further processing.

B. Pre-processing

The image thus obtained from above process may contain a certain amount of noise. The preprocessing has the sub-processes as noise removal, binarization of image, detection of edges, dilation and filling. Depending on the resolution of the scanner and the success of the applied technique in threshold process, the characters may be smeared or broken. Some of these defects may later cause poor recognition rates [4]. So we can use a pre-processor to eliminate those defects and to smooth the digitized characters. The smoothing implies both filling and thinning of data under consideration. Filling eliminates tinny breaks and unwanted gaps in the digitized characters, while the process of thinning reduces the width of the line, smoothening smoothes the characters. The most common technique used for smoothing is the window technique, a window moves across the binary image of the character, applying certain rules to the contents of the window. In addition to smoothing, pre-processing usually includes normalization. It is applied to obtain characters of uniform size, particular slant and particular rotation.

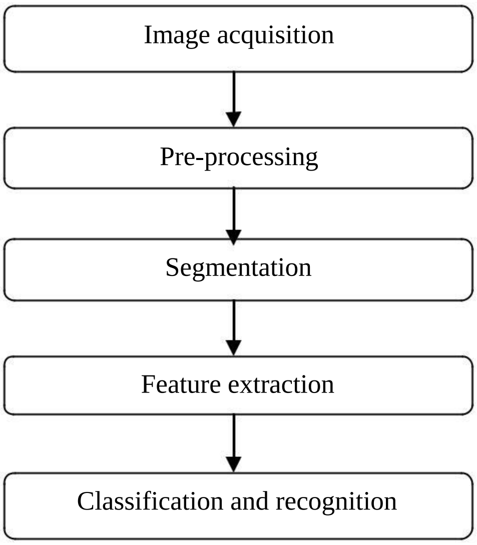

Fig. 1. Character Recognition System.

C. Segmentation

Segmentation of the text is mostly performed in
the following sequence:

1. Text is segmented into lines.
2. Lines are segmented into words.
3. Removal of header line.
4. Segmentation of characters from the word.

In Gurmukhi Optical character recognition, segmentation process is very tedious job because of its huge character set and large number of symbols available in Gurmukhi text. Error at this stage propagates to recognition stage and reduces efficiency there. Presence of overlapped and skewed characters causes difficulties in the process of segmentation. Presence of touching characters further elevates the problem in the process of segmentation.

D. Feature Extraction: change in our words full

In this stage, we define set of features which identify attributes of character. To identify the shapes, structural and statistical features are used. Zoning, moments, etc. come under statistical features. They have quite good tolerance but these are font dependent. On the other hand structural features consider shape attributes. This in turn depends on topological features and their inter-relationship. These include character holes, concavities in contour, end points etc. The feature set should make a clear distinction among the characters of different classes and at the same time it should be similar for the characters of the same class. The efficient functioning here improves the recognition rate and reduces the false classification which makes the further processing efficient.

E. Classification and Recognition

Character is identified and then assigned to its correct character class, in the classification process. The correct character class is selected based on the extracted features. The relationship between the characteristics may be of importance when deciding on class membership. For instance, if we know that a character consists of two vertical strokes, it may be a “ga” and the relationship between the two strokes is needed to distinguish the characters. A structural approach is then needed.

III CHALLENGES

OCR on the basis of the data acquisition process can be classified into two type viz. online data and off line data. Whether the acquired data are online or off-line, the problems always arise due to touching, merged or overlapped characters. In Figure 2, the two characters are overlapped characters. Two adjacent characters are in the region of each other if they do not touch each other and cannot be separated by a single vertical line due to the overlapping region. While segmenting such region overlapped characters, the problems occur. In "Segmentation of touching and fused Gurmukhi characters", the problem of overlapped characters region has been mentioned. The shadow characters may contain two characters of different width, in spite of having same font size and font style. Hence it causes the problem in their segmentation [7].

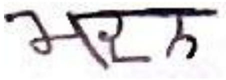

Fig. 2. Example of overlapped characters.

In Gurmukhi handwritten character recognition is a very difficult task due to improper handwriting people write the characters one over the other.

This causes the difficulties in segmentation and identification of each individual character from the overlapped one. The

Gurmukhi character set includes consonants, vowels, auxiliary signs, half characters etc. In segmentation of such characters due to lack of vertical white space between the two characters, the complexity increases and it requires more computation to segment out them accurately. Thus, it is very tedious job to segment and identify such overlapped characters in the Gurmukhi text. Most of the handwritten documents have the characters written very close to each other leaving no space between the two. Such an example is given in figure 3. In the segmentation process these are segmented as a single character and next in classification process it gets rejected as no such character is present in the database.

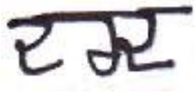

Fig. 3. Example of touching characters.

In Gurmukhi each word can be divided into three regions [9] viz. Upper region, middle region and lower region. Upper and lower regions hold the modifiers and the middle region holds the characters. Many times in writing the documents these regions get overlapped on each other as shown in figure. 4. This overlapping results in problems for character recognition process. Also the lower region of upper string gets inserted into the upper region of lower string and makes it difficult to segment them.

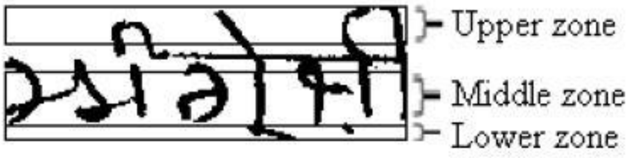

Fig 4. Three strips of a word in Gurmukhi script [6].

The multiple horizontally overlapping lines which are normally found in printed newspapers of almost every language, so is the case of Gurmukhi script. It is shown in the figure5 given below.

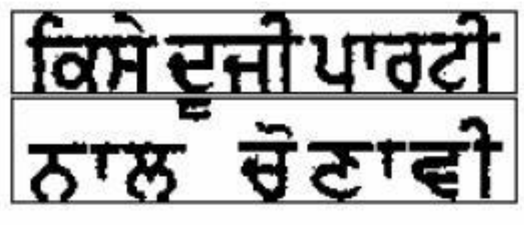

Fig. 5. Lower zone characters of one line touching with upper zone of next line.

Detecting overlapped handwritten characters between adjacent lines of text is presented using Neuro- Fuzzy recognition of overlapping handwritten text between adjacent lines of text using soft computing [5]. The presence of a horizontal line on the top of all characters forming a word is the most distinctive feature of Gurmukhi. This line is known as header line as shown in figure 5. The words can typically be divided into three strips: top, core, and bottom. The header line separates the top and core strips and a virtual base line separates the core and lower strips [6]. The top strip generally contains the top modifiers, and bottom strip contains lower modifiers.

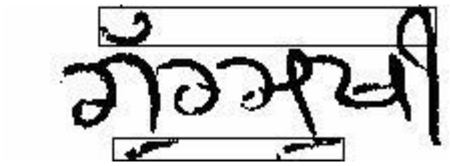

Fig. 6. Lower zone characters overlapped and connected with middle zone characters [7].

When two or more words appear side by side to form a big word in Gurmukhi, the header lines touch and generate a bigger header line. This is one of the most common overlapping issues. Two adjacent words appear as one single word. Also, sometimes due to styles of writing the characters gets overlapped on the header line. Figure 6 shows some middle strip characters get overlapped on header line due to special writing styles.

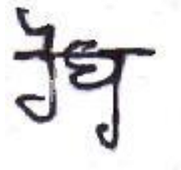

Fig. 7. Overlapping characters over header lines

IV PREVIOUS WORK AND RELATED RESEARCH

Sharma et al. [7] proposed a technique which segments the words in an iterative manner by focusing on presence of headline, aspect ratio of characters and vertical and horizontal projection profiles. . Handwritten text is prone to problem of half characters, overlapped, connected and merged characters within a word. The proposed approach of segmentation can be used for handwritten text of Indian language scripts like Devnagri, Bangla etc. having structural feature similar to Gurmukhi script. Bansal et al. [6] proposed the technique for Isolated Handwritten Word Segmentation in Gurumukhi Script**.** Segmentation of handwritten words is difficult primarily because of structural features of the script and varied writing styles. Handwritten words are prone to the problem of overlapped, connected, merged and broken characters. Based on certain properties of Gurmukhi script, different zones across the height of word are detected. Further, different categories of overlapping and touching characters in all the three zones of handwritten words in Gurmukhi script have been identified on the basis of structural properties of Gurmukhi script. A method for segmenting overlapping characters in middle zone has been discussed. Jindal et al. [8] discusses about the multiple horizontally overlapping lines which are normally found in printed newspapers of almost every language due to high compression methods used for printing of the newspapers. This paper, we have discusses a solution for segmenting horizontally overlapping lines. According to this Whole document is divided into strips and algorithm is applied for segmenting horizontally overlapping lines and associating small strips to their respective lines. The algorithm works well for the Gurmukhi script. It shows a notable improvement for recognizing printed text from Gurmukhi newspapers. Kumar et al. [9] proposed the method for the line, word, character and top character segmentation for printed Hindi text in Devanagari script. And also describe the line and word segmentation for printed text in Gurumukhi script. This paper, presents a modified algorithm for segmentation of line, word, character, top character for Devanagari Script and Segmentation of line, word for Gurmukhi Script. The overall successful segmentation achieved through the algorithm is better than previous result. At few point segmentation is good but at few point it is not up to the expectations. This may be because of the shape of characters. Mangla et al. [10] discusses the segmentation of touching and broken characters in handwritten Gurmukhi word. 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. This paper describes the method of segmentation for touching and broken characters of handwritten Punjabi text that is the Gurmukhi script. The segmentation technique described here is based on neighboring pixels for broken characters and increase the accuracy for touching characters. This technique is named as End detection algorithm. Thakral et al. [12] shows a new strategy for the segmentation of conjuncts, and overlapping characters in Devanagari script on Hindi language. The algorithm is focused around Cluster Detection technique and works for various input characters. Hindi text is complex task due to variations in handwritings, structural properties of language and presence of upper, lower modifiers. These challenges have been overcome here. Different issues have been solved and just a single algorithm will give anticipated results with different sort of inputs. The given technique segments the middle region of the word accurately. The table above represents the comparison among the various existing techniques till date. The accuracy is represented and the technique used is mentioned.

TABLE I COMPARISON OF EXISTING TECHNIQUES:

Ref.	**Technique used**	**Type of input**	**Accuracy**
Dharmveer et al.	Horizontal and vertical projection profile[7]	Simple Gurmukhi text	96.22%
Kumar et al.	Waterreservoir technique[9]	Isolated and touching Gurmukhi characters	93.5%
Kumar et al.	Variable size window[9]	Isolated characters	90%
Mangla et al.	Analysis of neighbouring pixels[10]	Touching and broken characters in Gurmukhi	95%
Binny et al.	Cluster detection technique[11]	Hinditouching, overlappingand conjuncts	94.5%

V CONCLUSION

Methods for treating the various problems related to the segmentation and classification of printed text in optical character recognition have developed remarkably in the last decade. A variety of techniques has also emerged for handwritten text, influenced by developments in related fields such as online and offline handwritten character recognition. Recognition of overlapped characters is a complex computational problem. Only a few works have been reported in the areas of Gurmukhi handwritten recognition. The existing systems are not very much capable of segmenting and recognizing overlapped handwritten Gurmukhi characters. In this paper, the problems occurred in OCR process of overlapped handwritten Gurmukhi scripts are discussed in detail. In future solutions to these problems add a great strength to OCR systems.

REFERENCES

[1] S. Mori, C. Y. Suen, and K. Yamamoto, “Historical review of OCR Research and development”, Proceedings of the IEEE, Vol. 80, No. 7, pp. 1029-1058, 1992.

[2] M. K Jindal, R. K. Sharma and G. S.Lehal., “A Study of DifferentKindsofDegradationin Printed Gurmukhi Script”, IEEE International Conference on Computing Theory and Applications, pp. 538-544, 2007.

[3] X.Wan,C.LiuandY. Zou, “On-line Chinese Character RecognitionSystemforOverlappingSamples”,2011 International Conference on Document Analysis and Recognition.

[4] K.Prasad,D.Nigam,A.LakhotiyaandD.Umre, “CharacterRecognitionUsingmatlab’sNeuralNetwork

Toolbox", International Journal of u- and e- Service, Science and Technology vol. 6, no. 1, 2013..

[5] R. Sheth, K. Patil, N. Thakur and K.T.Talele, "Neuro-Fuzzy Recognition of Overlapping Handwritten Text between Adjacent Lines of Text using Soft Computing".

[6] G. Bansal and D. Sharma, "Isolated Handwritten Words Segmentation Techniques in Gurmukhi Script", International Journal of Computer Applications, Vol. 1, No. 24, pp. 122-130, 2010.

[7] D. V Sharma and G. S, Lehal, "An Iterative Algorithm for Segmentation of Isolated Handwritten Words in Gurumukhi Script", IEEE International Conference ON Pattern Recognition, Vol. 2, pp. 1022-1025, 2006.

[8] M.K.Jindal.R.K.SharmaandG.S.Lehal, "Segmentation of Horizontally Overlapping Lines in Printed Gurmukhi Script", IEEE International Conference in Advanced Computing and Communications, pp. 226-229, 2006.

[9] M.Kumar.,M.K.Jindal,andR.K.Sharma, "Segmentation of Isolated and Touching Characters in Offline Handwritten Gurmukhi Script Recognition", International Journal of Information Technology and Computer Science, Vol. 6, No. 2, pp. 58-63, 2014.

[10] P. Mangla and H. Kaur., "An End Detection Algorithm for segmentation of Broken and Touching characters in Handwritten Gurumukhi Word", IEEE International Conference on Reliability, Infocom Technologies and Optimization, pp. 1-4. 2014.

[11] B. Thakral and M. Kumar, "Devanagari Handwritten Text Segmentation for Overlapping and Conjunct Characters: A Proficient Technique", IEEE International Conference on Reliability, Infocom Technologies and Optimization, pp. 1-4, 2014.

[12] R.CaseyandE.Lecolinet,"A Survey of Methods and Strategies in Character Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, No.7, pp 690-706. 1996.

[13] M. K. Jindal, "Degraded text recognition of Gurmukhi script", Doctoral dissertation, PhD Thesis, Thapar University, Patiala, India. 2008.

[14] V. Kumar and P.K. Sengar., "Segmentation of Printed Text in Devanagari Script and Gurmukhi Script", International Journal of Computer Applications, Vol. 3, No. 8, pp. 24-29, 2010.