Segmentation of Touching Characters in Handwritten Gurumukhi Script

Er.Naunita*
Assistant Professor
Department of Computer Science & Engineering
GZS-PTU Campus, Bathinda (Punjab)
naunitabansal@yahoo.com

Er.Amit Taneja**
Assistant Professor
Department of Computer Science & Engineering
GZS-PTU Campus, Bathinda (Punjab)
amit_t19@yahoo.com

Er.Meenu Chawla***
Assistant Professor
Department of Computer Science & Engineering
GZS-PTU Campus, Bathinda (Punjab)
meenuchawla011@gmail.com

Abstract— *Segmentation of a word into characters is one of the important challenges. This is even more challenging when we segment characters in an offline handwritten document. Touching characters make this problem more complex. In this paper, we have applied straight segmentation method for identification and segmentation of touching characters in handwritten Gurumukhi words. We could achieve 90.9% accuracy for character segmentation with this method. If the characters are neither broken nor overlapping, then this technique will produce better results.*

Keywords- *Character Segmentation, Projection analysis, Zone Segmentation.*

I. INTRODUCTION

Optical Character Recognition (OCR) and Document image analysis are two important topics in the field of pattern recognition. In a text document image, preliminary step is extraction of text lines from document. Then each text line is segmented into words, and then each word is segmented into individual character images. Finally, these character images are inputted to the feature extraction phase for deciding the relevant shape contained in the character. This process is shown in Fig I. Digitization is the process of converting the paper based handwritten document into electronic form. Digitization is the process whereby a document is scanned and an electronic representation of the original, in the form of a bitmap image, is produced. The process of digitization gives a digital image. This digital image is inputted to preprocessing phase. Skew detection/correction, skeletonization and noise reduction/removal are three important steps of preprocessing phase. Skewness exists in a digital image if the bitmapped image is tilted. The document that is to be scanned may contain different fonts. The process of skeletonization is used to have uniformity in the representation of these kind of fonts. In this process, the width of curves present in the representation is decreased and the width is reduced from many pixels to single pixel. In the noise removal process, the unwanted bit patterns that might occur in digitized image are removed. The preprocessing phase is followed by segmentation phase. Segmentation is an important phase in character recognition process. In this process the digital image is segmented into paragraphs, lines, words and characters. A digital image of Gurumukhi document can also be segmented into these parts. Once a digital image is segmented into characters, appropriate features are extracted from the image in

feature extraction phase. This is important to define and extract appropriate and efficient features from the digital image as these are very important for improving the performance of a recognition system. The features extracted in this phase are further used in classification phase. Classification phase is the decision making phase in which a class membership is assigned to each digital image. In this paper, a technique for segmentation of touching handwritten Gurumukhi characters has been presented.

There are number of problems in segmentation of handwritten documents, for example, existence of characters with different sizes and various styles of writing a document. This problem becomes much more difficult when characters are touching or overlapping. A good number of algorithms have been proposed in past for segmentation of characters. These are based on 1) classical approach 2) recognition based segmentation 3) holistic approach and 4) hybrid approach [1]. Classical approach consists of methods that divide the input image into sub images, which are then classified. The operation of decomposition of an image is called "dissection". In this approach, the segmentation of characters is based on structural features. Dissection here means cutting up of the image into meaningful components base on general features like approximate character size, white space etc. In general, the criterion for good segmentation is the agreement of general properties of characters such as height, width, separation from neighboring components, etc. In recognition based segmentation, a search is made for image components that match with the character classes in the alphabet.

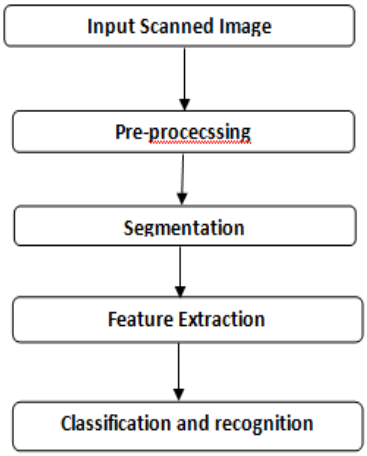

Fig.I

In the holistic approach, we want to recognize the word as a whole and thus avoid segmenting this into characters. Hybrid approach employs dissection together with recombining of rules to define segments. In this approach, there is a continuous space of segmentation strategies rather than a discrete set of classes with well-defined boundaries. Bansal and Sinha [2] have proposed a two pass algorithm for the segmentation and decomposition of Devanagari composite characters / symbols into their constituent symbols. Their algorithm uses structural properties of the script. In the first pass, words are segmented into easily separable characters. Statistical information about the height and width of each separated box is used to hypothesize whether a character box is composite. In the second pass, the hypothesized characters are further segmented. The algorithm is designed to segment a pair of touching characters. Jindal et al. [3] proposed a complete solution for segmenting touching characters in all the zones of printed Gurumukhi script. A study of touching Gurumukhi characters is carried out and these characters have been divided into various categories after a careful analysis. Structural properties of Gurumukhi characters are used for defining the categories. New algorithms have been proposed segment the touching characters in all three zones. These algorithms have been shown a reasonable improvement in segmenting the touching characters in degraded printed Gurumukhi document. Bansal and Sinha [4] have given a method for segmentation of the printed text in Devanagri. Their approach is a hybrid approach, where in they try to recognize the parts of the conjunct that form part of a character class. They use a set of filters that are robust and two distance based classifiers. They have presented a two level portioning scheme and search algorithm for the correction of optically read Devanagri characters of text recognition system. Chaudhuri and Garain [5] have given a technique based on fuzzy mul-factorial analysis. A predictive algorithm is developed for effectively selecting cut-points to segment touching characters. Pal and Datta [6] have used a water reservoir principle for Bangla handwritten text segmentation. Ikeda et al. Jindal et al. [9] have discussed an algorithm for segmentation of touching characters in upper zone of Gurmukhi script. Saba et al. [10] have provided survey on methods for touching character segmentation. They divide the touching character segmentation techniques

into two classes that perform explicit or implicit character segmentation.

This paper is divided as follows: Section II describe the zones in Gurumukhi script and Section III shows different types of character that may exist in a Gurumukhi script documents. Section IV contains the features of data collected and section V describes the hybrid approach used in this work for segmentation process and describes the problems of character segmentation in Gurumukhi script and experimental results and discussions are showed in section VI.

II. ZONE SEGMENTATION OF GURMUKHI SCRIPT

A line of Gurumukhi script can be partitioned into three horizontal zones (Fig.1I) upper zone, middle zone and lower zone. The middle zone generally consists of the consonants. The upper zone represents the region above the headline, while the middle zone represents the area just below the headline and above the lower zone. The lower zone is the lowest part which contains some vowels and half vowels appear in the higher, middle or lower zone only. The horizontal line is present at upper part in Punjabi language called headline. Punjabi is written from left to write. We needs to perform the following tasks:

(i) To find the header line.
(ii) To find the base line.
(iii) To define the upper zone.
(iv) To define the lower zone.
(v) To define the middle zone.

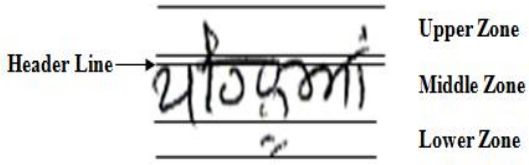

Fig.II

III. DIFFERENT TYPES OF CHARACTERS

3.1 Touching Characters

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 fig. shows this problem:

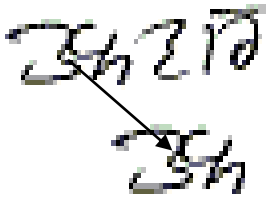
Fig III: Touching character

3.2 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 IV: Broken characters

3.3 Overlapping Characters

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. Image is shown below

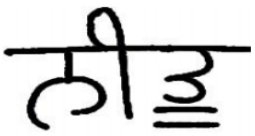
Fig V: Overlapped Character

IV. DATA COLLECTION

In this study, we have collected 20 handwritten Gurumukhi script documents. We have collected same handwritten Gurumukhi script document different writers All these documents are scanned at 300 dpi resolution. A good number of these documents contained touching characters. As such, a sufficiently large database has been collected for handwritten Gurumukhi script documents.

VI. CHARACTER SEGMENTATION IN GURUMUKHI SCRIPT

Segmentation of handwritten Gurumukhi characters is a challenging task because of structural properties of the Gurumukhi script and varied writing styles. We have tested the performance of the OCR and different algorithms for characters segmentation in collected handwritten Gurumukhi script documents. In Gurumukhi script, characters contain a horizontal line at the upper of the middle zone, is called headline. The headline helps in the recognition of script line positions. Segmentation of individual characters in handwritten Gurumukhi script is a straightforward process.

In our system, we have used HYBRID APPROAC i.e a combination of two approaches which are as follows:

1. Vertical Profile Projection Technique
2. Horizontal Profile Projection Technique.

Horizontal Profile projection: 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.

With the help of Horizontal Profile projection technique Lines are extracted from a given paragraph for further use [7]

Vertical Profile Projection: For a binary image of size H*W where H is the height of the image and W is the width of the image, the vertical projection has been defined as VP (k), k=1, 2 ...W.

This operation counts the total number of black pixels in each vertical column.

With the help of Vertical profile projection technique words from a segmented line can be extracted and then character can be segmented from the extracted words The processed word is the outcome of segmentation process. The segmentation process extracts constituent symbols images from a Gurumukhi script word and performs the following tasks:

(i) To find the header line. This is accomplished by finding maximum number of black pixels in a row.
(ii) To remove the header line.
(iii) Detect the end of character using vertical profile projection technique and save that end point.

Fig.VI: Touched character

As shown in Figure VI, 1st and 2nd characters touched with each other in the given word. So, vertical projection profile of these characters is also touching with each other. Now, detect the end of character using vertical profile projection technique and save that end point for touching character. In this technique, we first identify touching characters in the given word. After this, the touching characters of the word are segmented.

Fig.VII: Segmented Touched Character

The horizontal and vertical projection profiles techniques together have been applied on all the Gurumukhi script documents, which have been collected in this study. The combined results of these techniques are given in next section.

VI. RESULTS AND DISCUSSION

In order to detect and segment characters in scanned handwritten Gurumukhi script documents, as mention in earlier section, we have used hybrid technique. These techniques have been applied on the documents. The results of segmentation accuracy are given in Table.

Accuracy of Segmentation

Total Number of document images	Total Number of Touching Characters	Total Characters correctly Segmented	% Accuracy
20	265	241	90.9%

VII. CONCLUSION

The practical importance of OCR applications, as well as the interesting nature of OCR problem, has led to great research interest and measurable advances in the field.

Without any particular approach, it becomes very cumbersome job to segment each character. But with the application of proposed hybrid approach, the efforts done for character segmentation reduce tremendously. Firstly we extract the line. After that we checked that the character is touched or not and finally we segment the character using these techniques.

REFERENCES

[1] Casey R G, Lecolinet E. A Survey of Methods and Strategies in Character Segmentation. IEEE Transactions on Pattern Analysis.

[2] Bansal V, Sinha R M K. Segmentation of Touching and Fused Devanagari Characters. Pattern Recognition. 2002, 35(4):875-893.

[3] Jindal M K, Lehal G S, Sharma R K. Segmentation Problems and Solutions in Printed Degraded Gurmukhi Script. International Journal of Signal Processing. 2005, 2(5):258-267.

[4] Bansal V, Sinha R M K. A Devanagari OCR and A Brief Overview of OCR Research for Indian Scripts. In Proceedings of STRANS01, IIT, Kanpur, India.

[5] Garain U, Chaudhuri B B. Segmentation of Touching Characters in Printed Devanagari and Bangla scripts using Fuzzy Mulfactorial Analysis. IEEE Transactions on Systems, Man and Cybernetics-A. 2002, 32: 449- 459.

[6] Pal U, Datta S. Segmentation of Bangla Unconstrained Handwritten Text. In the proceedings of 7th International Conference on Document.

[7] Garg N. K.; Kaur L; Jindal M.K, "Segmentation of handwritten Hindi Text", IJCA (0975-8887) Volume1-No: 4 2010, pp 19-23.

[8] Jindal M K, Lehal G S, Sharma R K, Segmentation of Touching Characters in Upper Zone in printed Gurmukhi Script. In Proceedings of 2nd Bangalore Annual Compute Conference on 2nd Bangalore Annual Compute Conference, (Bangalore, India, January 09 - 10, 2009). COMPUTE '09. ACM, New York, NY, 1-6. DOI= http://doi.acm.org/10.1145/1517303.1517313

[9] Saba T, Sulong G, Rehman A. A survey on methods and strategies on Touched character segmentation. International Journal of Research and Reviews in Computer Science. 2010, 1(2):103-114.

[10] Naunita; Kaur R., "Problems of Character Segmentation in Handwritten Text Document in Gurumukhi Script", IJERT, 2013, Vol.2 (8):2529-2532.

[11] Jindal M. K.;Lehal G. S.; Sharma R. K., "A Study of Touching Characters in Degraded Gurumukhi Text", PWASET, 2005,Volume 4 :121-124.