I.J. Information Technology and Computer Science, **2014, 02, 58-63**
Published Online January 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2014.02.08

Segmentation of Isolated and Touching Characters in Offline Handwritten Gurmukhi Script Recognition

Munish Kumar
Department of Computer Science, Panjab University Rural Centre, Kauni, Muktsar, Punjab, India
E-mail: munishcse@gmail.com

M. K. Jindal
Department of Computer Science & Applications, Panjab University Regional Centre, Muktsar, Punjab, India
E-mail: manishphd@rediffmail.com

R. K. Sharma
School of Mathematics & Computer Applications, Thapar University, Patiala, Punjab, India
E-mail: rksharma@thapar.edu

Abstract**—** Segmentation of a word into characters is one of the important challenges in optical character recognition. 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 water reservoir based technique for identification and segmentation of touching characters in handwritten Gurmukhi words. Touching characters are segmented based on reservoir base area points. We could achieve 93.51% accuracy for character segmentation with this method. If the characters are neither broken nor overlapping, then this technique shall produce even better results**.**

Index Terms**—** Character Segmentation, Water Reservoir Method, Zone Segmentation, Handwritten Text Recognition

I. Introduction

Document image analysis and Optical Character Recognition (OCR) 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 isolated 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 illustrated in Fig 1. 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 that are performed in preprocessing phase. Skewness exists in a digital image if the bitmapped image is titled. Skewness can be caused by different factors, including, errors in scanning. The document that is to be scanned may contain different fonts, including fonts with bold face. The process of skeletonization is used to have uniformity in the representation of these 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 pattern(s) 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 (akhars). A digital image of Gurmukhi document can also be segmented into these units. 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 critical 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 isolated and touching handwritten Gurmukhi characters has been presented.

Segmentation of handwritten text document into lines, words and characters is one of the most important and challenging tasks in a handwriting recognition system.

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. Segmentation of individual characters which constitute a written string, is a straightforward process when characters are well spaced. 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, pitch, white space etc. In general, the criterion for good segmentation is the agreement of general properties of characters such as height, width, separation from neighbouring components, disposition along a baseline, etc. In recognition based segmentation, a search is made for image components that match with the character classes in the alphabet. In this technique, recognition is done iteratively from left-to-right scan of words, while searching for a "satisfactory" recognition result. In this method, the criterion is recognition confidence that includes syntactic or semantic correctness of the overall result.

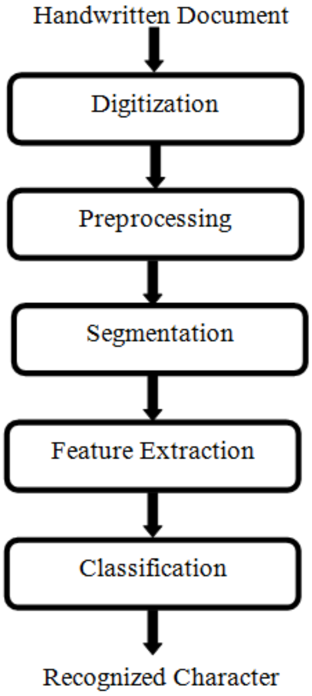


Fig. 1: Handwritten character recognition system

In the holistic approach, we attempt to recognize the word as a whole and thus avoid segmenting this into characters. Hybrid approach employs dissection together with recombination rules to define potential 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 extensively uses structural properties of the script. In the first pass, words are segmented into easily separable characters/composite 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 composite 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 three zones of printed Gurmukhi script. A study of touching Gurmukhi characters is carried out by them and these characters have been divided into various categories after a careful analysis. Structural properties of Gurmukhi characters are used for defining the categories. New algorithms have been proposed by them to segment the touching characters in middle zone, upper zone and lower zone. These algorithms have been shown a reasonable improvement in segmenting the touching characters in degraded printed Gurmukhi document. Bansal and Sinha [4] have given a complete 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 to classify the segmented images into known classes. They have presented a two level portioning scheme and search algorithm for the correction of optically read Devanagri characters of text recognition system for Devanagri script. 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.* [7] have proposed a method for recognizing Japanese handwritten characters including touching ones. The touching characters are segmented by cutting the connected components in the pre-segmentation process. Zhong [8] has given a novel method of circle scanning to detect intersection regions and to extract embedded/line touching character objects. 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. Reddy *et al.* [11] have proposed split profile algorithm for character segmentation in Telugu script. They analyze the statistical behavior of cursive components.

This paper is organized as follows: Section 2 describes the zones, in Gurmukhi script and section 3 depicts different types of character that may exist in a Gurmukhi script document. Section 4 contains the features of data collected in this work and section 5 describes the water reservoir method used in this work for segmentation process. Section 6 describes the problems of character segmentation in Gurmukhi script and experimental results and discussions are presented in section 7.

II. Zone Segmentation of Gurmukhi Script

A line of Gurmukhi script can be partitioned into three horizontal zones namely, upper zone, middle zone and lower zone. Consonants are generally present in the middle zone. 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. In the process of Gurmukhi script recognition, one 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.

III. Different Types of Characters

3.1 Isolated Characters

When two or more characters do not touch with each other, these are classified as isolated characters. Character segmentation is a straight forward process whenever characters are well spaced as shown in Fig. 2.

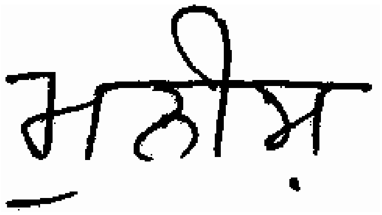


Fig. 2: Gurmukhi word with well-spaced characters

3.2 Touching Characters

In a handwritten Gurmukhi documents this is highly probable that characters touch each other and separation of such touching characters is a complex problem as shown in Fig. 3.

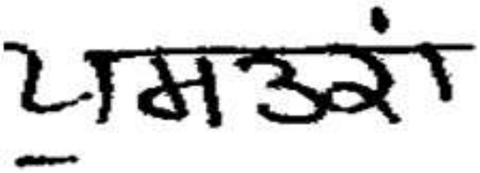


Fig. 3: Gurmukhi word with touching characters

As shown in Figure 3, third and fourth characters are touching with each other in the given word.

3.3 Overlapping Characters

In handwritten Gurmukhi documents, the characters can overlap with each other as shown in Fig. 4. As such, vertical projection of these characters will also be overlapping with each other.

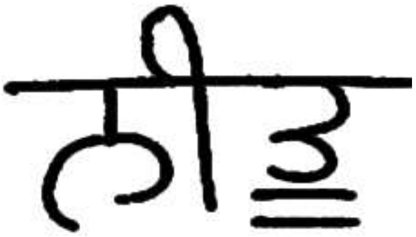
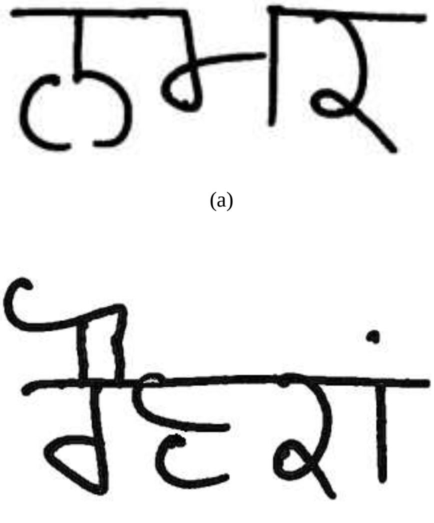


Fig. 4: Gurmukhi word with overlapping characters

3.4 Broken Characters

In handwritten Gurmukhi documents, some portion of characters in text may be missing as shown in Fig. 5. Fig. 5 (a) contains horizontally broken character and Fig. 5 (b) contains vertical broken character. It has been seen that most of the times, each broken character will have aspect ratio less than that of a single isolated character.



(a)

(b)

Fig. 5: (a) Horizontally broken characters (b) Vertically broken characters.

IV. Data Collection

In this study, we have collected 300 handwritten Gurmukhi script documents. These documents are of three categories. We have collected same handwritten Gurmukhi script document by 100 different writers and have put these documents in category 1. For category 2, we have again collected 100 documents written by a single writer. We have also collected 10 different documents by 10 different writers. These documents have been put in category 3. 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 Gurmukhi script documents. These 3 categories have further been analyzed and discussed in this paper.

V. Water Reservoir Method

The water reservoir method is as follows [12]. If water is poured from top and bottom of the character, the cavity regions of the characters where water will be stored are considered as reservoirs. For illustration see Fig. 6. Here by top reservoirs we mean the reservoirs obtained when water is poured from top. All reservoirs obtained in this way may not be considered for further processing. Those reservoirs whose heights are greater than a threshold are considered for future processing. The value of threshold is 1/10 of the character.

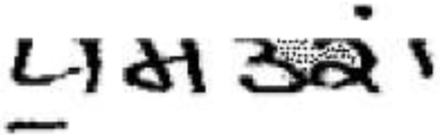
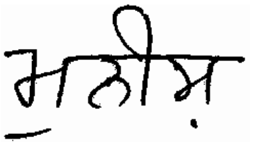


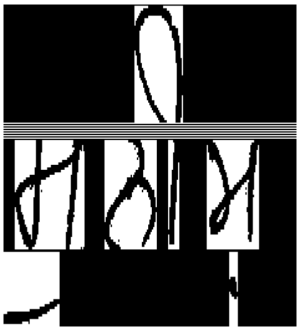
Fig. 6: A reservoir obtained from water flow from top is marked by dots

VI. Character Segmentation in Gurmukhi Script

Segmentation of handwritten Gurmukhi characters is a challenging task primarily because of structural properties of the Gurmukhi script and varied writing styles. We have tested the performance of the OCR and different algorithms for characters segmentation in collected handwritten Gurmukhi script documents. In Gurmukhi script, most of the 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 and character segmentation. Segmentation of individual characters in handwritten Gurmukhi script recognition, is a straightforward process when characters are well spaced as show in Fig. 7.



(a)



(b)

Fig. 7: (a) Gurmukhi word with well-spaced characters (b) Processed word

Definition 1.

(Horizontal projection): For a given binary image of size $M \times N$ where M is the height and N is the width of the image, the horizontal projection is defined by [2] as:

$$HP\ (i),\ i = 1, 2, 3, \dots, M \tag{1}$$

where *HP (i)* is the total number of black pixels in i^{th} horizontal row.

Definition 2.

(Vertical projection): For a given binary image of size $M \times N$ where M is the height and N is the width of the image, the vertical projection is defined as [2]:

$$VP\ (j), j = 1, 2, 3, \dots, N \tag{2}$$

where *VP (j)* is the total number of black pixels in j^{th} vertical column.

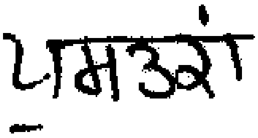
The processed word is the outcome of segmentation process. The segmentation process extracts constituent symbols images from a Gurmukhi script word and performs the following tasks:

(i) To find the header line. This is accomplished by finding maximum number of black pixels in a row.

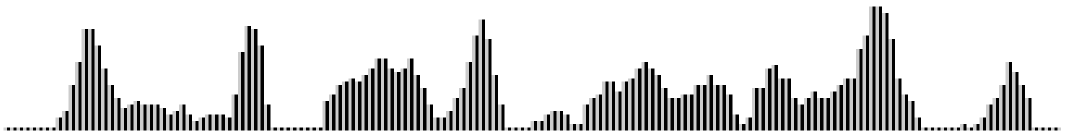
(ii) To remove the header line.

(iii) To extract sub images those are vertically separated from their neighbours. These sub images may contain more than one connected component.

In a handwritten Gurmukhi script documents this is highly probable that characters touch each other and separation of such touching characters is a complex problem as explained in Figure 8 (a) and (b). We have used statistical analysis based projection profiles technique for touching character segmentation.



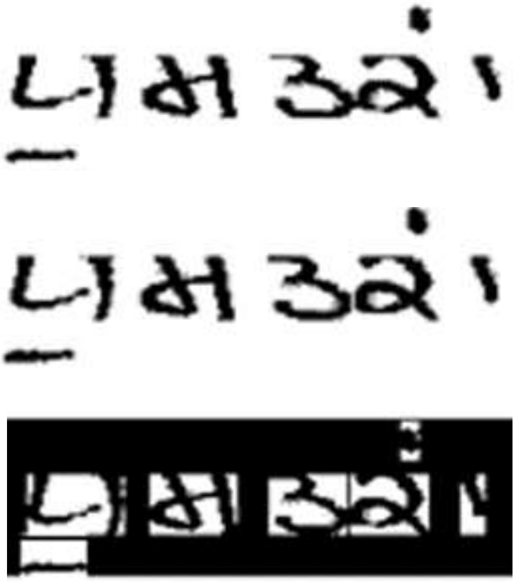
(a)



(b)

Fig. 8: (a) Touching characters (b) Vertical projection profile

As shown in Figure 8 (a), third and fourth characters are touching with each other in the given word. So, vertical projection profile of these characters is also touching with each other. We have used water reservoir based method for segmenting such touching characters. In this technique, we first identify isolated and touching characters in the given word. After this, the touching characters of the word are segmented based on reservoir base area as illustrated in Figure 9 (a), (b) and (c).



(a)

(b)

(c)

Fig. 9: (a) Gurmukhi handwritten word without headline (b) Touching character segmentation (c) Complete segmentation of word

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

VII. Results and Discussion

In order to detect and segment characters in scanned word of handwritten Gurmukhi script documents, as mention in section 4, we have used vertical projection profile technique and water reservoir base area point's technique. These techniques have been applied on the documents of three different categories. The category wise results of segmentation accuracy are given in Table 1-3.

Table 1: Same document written by different writers

Document	Accuracy
Doc 1	93.91%
Doc 2	92.24%
Doc 3	91.43%
Doc 4	92.62%
Doc 5	94.16%
Doc 6	97.46%
Doc 7	95.42%
Doc 8	92.68%
Doc 9	94.42%
Doc 10	96.12%

Average Accuracy = 92.53 %

Table 2: Different documents written by same writer

Document	Accuracy
Doc 1	88.9%
Doc 2	91.31%
Doc 3	90.23%
Doc 4	90.42%
Doc 5	93.26%
Doc 6	97.87%
Doc 7	91.41%
Doc 8	92.32%
Doc 9	91.96%
Doc 10	97.62%

Average Accuracy = 93.97 %

Table 3: 10 Different documents written by different writers

Document	Accuracy
Doc 1	92.12%
Doc 2	91.27%
Doc 3	94.53%
Doc 4	91.86%
Doc 5	95.46%
Doc 6	97.75%
Doc 7	94.46%
Doc 8	93.42%
Doc 9	92.16%
Doc 10	96.72%

Average Accuracy = 94.04 %

As such, we have achieved an overall average accuracy of 93.51% for the segmentation of isolated and touching characters in the present work.

References

[1] Casey R G, Lecolinet E. A Survey of Methods and Strategies in Character Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 18(7): 690-706.

[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 Analysis and Recognition, 2003, 1128-1132.

[7] Ikeda H, Ogawa Y, Koga M, Nishimura H, Sako H, Fujisawa H. A Recognition Method for Touching Japanese Handwritten Characters. In the proceedings of 5th International Conference on Document Analysis and Recognition, 1999, 641-644.

[8] Zhang D X. Extraction of Embedded and/or Line-Touching Character like Objects. Pattern Recognition. 2002, 35(11): 2453-2456.

[9] 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

[10] 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.

[11] Reddy L P, Babu T R, Rao N V, Babu B R. Touching Syllable Segmentation using Split Profile Algorithm. 2010, 7 (3):17-26.

[12] Pal U, Belaid A and Choisy Ch. Touching numeral segmentation using water reservoir concept, *Elsevier Science Inc*, 2003, 24(1-3): 261-272.

Authors' Profiles

Munish Kumar received his Bachelors degree in Information Technology from Punjab Technical University, Jalandhar, India in 2006 and Post Graduate degree in Computer Science & Engineering from Thapar University, Patiala, India in 2008. He started his carrier as Assistant Professor in computer application at Jaito centre of Punjabi university, Patiala. He is working as Assistant Professor in Panjab University Rural Centre, Kauni, Muktsar, Punjab, INDIA. He is currently pursuing PhD degree from Thapar University, Patiala, Punjab, India. His research interests include Character Recognition.

Professor Manish Kumar Jindal received his Bachelors degree in science in 1996 and Post Graduate degree in Computer Applications from Punjabi University, Patiala, India in 1999. He received his PhD degree in computer science & engineering from Thapar University, Patiala, India in 2008. He is working as Associate Professor in Panjab University Regional Centre, Muktsar, Punjab, INDIA. His research interests include Character Recognition.

Professor Rajendra Kumar Sharma received his PhD degree in mathematics from the University of Roorkee (Now, IIT Roorkee), India in 1993. He is currently working as Professor at Thapar University, Patiala INDIA, where he teaches, among other things, queuing models and its usage in computer networks. He has been involved in the organization of a number of conferences and other courses at Thapar University, Patiala. His main research interests are in traffic analysis of Computer Networks, Neural Networks, and Pattern Recognition.

How to cite this paper: Munish Kumar, M. K. Jindal, R. K. Sharma,"Segmentation of Isolated and Touching Characters in Offline Handwritten Gurmukhi Script Recognition", International Journal of Information Technology and Computer Science(IJITCS), vol.6, no.2, pp.58-63, 2014. DOI: 10.5815/ijitcs.2014.02.08