A Technique for Segmentation of Gurmukhi Text

G S Lehal[1] and Chandan Singh[2]
[1]Department of Computer Science and Engineering, Thapar Institute of Engineering & Technology, Patiala, India. gslehal@mailcity.com
[2]Department of Computer Science and Engineering, Punjabi University, Patiala, India.

Abstract. This paper describes a technique for text segmentation of machine printed Gurmukhi script documents. Research in the field of segmentation of Gurmukhi script faces major problems mainly related to the unique characteristics of the script like connectivity of characters on the headline, two or more characters in a word having intersecting minimum bounding rectangles, multi-component characters, touching characters which are present even in clean documents. The segmentation problems unique to the Gurmukhi script such as horizontally overlapping text segments and touching characters in various zonal positions in a word have been discussed in detail and a solution has been proposed.

1. Introduction

Text segmentation is a process in which the text image is segregated into units of patterns that seem to form characters. All recognition algorithms depend on the segmentation algorithm to break up the image into individual characters Many papers concerning the segmentation of strings consisting of English letters and numerals have been published. The recent surveys on this topic can be found in references [1-2]. Some papers dealing with segmentation of different Indian language scripts such as Bangla, Devnagri and Gurmukhi have also appeared in literature[3-7], but none of the paper has dealt in detail with the practical problems peculiar to the Indian language scripts such as horizontally overlapping text segments and touching characters in various zonal positions in a word.

2. Characteristics Of Gurmukhi Script

Gurmukhi script is used primarily for the Punjabi language, which is the world's 14th most widely spoken language. Some of the major properties of the Gurmukhi script are:

- Gurmukhi script alphabet consists of 41 consonants, 12 vowels and 3 half characters which lie at the feet of consonants (Fig 1).
- A majority of the characters have a horizontal line at the upper part (Fig 1). The characters of words are connected mostly by this line called head line and so there is no vertical inter-character gap in the letters of a word and formation of merged characters is a norm rather than an aberration in Gurmukhi script The words are, however, separated with blank spaces.
- A word in Gurmukhi script can be partitioned into three horizontal zones (Fig 2). The upper zone denotes the region above the head line, where vowels reside, while the middle zone represents the area below the head line where the consonants and some sub-parts of vowels are present. The lower zone represents the area below middle zone where some vowels and certain half characters lie in the feet of consonants. A statistical analysis of Punjabi corpus has shown the zone-wise percentage distribution

of Gurmukhi symbols in printed text as : upper zone(25.39%), middle zone (70.41%) and lower zone (4.20%).

- The bounding boxes of 2 or more characters in a word may intersect or overlap vertically.
- The half characters in the lower zone frequently touch the above lying consonants in the middle zone. Similarly closely lying upper zone vowels frequently touch each other.
- There are many multi-component characters in Gurmukhi script. A multi-component character is a character which can decompose into isolated parts. (e.g. ਸ਼, ਖ਼, ਜ਼, ਗ਼, ਫ਼, ੵ)

Fig 1: Gurmukhi script character set

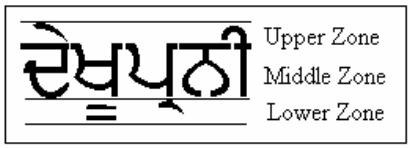

Fig 2: Three zones of a Gurmukhi word

3. Proposed Technique

After digitization of the text, the text image is subjected to pre-processing routines such as noise removal, thinning and skew correction. The thinned and cleaned text image is then sent to the text segmenter, which segments each uniform text zone into text lines and text lines into words. Words are further segmented into characters and sub-characters. To simplify character segmentation, since it is difficult to separate a cursive word directly into characters, a smaller unit than a character is preferred. In our current work, we have taken an 8-connected component as the basic image representation throughout the recognition process and thus instead of character segmentation we have performed *connected component segmentation*. A combination of statistical analysis of text height, horizontal projection and vertical projection and connected component analysis is performed to segment a text image into connected components.

3.1 Line segmentation

Horizontal projection, which is most commonly employed to extract the lines from the document[3-5], fails in many cases when applied to Gurmukhi text and results in over segmentation or under segmentation. Over segmentation occurs when the white space breaks a text line into 2 or more horizontal text strips (Fig 3a). Under segmentation occurs when one or more vowel symbols in upper zone of a text line overlap with modifiers present in lower zone of previous line As a result, white space no longer separates 2 consecutive text lines and two or more text lines may be fused together (Fig 3b). So special care has to be taken for these cases.

The text image is broken into horizontal text strips using horizontal projection in each row. The gaps on the horizontal projection profile are taken as separators between the text strips. Each text strip could represent a) Core zone of one text line consisting of upper, middle zone and optionally lower zone (core strip). b)Upper zone of a text line (upper strip). c)Lower zone of a text line (lower strip). d)Core zone of more than one text line (multi strip). The next task is to identify the type of each strip.

(a)

(b)

Fig 3 : Example of a) Over Segmentation b)Under Segmentation. Red line indicates the overlap region.

For this we first calculate the estimated average height of the core strip. We found that we cannot directly take the arithmetic mean of all the strips as the average height of the core strip, since the height of the upper strips, lower strips and multi strips present in the document image can greatly influence the overall figure. From experiments (table 1), it was observed that more than 60% of the text strips were core strips and thus 75 percentile of height of all the strips closely represented the average height of a core strip. We call this height as *AV*. Once the average height of core strip is found, then the class of the strips is identified. If the height of a strip is lesser than 33% of *AV*, then the strip is a lower strip or an upper strip. If the height is greater then 150% of *AV*, then the strip is a multi strip. Otherwise the strip is core strip. To distinguish between upper and lower strips, we look at the immediate next core strip. We determine the spatial position of headline in the immediate next core strip, where the headline is found by locating the row with maximum number of black pixels. If the headline is present in upper 10% region of the core strip, then the previous strip is an upper strip else it is a lower strip. Next determine the accurate average height of a core strip (ACSH) by calculating the arithmetic mean of all core strip. This information will be used to dissect the multi strip into constituent text lines. The average consonant height (ACH) is also estimated by calculating the average height of the text lying below the headline in each core strip. This information is needed in the other segmentation phases. Some statistics of horizontal strips generated from 40 document images from books, laser print outs and newspaper are tabulated in table 1.

Table 1 : Statistics of Horizontal Strips

Strip Type	**Min height (Pixels)**	**Max height (Pixels)**	**Average Height (Pixels)**	**Percentage of occurrence**
Lower Strip	1	13	2.6	35.7%
Upper Strip	6	8	6.5	0.2%
Core Strip	36	63	46.7	60.5%
Multi Strip	73	101	87.2	3.6%

3.2 Sub division of Strips into smaller units

In the next stage, all the text strips are processed from top to bottom in the order in which they occur in text and divided into smaller components. If a strip is an upper or lower strip, then it is entirely made up of disconnected characters or sub characters. These characters can easily be isolated by scanning from left to right till a black pixel is found and then using a search algorithm finding all the black pixels connected to it. Each such connected component represents one character or sub character. The smallest connected component of a core strip is a word since all the consonants and majority of upper zone vowels in the word are glued with the headline , so there is no inter character gap and white space separates words.. The word may not contain the complete character images, as some of the characters or their parts may be present in the neighbouring strips. For segmentation of the strip into words vertical projection is employed by counting the number of black pixels in each vertical line, and a gap of 2 or more pixels in the histogram is taken to be the word delimiter. The word is then broken into sub-characters. First the position of the headline in the word is found by looking for the most dominant row in the upper half of the word. The connected component segmentation process proceeds in 3 stages. In the first stage the connected component touching the headline and present in the middle and upper zone are isolated. In the second stage the character sub-parts not touching the headline and lying in upper zone are segmented while the characters in lower zones are isolated in third stage. The black pixels lying on the headline are not considered while calculating the connected components, otherwise all the characters glued along the headline will be treated as a single connected component. Each connected component now represents a) a single character or b) a part of character lying in one of upper, middle or lower zone. The zonal position of the connected components, coordinate of the left most pixel in the bitmap and the amount of overlapping with other components in other zones is later used to cluster the connected components into characters.

A multiple core strip is made of multiple overlapping text lines, which cannot be separated by horizontal projection. To segregate the text lines, the statistics generated in first pass will be used. The zonal height is divided by Average core strip height (ACSH) to get an idea about the number of text lines present in the strip. To extract the first row, an imaginary cut is made at 0.75*ACSH. We have deliberately made a cut at 0.75*ACHS instead of ACSH, so that by chance the line does not cross any character lying in next text row. This cut will be slicing most of the words into two parts, but that does not create any problem since we are looking for connected components only. The portions of the words which have been sliced and are lying in next sub-strip will also be added to the connected component of current word, since physically they are still connected. The sub-strip is then split into words by vertical projection analysis using the same method as used in segmentation of core strip. For the next sub-strip a cut is made at ACSH height and the words are extracted in that sub-strip. This process continues till all the sub-strips have been segmented into words. Next the connected components present in the upper zone of the word are identified. The upper zone can also contain lower zone vowels of words lying in previous line. So a distinction is made using the distance of minimum bounding rectangle with the headline. We do not search for connected components in the lower zone, since if accidentally a connected component of upper zone of a word present in next line is encountered, then the search will lead to words present in next line and they will all be identified as lower zone symbols of current word.

4. Touching Characters

Segmentation of touching characters has been the most difficult problem in character segmentation. Various papers have appeared concerning the segmentation of touching characters[1, 8-9]. All these papers have dealt with Roman script text only. As already mentioned segmentation process for Gurmukhi script proceeds in both x and y directions since two or more characters of a word may be sharing the same x coordinate. So for segmentation of touching characters in Gurmukhi script, the merging points of the touching characters have to be determined both along the x and y axes. During our experiments we found a large percentage of touching characters in even clean machine printed Gurmukhi text. These touching characters can be categorized as follows as a)Touching characters in upper zone b)Touching characters in middle zone c)Lower zone characters touching with middle zone characters and d)Lower zone characters touching with each other (Fig. 4)

Fig 4 : Examples of touching characters a) touching characters in upper zone b)touching characters in middle zone c) Lower zone characters touching with middle zone characters d) Lower zone characters touching with each other

4.1 Touching characters/Connected Components in upper zone

Closely lying upper zone vowels frequently touch each other in even clean documents. Another common problem encountered is the merging of the dot symbol with other vowels or headline. In our experiments we found that about **6.9%** of the upper zone vowels were touching other vowels or merging with the headline.

A connected component(CC) in upper zone is a candidate for further splitting if it satisfies one of the following two conditions:

a) Width of the CC is more than 75% of ACH: In normal cases the width of a vowel in upper zone is less than 75% of average height of a consonant (ACH). So if the width of the CC exceeds 75% of ACH then it is assumed that we have multiple fused vowels. If the width of the CC is x, then the potential cutting point is found in the region x/4 to 3x/4. For determining the cutting point, the algorithm suggested by Kahan at el8] is used.
b) Presence of an eastward oriented stroke in the second half of the CC along the x-axis: A careful analysis of the shapes of upper zone vowels (Fig 1) reveals that there is no vowel in upper zone which has a junction in the second half along the x-axis and a stroke originating from that junction which is oriented in eastern direction and not touching the headline. Thus if there exists one such stroke, then it is not part of the vowel. We use this property to search for a joint in the second half of the CC. If it is present then check for existence of a stroke emerging from this joint which is oriented in eastern direction and not touching the headline. If such a stroke is present then it is disconnected from the main connected component. Using this technique we were able to successfully separate the merged dot symbol from the CC of an upper vowel (Fig. 5).

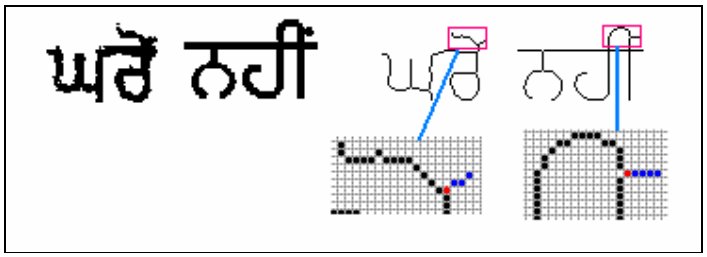

Fig 5 : Separation of touching connected components in upper zone. On the left side are the scanned images of text and their thinned versions are on the right. Blue segments represent the detected touching connected components

4.2 Touching characters/Connected Components in middle zone

On clean machine printed text, the frequency of occurrence of touching characters in the middle zone is found to be quite low **(0.12%)**. This is also the region, where the touching characters, if present, are most difficult to detect and split. Since the chances of touching characters in middle zone are very low, so we normally do not test for occurrence of touching characters in this region. The testing is done only if :
i) The width of a CC is more than 175% of ACH ii) The classifier fails to recognize the CC. To detect the cutting point, the method suggested by Kahan et al[8] is used. But a new problem was faced, as we are dealing with thinned images, it was found that sometimes a) The cutting point was not correctly found. This is because of thinned image the number of pixels is reduced and there is a shift in the direction of pixels because of thinning. This is illustrated in fig 6, where the image of touching character pair ਮਾ is incorrectly segmented . b) The cutting point may be correctly found but after separation into two CC's, the shape of the CC is so badly deformed that the classifier fails to recognize it. This can also be seen in fig 6, where the CC of the thinned image of touching character pair ਬਮ sans headline, is correctly segmented but the shape of the CC corresponding to ਮ is disfigured and it cannot be correctly recognized.
These twin problems were overcome by considering the unthinned version of the CCs. The original image is retained along with its thinned version and as there are sufficient number

of pixels in original scanned images, so the touching point is more easily identified. The image is then split at the touching point and the separated images are then thinned and sent to the classifier for identification (Fig 6).

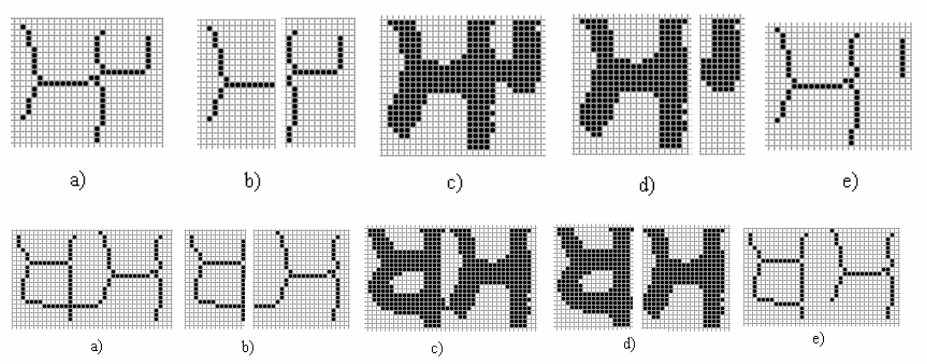

Fig 6 a) Touching Connected components b) Components separated by the algorithm c) Unthinned versions of connected components of fig 6a. d) Images of fig 6c split by Kahan method e) Thinned images of segmented images of fig 6d.

4.3 Connected Components in lower zone touching the above lying connected components

It was observed that the half characters and the vowels lying in lower zone frequently touch the above lying consonants. The frequency of lower zone characters touching the middle zone characters is found to be **19.1%**. For segmentation of these half characters, the average height of a consonant was used. It can be observed that for same font and size, the height of all consonants is almost same. So if the height of a consonant is found to be more than 120% of ACH then the consonant is topologically disconnected into two parts at the row near ACH with minimum number of black pixels. By default the cutting point is taken as the immediate row below the ACH number of pixels in the y-axis and we look up and below 10% of ACH number of rows for any row with fewer number of pixels. In case there exists such row, that row is taken as the cutting row. This method works well for most of the cases but fails in case where the touching lower zone character (such as vertical line like character such as the character ਼ ਼)does not much increase the height of the middle zone consonant (Fig 7).

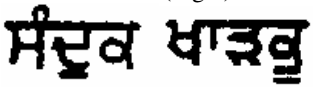

Fig 7 Scanned touching images of words ਸੰਦੂਕ ਖਾੜਕੂ

4.4 Connected Components in lower zone touching each other

In some rare instances the lower zone characters were found touching each other. The frequency of such occurrence though is very low **(0.031%)**. For segmentation of such character pairs, the simple technique of splitting the connected component at the middle of the horizontal axis served the purpose. For identification of such merged character pairs, the same criteria as used for upper zone vowels is used, that is if the width of a connected component in lower zone exceeds 75% of ACH then it is assumed that we have multiple fused vowels/half characters.

5. Conclusion

We have presented a scheme for decomposing a text image in Gurmukhi script into sub-characters or connected components. These connected components are then recognized by the classifier and merged to form characters. The various complexities such as absence of vertical inter-character gap in a word, horizontally overlapping text lines, multi-component characters, overlapping and intersecting circumscribing rectangles of characters, touching characters in various zonal positions in a word, presence of a lower character of a word in upper zone of a word in next line etc. have been taken care of by using this scheme. This is the first time that the problem of touching characters in Gurmukhi text has been studied in detail and solutions suggested for tackling the touching characters in various zonal positions of a word. Table 2 shows the accuracy rate of detecting and correctly segmenting the touching characters. These results are obtained by implementing and testing the proposed technique on about 40 machine printed Gurmukhi documents.

Table 2 : Accuracy rate of detection and segmentation of touching characters

Type of touching characters/connected components	**% of correct detection and segmentation**
Touching/merging upper zone vowels	92.5%
Touching middle zone consonants	72.3%
Touching middle zone and lower zone characters	89.3%
Touching lower zone characters	95.2%

References

[1] Lu, Y. : Machine Printed Character Segmentation – an Overview. Pattern Recognition, Vol. 28, (1995) 67-80

[2] Casy, R. G., Lecolinet, E.: A survey of methods and strategies in character segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, (1996) 690-706

[3] Chaudhuri, B. B., Pal, U. :A complete printed Bangla OCR system. Pattern Recognition, Vol. 31, (1998) 531-549

[4] Pal, U., Chaudhuri , B. B. : Printed Devnagri Script OCR System. Vivek, Vol. 10, (1997) 12-24

[5] Bansal, V. : Integrating knowledge sources in Devanagri text recognition. Ph.D. thesis, IIT Kanpur, INDIA. (1999).

[6] Goyal, A. K., Lehal, G. S., Deol, S. S. : Segmentation of Machine Printed Gurmukhi Script. Proceedings 9th International Graphonomics Society Conference, Singapore, (1999) 293-297

[7] Lehal, G. S., Singh, S. : Text segmentation of Machine Printed Gurmukhi Script. Document Recognition and Retrieval VIII, Paul B. Kantor, Daniel P. Lopresti, Jiangying Zhou, Editors, Proceedings SPIE, USA, Vol. 4307, (2001) 223-231

[8] Kahan, S., Pavlidis, T., Baird, H. S. : On the recognition of printed characters of any font and size. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 9, (1987) 274-287

[9] Liang, S., Shirdhar, M., Ahmed, M. : Segmentation of touching characters in printed document recognition. Pattern Recognition, Vol. 27, No. 6, (1994) 825-840