Isolated Handwritten Words Segmentation Techniques in Gurmukhi Script

Galaxy Bansal Dharamveer Sharma

ABSTRACT

Segmentation of handwritten words is a challenging task primarily because of structural features of the script and varied writing styles. Handwritten words are also 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. Segmentation accuracy of 72.6% has been achieved with the use of the algorithms for segmenting all types of words. Segmentation accuracy of 88.1% has been achieved for segmenting all types of handwritten words in Gurmukhi script. Further, different categories of overlapping and touching characters in all the three zones (upper, middle and lower zone) 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 proposed.

1. Introduction

The Gurmukhi script alphabet consists of 38 consonants (vianjans), 3 vowels, 9 vowel modifiers (laga or matra), 3 Auxiliary signs and 3 half characters. Handwritten Gurmukhi word segmentation is a challenging task since characters of handwritten words do not have fixed size and shape. So they are quite different from the printed characters. In the case of printed words, vertical bar of a character occupies a single column whereas that of handwritten words might occupy more than one column. The methods of printed words fail to work on the handwritten words. We are therefore inclined to develop methods that can work both on the printed and handwritten words.

Segmentation is a technique, which partitions handwritten Gurmukhi text or words into individual characters. Since recognition heavily relies on isolated characters, segmentation is a critical step for character recognition because incorrect segmentation may lead to incorrect character recognition. In character segmentation, broken characters and touching characters are responsible for the majority of errors in automatic reading of both machine-printed and handwritten texts.

In handwritten text, the characters or symbols overlap, touch or even merge with each other. The characters or symbols can be skewed. The headline can be skewed, broken or uneven or the word can be without headline. So, the objective is to segment those words into characters or symbols. The simple technique of using inter-character gap for segmentation is generally useful for good quality printed documents, but this technique fails to give satisfactory results if the input text is handwritten.

Some segmentation techniques for machine printed text can be found in references [1, 6, 10]. Many techniques have been proposed on segmenting words into characters in other Indian scripts like Bangla, Tamil Oriya and Hindi having similar properties of Gurmukhi script [2-5, 11]. Segmentation techniques for touching characters can be found in references [7, 9].

This paper is organized as follows: section 2 covers properties of Gurmukhi script, section 3 covers problems associated with structure of script and in handwritten words. In section 4, segmentation techniques for different types of words are explained. Section 5 covers the results and discussion.

2. Problems Associated with Structure of Script

1) Characters having small dot (considered as noise)

Another problem with Gurmukhi script is that it has six characters (**ਸ਼, ਖ਼, ਗ਼, ਜ਼, ਫ਼, ਲ਼**) having small dot known as bindi (.) at the lower portion of the character. So these small dots cannot be simply removed but must be saved. During recognition, when a character from these five characters is found, the location of noise pixel around this character can be used to find whether the character has bindi (.) or not. The head line can be used to separate characters from the word. From analysis, it is found that each character touches head line only once or at most two times. All pixels are connected except bindi (.). If the character touches only once then that position is the rightmost part or the middle of the character and has continuous head line. So, these characters can be easily separated.

2) Presence of head line

Segmentation focuses on separating text into lines and words and by determining the location of inter word gap. Handwritten text differs from printed text in that it lacks uniform spacing between two lines or words. Segmentation of Gurmukhi is totally different from that of roman script because there is no inter character gap in Gurmukhi as in Roman because all characters in a word are attached by a

horizontal line above characters called head line along with some symbols Kanna(ੰ), Bihari(ੀ), Sihari(ਿ) so there is only inter word gap and it is used for word separation. But separating characters from words is little bit difficult. Gurmukhi scripts are connected at the upper half of the word. So, upper half has maximum number of pixels in a row.

3) Gagga treated as Rarra and Kanna

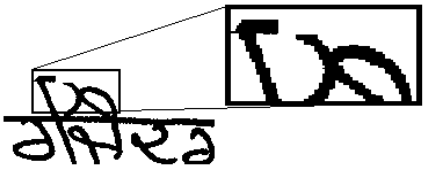
The characters which are connected to head line at two locations have a path from first location to second location except Gagga (ਗ). Using this path we can move from right location and separate the character. But Gagga (ਗ) is treated as two characters each touching the head line once. This is a combination of Rarra (ਰ) and Kanna(ੰ). So during recognition, when Rarra (ਰ) and next Kanna (ੰ) will be found, then this will be considered as Gagga (ਗ).

4) Touching characters

In handwritten Gurmukhi, adjacent characters touch each other and separation of such touching characters is a major problem. The problem in touching characters is that the characters are also attached in a zone other than the base line, so the vertical histogram of the word shows no white space between characters through which we can separate them. Because during segmentation of that character when we follow a path from the location where the character is attached to the head line but whenever we reach at the connected area then it seems to be different character so its segmentation is difficult. So, the connected characters result in under-segmentation.

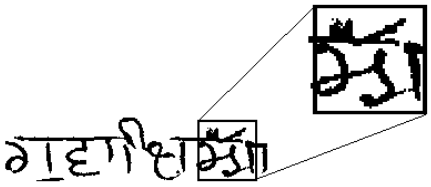
Existence of touching characters in any document decreases the recognition accuracy of OCR drastically. On statistical analysis of touching characters, we have made the following observations:

- Touching characters are found in all three zones (*i.e.*, upper, middle and lower zones) of a handwritten word in Gurmukhi script. Further touching characters can be divided into 5 categories:
  a) Upper zone characters touching with each other (as shown in Figure 1(a)).



(a)

  b) Upper zone characters touching with middle zone characters (as shown in Figure 1(b)).

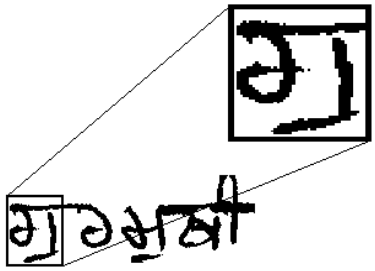


(b)

  c) Middle zone characters touching with each other (as shown in Figure 1(c)).

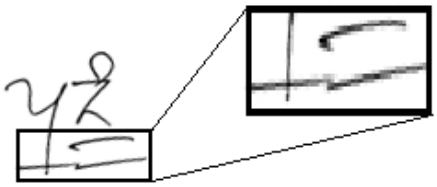
(c)

  d) Middle zone characters touching with lower zone characters (as shown in Figure 1(d)).



(d)

  e) The lower zone characters touching with each other (as shown in Figure 1(e)).



(e)

Figure 1: Touching characters (a) Upper zone- Upper zone, (b) Upper zone- Middle zone, (c) Middle zone- Middle zone, (d) Middle zone- Lower zone and (e) Lower zone- Lower zone

- Characters touch each other mostly at the centre of the middle zone, less frequently at top of the middle zone and rarely at the bottom of the middle zone.
- Touching characters have larger aspect ratio than that of individual character except the characters aira (ਅ) and ghaga(ਘ).

- In a single word, only two characters touch each other. The possibility of more than two touching characters is rare.
- In most of the cases, the vertical thickness of the blob at touching position is small as compared with the thickness of the stroke width. But in some cases, thickness may be equal or greater than the stroke width.
- Most of the characters of Indian scripts contain sidebars at their right end, *e.g.*, in Gurmukhi script 12 consonants have side bars at their right end. The possibility of touching is very high at this position.
- Another peculiar problem found in Gurmukhi text is that of characters touching across the neighboring lines. This introduces complexity not only in character segmentation but also in line segmentation.

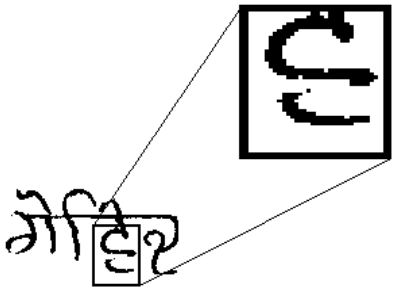
5) Broken characters

Missing character is another problem with the handwritten Gurmukhi script because during writing, some portion of the character may be missing. It is also difficult to segment the character because when we start tracing the path some portion may be missing so this leads to the error in the segmentation. So, the missing characters result in over-segmentation. Due to presence of broken characters the performance of any OCR may further decrease.

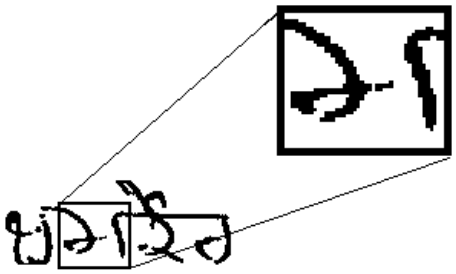
We have made the following observations:

- One character may be broken in more than one fragment.
- If spacing between the fragmented characters is less, it becomes difficult to determine which fragment belongs to which character.
- Most of the times, each fragment of broken character will have aspect ratio less than that of a single isolated character.
- Broken characters are mostly found in middle zone, less in upper zone and rarely in lower zone.
- The fragment of a character is generally not similar in shape of some other individual character.

The characters can be broken horizontally or vertically as shown in Figure 2(a) and Figure 2(b) respectively.



(a)



(b)

Figure 2: Broken characters: (a) Horizontally and (b) Vertically

6) Overlapping characters

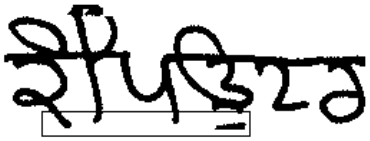
In handwritten text, the characters overlap each other most of the times i.e. vertical projection of any of the two characters or more overlap with each other.

On statistical analysis of 64 overlapping characters, we have made the following observations:

- Middle zone characters overlap with each other (as shown in Figure 3(a)).

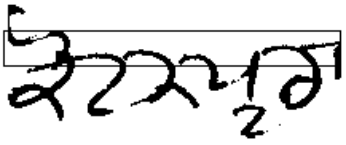
(a)

- Middle zone character overlaps with lower zone character (as shown in Figure 3(b)).



(b)

- Middle zone character overlaps with upper zone character (as shown in Figure 3(c)).



(c)

Figure 3: Overlapping characters (a) Middle zone- Middle zone, (b) Middle zone- Lower zone and (c) Middle zone- Upper zone

7) Words with skewed, uneven headline

In handwritten text, the recognition of head line is not easy as it may not be straight, it may not be continuous and at times, it may not exist at all. Some have habit of writing

without headline. So in these cases, it becomes difficult to identify headline and wrong segmentation is done by identifying wrong headline.

On statistical analysis of words with uneven headline we have made the following observations:

- Word can have uneven headline i.e. some portion of headline is missing, some is skewed, etc (as shown in Figure 4(a)).
- Word can have skewed headline (as shown in Figure 4(b)).
- Word can have broken headline (as shown in Figure 4(c)).
- Word can be without headline (as shown in Figure 4(d)).

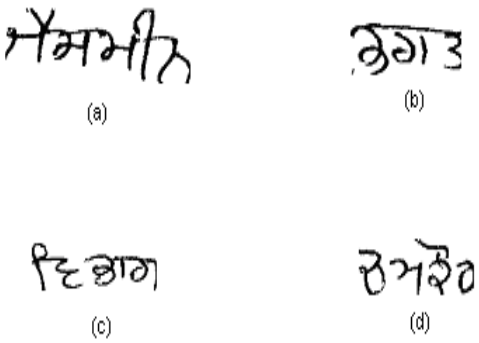


Figure 4: Uneven Headline (a) missing , skewed and broken headline, (b) skewed headline, (c) broken headline and (d) without headline

8) Words skewed to left or right

In handwritten text, the characters are skewed to left or right some times. By this skewness, the projections of two or more characters or symbols overlap with each other. These types of skewness of word are shown in Figure 5.

Figure 5: Word Skewness (a) Left skewed and (b) Right skewed

3. Segmentation of handwritten words in Gurmukhi script

Segmentation process involves separation of a word into characters, vowel modifiers, half characters. Before segmentation, the whole word is divided into three parts: upper, middle and lower zone. The zones are as shown in Figure 6 .

a) The upper zone denotes the region above the headline containing vowels, so identification of headline solves the purpose of finding vowels above headline.

b) The middle zone denotes the area below the headline where consonants and some subparts of vowels are present. Vowel in middle zone has single connectivity with headline, but consonant can be connected to headline at one or two locations.

c) The lower zone denotes the area below middle zone where some vowels and certain half characters lie in the foot of full character, so area below the middle zone is of minimum density.

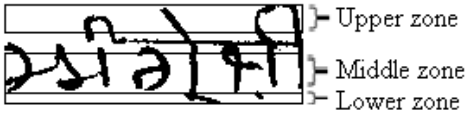


Figure 6: Classification of word into zones

Before discussing the problems of segmenting handwritten words in Gurmukhi script and proposing their solutions, we hereby give some definitions and notations used in various algorithms:

a) Horizontal Projection (HP): For a given binary image of size X × Y, where X is the width and Y is the height of the image, the horizontal projection is defined as:
HP(i), i = 1, 2, 3, …, Y.
where HP(i) is the total number of black pixels in ith row.

b) Vertical Projection (VP): For a given binary image of size X × Y, where X is the width and Y is the height of the image, the vertical projection is defined as:
VP(j), j = 1, 2, 3, …, X.
where VP(j) is the total number of black pixels in jth column.
Vertical histogram of word is shown in Figure 7.

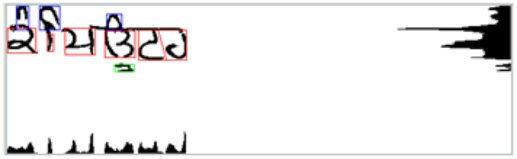
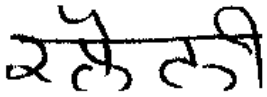


Figure 7: Horizontal and Vertical Histograms

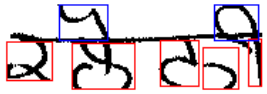
c) Headline Rectangle (HR): To overcome the problem of uneven headline, the row having maximum horizontal profile is displaced with a threshold value of 6 rows above and 6 rows below it. This rectangle is taken as headline rectangle. Its height will always be 13.

Detection of Headline

The header line is the most visible and distinguishing part of a word. By separating the header line we can obtain the upper and core-bottom parts of a word. For separation of header line, the horizontal projection (i.e., the number of pixels in each row) of the word is calculated and the region with maximum number of pixels of the word is identified. Since the header line covers the entire word, the region with the highest pixel density will give us the position of the line. In the handwritten characters, the header line covers multiple rows in contrast to printed characters whereas it covers a single row. The segmentation of word by detecting headline is shown in Figure 8.



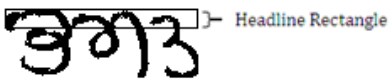
(a)



(b)

Figure 8: (a)Word with headline, and (b) correctly segmented

If the word is without headline, the maximum HP is compared with HPs starting from y=0 to y=h/3 where h is height of word. If there is not much difference between the values of HPs, then position of headline is replaced by that position of y. In Figure 9(a), headline rectangle is detected. The segmentation of word without headline is shown in Figure 9(b).



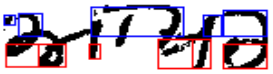
(a)

(b)

Figure 9: (a) Word without Headline and (b) Correct segmentation

Sometimes there are words having most of the characters without headline. So, in that case the headline detection gets wrong that is shown in Figure 10(a). The number of pixels in row that should be the headline is much less than the maximum number of pixels. The example of wrong segmentation is shown in Figure 10(b).

(a)



(b)

Figure 10: (a)Wrong headline detection and (b) Wrong segmentation

In this word, the characters aira(ਅ) and khakha(ਖ) are without headline.

Zone Detection

The row immediately below headline rectangle is taken as top of middle zone and the row immediately above headline rectangle is taken as bottom of upper zone. The bottom of middle zone is calculated by checking the value of HP of y starting from top of middle zone upto h. If the value of HP is less than 1, then that row is the bottom of middle zone.

Segmentation of middle zone characters or symbols

In middle zone, First of all, VP of each column starting from x = 0 to x = w (width) is taken. The characters or symbols can be straight, broken, connected, skewed or overlapped.

If the characters are straight i.e. without any problem like gaps in symbols then they are easily segmented by evaluating VPs. If VP is 0 or 1 for adjacent columns then that is the segmentation column.

The segmentation of middle zone characters or symbols is shown in Figure 11.

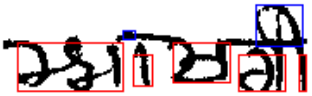


Figure 11: Segmentation of middle zone sub-symbols

Segmentation of overlapping characters in middle zone

If the characters or symbols are skewed, they are not segmented by evaluating VPs because their VPs overlap. So, to overcome the problem of overlapping characters or symbols, contour

tracing is done. The overlapping characters are shown in Figure 12(a).

Figure 12: (a) Overlapping sub-symbols and (b)correct segmentation

In this word, the character rara(ੜ) is overlapping with the symbol kanna(ਾ). The square tracing algorithm cannot be applied here. The reason is that there is concavity at the bottom of most of the Gurmukhi characters like ਓ, ਚ, ਝ, ਠ. If square tracing algorithm is applied to these type of characters, only the value of column will increment and value of row will remain same when it will reach the last row of character. It will not trace the whole character. So, we have applied a new technique for tracing. In this technique, the values of all the 8 pixels surrounding the current pixel are calculated. Correct segmentation by using contour tracing algorithm is shown in Figure 12(b).

The steps for contour tracing are shown as follows:

1. repeat steps while y > y_displace and x < right
2. check 8* 8 neighbourhood of each pixel starting from left bottom of segmented box
3. if any of these nine pixels is black
   then break.
   else
   increment the value of x by 1 and decrement value of y by 1.
4. proceed further according to the position of black pixel. If code is getting stuck because of repeating values then increment x by 2.

Joining of broken sub-symbols

If the characters or symbols are broken into sub-symbols, then to join them we calculated the number of pixels in the column three less than the left column of new sub-symbol. If the number of pixels is greater than 0, then earlier symbol is the part of new sub-symbol. Left of the earlier sub-symbol is made the left of new sub-symbol and top and bottom are adjusted accordingly. Without joining broken sub-symbols, the example is shown in Figure 13:

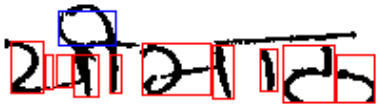


Figure 13: Segmentation without joining broken sub-symbols

Segmentation of middle zone sub-symbols from lower zone sub-symbols

Sometimes middle zone characters or symbols overlap with each other. To segment middle characters or symbols from lower symbols, we calculated the number of pixels in each row of segmented rectangle starting from rectangle's top. If this number of pixels is 0 in any row then that row is segmentation row. Segmentation of character mamma (ਮ) from symbol aunkar(ੁ) is shown below in Figure 14.

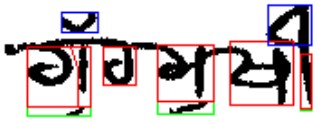


Figure 14: Segmentation of overlapping middle zone- lower zone sub-symbols

We approached a technique to segment middle zone characters touching with lower zone symbols. The technique is to find a row starting from rectangle's bottom to top so that the difference between the value of pixels of row below it and the row itself should be about 10. Segmentation is done as shown in Figure 15.

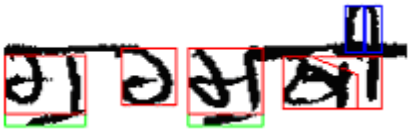


Figure 15: Segmentation of overlapping and touching sub-symbols

But the problem here is that most of the words in Gurmukhi have same difference in number of pixels. The example is shown in Figure 16.

Figure 16: Problem in segmentation of touching middle zone and lower zone sub-symbols

After segmentation of word into symbols or sub-symbols, binary image is created and each segment is stored as a new image.

4. Results and discussion

For implementing the algorithms proposed in this chapter, we selected a set of 2148 words has been considered.

The main objective was to discuss a complete solution for character segmentation phase of handwritten words in Gurmukhi script. As the problem of word segmentation is trivial, we have discussed in detail character segmentation problems which include methods to segment the overlapping characters or symbols and joining broken sub symbols present in all the three zones in handwritten words of Gurmukhi script.

For segmenting overlapping characters or symbols, we have scanned a number of words in different handwritings of Gurmukhi script.

Results for segmentation of words in various handwritings are given in .

Table 1.

Table 1: Percentage accuracy for different handwritings**Error! Reference source not found.**

Hand writing	Words without any overlapped, connected or merged characters: Correctly segmented/Tested	Words with overlapped, connected or merged characters: Correctly segmented/Tested	Over-segmented words: Correctly segmented/ Tested	%age accuracy
H1	9/12	10/13	9/15	76%
H2	17/19	7/10	14/16	82.8%
H3	15/21	7/13	11/12	64.7%
H4	7/8	6/15	5/14	56.5%
H5	27/32	3/6	4/10	78.9%
H6	18/23	4/7	11/17	73.3%

The ratio of correctly segmented words to total words tested of different types is given in columns 2, 3 and 4.

Sample words from Handwriting H1 are shown in Figure 17: Words from Handwriting H1Figure 17:

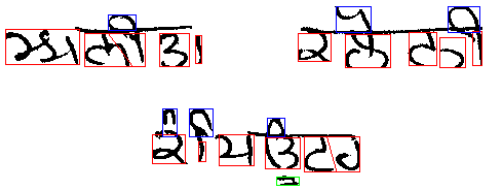


Figure 17: Words from Handwriting H1

Sample words from Handwriting H2 are shown in Figure 18:

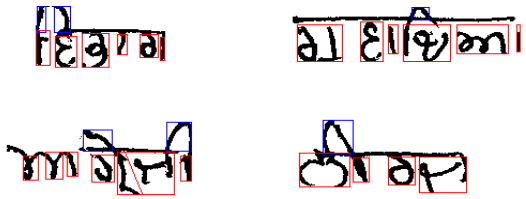


Figure 18: Words from Handwriting H2

Sample words from Handwriting H3 are shown in Figure 19:

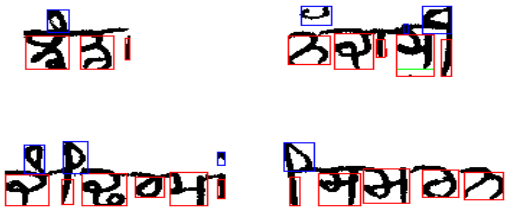


Figure 19: Words from Handwriting H3

In handwriting H3, most of the words have uneven headline.

Sample words from Handwriting H4 are shown in Figure 20:

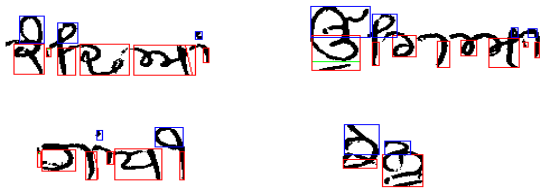


Figure 20: Words from Handwriting H4

Sample words from Handwriting H5 are shown in Figure 21:

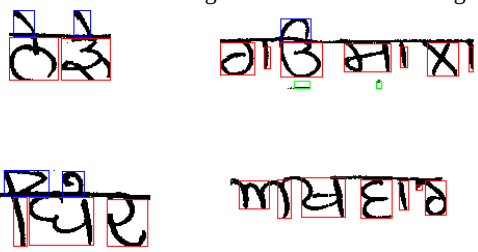


Figure 21:Words from Handwriting H5

Sample words from Handwriting H6 are shown in Figure 22:

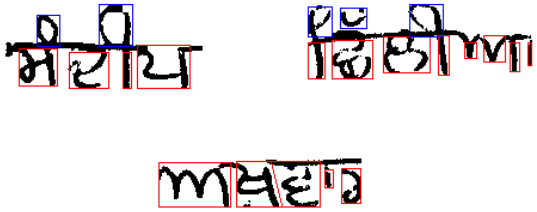


Figure 22: Words from Handwriting H6

The handwriting H4 has most of the words without headline. Its percentage accuracy is less because of wrong detection of headline. So, percentage accuracy is different for different handwritings.

Results of segmentation of all words tested are given in Table 2.

Table 2: Percentage accuracy for all types of words

Phase	Total words tested	Correctly segmented	%age accuracy
Words without any overlapped, connected or merged characters	1380 (1)	1116 (2)	80.9%
Words with overlapped, connected or merged characters	768 (3)	444 (4)	57.8%
Over-segmented words	1008	648	64.3%
Overall segmentation (1)+(3)	2148	1560	72.6%

In this database, most of the words have half character rara (ਾ ਿ ੀ) in lower zone that is touching with middle zone characters (as in Figure 18), character aira(ਅ**) that is overlapping with headline (as in** Figure 18 **to** Figure 22**) or words are without headline (as in** Figure 20**). Including these type of words is decreasing percentage accuracy.**

REFERENCES

[1] Sargur N. Srihari, "Machine Printed Character Segmentation Method using Side Profiles", Proc. SMC' 99, Vol. 6, pp. 863-867, 1999.

[2] U. Pal and S. Datta, "Segmentation of Bangla Unconstrained Handwritten Text", Proc. 7th ICDAR, pp 1128-1132, 2003.

[3] N. Shanthi and K. Duraiswamy, "Preprocessing Algorithms for the Recognition of Tamil Handwritten Characters", Proc. 3rd International CALIBER, Cochin, pp. 77-82, 2005.

[4] N. Tripathy and U. Pal, "Handwriting segmentation of unconstrained Oriya text", Proc. Sadhana Academy of Engineering Sciences, Vol. 31, Part 6, pp.755-769, 2006.

[5] M. Hanmandlu and P. Agrawal, "A structural approach for segmentation of handwritten Hindi text", Proc. The International Conference on Cognition and Recognition, pp. 589-597.

[6] M. K. Jindal, R. K. Sharma and G. S. Lehal, "Segmentation of Horizontally Overlapping Lines in Printed Indian Scripts", Proc. International Journal of Computational Intelligence Research, Vol. 3, No. 4, pp. 277-286, 2007.

[7] M. K. Jindal, R. K. Sharma and G. S. Lehal, "A study of touching characters in a degraded Gurmukhi text", Proc. World Academy of Science, Engineering and Technology, Vol. 4, pp.121-124, 2005.

[8] LI Yi, Yefeng Zheng, David Doermann, Stefen Jaeger, "Script independent text line segmentation in freestyle handwritten documents", pp. 1-28, 2006.

[9] Dharam Veer Sharma and G. S. Lehal, "An Iterative Algorithm for Segmentation of Isolated Handwritten Words in Gurmukhi script", Proc. 18th ICPR, pp. 1022-1025, 2006.

[10] M. K. Jindal, R. K. Sharma and G. S. Lehal, "Segmentation of Horizontally Overlapping Lines in Printed Gurmukhi Script", Proc. International Journal of Computational Intelligence and Research (IJCIR), pp. 226-229, 2006.

[11] A. Bishnu and B. Chaudhuri, "Segmentation of Bangla Handwritten Text into Characters by recursive Contour Following", 5th ICDAR, pp.402-405, 1999.

[12] G. S. Lehal and Chandan Singh, "A Gurmukhi Script Recognition System", Proc. 15th ICPR, Vol. 2, pp. 557-560, 2000.

[13] Manish Kumar, Dr. R. K. Sharma (TIET, Patiala) and Dr. G. S. Lehal (Punjabi University, Patiala), Ph.D. Thesis on "Degraded Text recognition of Gurmukhi Script", March 2008.