An Iterative Algorithm for Segmentation of Isolated Handwritten Words in Gurmukhi Script

Dharam Veer Sharma[1] and Gurpreet Singh Lehal[2]
Department of Computer Science, Punjabi University, Patiala-147002
Punjab, INDIA
dveer72@hotmail.com[1], gslehal@gmail.com[2]

Abstract

Segmentation of handwritten text in Gurmukhi script is an uphill task primarily because of the structural features of the script and varied writing styles. The presence of a horizontal line connecting characters of a word (i.e. head line), half characters and overlapping of some vowel between middle and lower zone of a word make the task even more difficult. Handwritten text is also prone to the problem of overlapped, connected and merged characters with in a word. Structural features are helpful in segmentation of machine printed text but these are of little help for segmentation of handwritten words. The proposed technique segments the words in an iterative manner by focusing on presence of headline, aspect ratio of characters and vertical and horizontal projection profiles. The proposed approach of segmentation can be used for handwritten text of Indian language scripts like Devnagri, Bangla etc. having structural feature similar to Gurmukhi script.

1. Introduction

Recognition of text heavily depends on proper segmentation of text into lines, words and then individual characters or sub-characters for feature extraction and classification of these characters. An error in segmentation may lead to wrong recognition of text and the system may be rendered useless. In this present work we have proposed a method of segmenting handwritten words in Gurmukhi script. Segmentation of handwritten words in Gurmukhi script is a challenging task because of the structural properties of Gurmukhi character set and writing styles of individuals.

Some surveys on segmentation techniques, for machine printed text, can be found in references [1, 2]. For segmentation of handwritten text also a survey paper is available [3]. Considerable amount of work has been carried out to segment words of machine printed Roman script and there are varied and some well developed techniques [4-8]. There are standard references available for segmentation of handwritten text in Roman script [9-11] as well. But very little work has been carried out for Indic scripts like Devnagari, Bengali, Gurmukhi etc. Only a few papers are available for segmentation of machine printed Devnagari [12,13] and Bangla [13] scripts, handwritten Bangla script [14,15] and machine printed Gurmukhi [16-18] scripts.

The next hurdle in this context is segmentation of handwritten text of Indic scripts. Through the present work, an effort has been made to clear this hurdle, especially for handwritten text of Gurmukhi script. There are some techniques available for segmentation of cursive text, which is inherently connected. But these techniques can not be successfully applied to Gurmukhi text, in which characters are connected through headline but there may be overlapping characters in upper and lower zones. The unique physical structure of Gurmukhi words such as connection of most of the characters of a word with headline and vertically and horizontally overlapping characters make the segmentation process more complicated as compared to other scripts.

This paper is organized as follows: section 2 covers previous work and related research, while the proposed algorithm is presented in section 3 and section 4 covers experimental results and conclusions of our research. For details on character set of Gurmukhi script and its properties refer [16-18].

2. Previous Work and Related Research

In the segmentation stage, the major challenge is to segment the overlapped (fig 1.a), connected (fig 1.b) and merged (fig 1.c) characters.

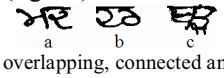


Fig.1: Examples of overlapping, connected and merged characters

In case of connected and merged characters the touching portions of the characters produce different

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

shape patterns as compared to primitive shape patterns, whereas patterns on the non-touching side of the characters are unaffected. The methods based on this feature of connected characters [7] is primarily intended for machine printed text and it fails for handwritten text, where the non connected sides of touching characters may be producing uneven patterns, extraneous strokes and overlapping if not connected. The procedure for segmentation of touching and fused characters of Devnagari script [12] is meant for machine printed text, where the presence of headline is easier to detect. This technique can not be applied to handwritten text as the headline may be uneven, broken or absent all together. Moreover, the decision that whether a region is containing connected characters or is consisting of one single character is difficult to make, because in handwritten text some characters may be having unusual width which may give impression of connected characters. Width of handwritten characters vary a lot, moreover characters in Gurmukhi script have huge dissimilarity in their width (as in fig. 2). The technique of identifying touching characters proposed in [13] also can not be successfully applied to handwritten text because in handwritten text even one character (as in fig. 2) may have aspect ratio greater than that of two connected characters.

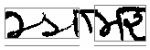


Fig.2: Characters with highly dissimilar aspect ratios.

Contour code feature based segmentation of handwritten text [10,14], alone, also fails to produce any acceptable results in case of handwritten words of Gurmukhi script and a segmentation accuracy of less then 70% is achieved, primarily because of broken, merged characters and dissimilar contours produced for the same characters written in different handwritings. Water reservoir methods of segmenting handwritten text in Bangla script [15] doesn't deal with the broken characters and because of characteristics of Bangla script most of the connected characters touch each other near headline area, where as in case of Gurmukhi script characters may connected in upper, middle or lower zones. The segmentation techniques proposed for machine printed Gurmukhi script [16-18] can also fail to segment handwritten text.

3. Segmentation of Handwritten words in Gurmukhi Script

Handwritten words in Gurmukhi script are prone to problems of overlapping characters in upper and lower zone, characters spanning upper and middle zones, connected and merged characters, characters of lower zone touching the middle zone characters. The proposed algorithm uses horizontal and vertical projections (histograms) from different zones of the word.

Horizontal Projection Profiles (HPPs) represent the horizontal count of pixels in each row of a rectangular region and is represented as HPP[i]; i = 1,2 …, H; where H is the height of the rectangular region.

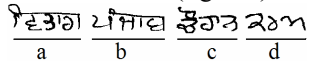
Vertical Projection Profiles (VPPs) represent the vertical count of pixels in each column of a rectangular region and is represented as VPP[i]; i = 1,2 …, W; where W is the width of the rectangular region.

Segmentation is carried out in three phases. Under first phase the word is segmented, under-segmented words are further segmented to the maximum possible extent in the second phase and in the third phase over segmentation is handled resulting from the first phase, which primarily exists because of broken characters n handwritten text.

Because of uneven strokes in handwritten text of Gurmukhi script, use of thinning is discouraged as this may lead to wrong over segmentation of words. Problems of segmentation of text, after thinning, has also been reported in [16-18].

3.1 Phase-I: Basic Segmentation

1. Handwritten words can be classified into three categories:
   a. Words with clear headline (fig.3.a),
   b. Words slanted (fig 3.b) or with broken (fig.3.c) headline and
   c. Words without headline (fig. 3.d).



a b c d

Fig.3: Different types of headlines

   Presence or absence of headline can be detected using horizontal projection profiles (HPPs) identify the position of the headline of the word. If the headline is not identified because of its absence then vertical projection profiles (VPPs) are created to identify gaps for segregation. In case of overlapping or connected characters phase 2 of the algorithm is applied. If the headline is not detected because of slant in the word then using VPPs the word is first bifurcated into N sub words and the segregation is done at the points where the value of VPPs is the least. Continuous runs of least values of VPPs are treated as gaps in the adjoining characters. Then these sub-words become target of segmentation and the whole process is reapplied.
2. Having found the location of headline, divide the word in two horizontal zones, one just above the headline and the other below the headlline.

3. This division will make gaps in the characters of the word which are segmented by creating vertical projection profiles.
4. Middle zone characters are identified using vertical projection profiles, created by considering the area just below the headline as the top starting point. A distance threshold value is used to prevent over-segmentation as some characters may be broken. This step gives us the number of columns formed in the word, which helps in associating characters in the upper and lower zone with characters in the middle zone.
5. For upper zone characters, consider the area just above the headline as the bottom and create VPPs. Any gap in the VPPs represents the cut point for segmentation of upper zone characters.
6. From the lower portion, just below the headline, again create horizontal projection profiles and from the bottom move upwards to find any gaps in the HPPs. Presence of any gap in the HPPs represents presence of characters in the lower zone. If no gaps are found up to a threshold then either the lower zone characters are not present or they may be overlapped or connected with the middle zone characters as shown in fig.4. Character (a) is overlapping two characters of the middle zone and character (b) is connected with a character

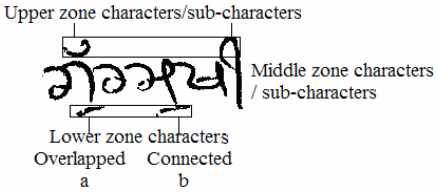


Fig. 4: Lower zone characters overlapped and connected with middle zone characters

7. A threshold value is set for area of a character and any value below that is ignored as that can be result of over segmentation or noise.

3.2 Phase-II: Handling under-segmentation

1. Use of VPPs does not help in segmenting overlapped or connected character.
2. For middle zone characters a threshold value of aspect ratio is used to identify whether the characters have been segmented properly or they are under segmented.
3. If the aspect ratio is greater than the threshold value then three different techniques can be applied.
   a. First, the under segmentation may be because of presence of a character in lower zone which is overlapping the area of two characters, as given in fig. 4(b).
   b. If there is no character in lower zone then for overlapping characters, starting from left-bottom and moving towards right, contour tracing is done to identify path between the characters (as given in fig. 5).

Fig.5: Overlapped characters in middle zone

If the path is found then the overlapped characters are segregated, as in fig.6.

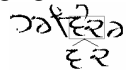


Fig.6: Segmentation of overlapped characters

If contour tracing leads to the right-bottom part of the character then the characters may be connected, as given in fig.7.

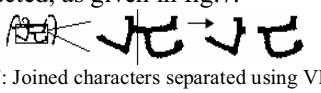


Fig.7: Joined characters separated using VPPs

For connected character again the procedure of VPPs is applied, ignoring the headline, and segregation is made starting from the middle of the connected characters and moving in both the directions, looking for minimum value of VPP, for making the cut. If traversal in both directions leads to extreme ends then either the characters are connected at more then one points or they are merged.

   c. For characters merged at more than one place, recognition based segmentation, or using heuristics is applied which is also called segmentation free recognition [8,9]
4. However, there are two characters (ਅ Aira) and (ਘ Ghagha) whose aspect ratio is far more than that of other characters and exceed the threshold value. But since both of these do not have headline over them, these are not considered as under-segmented.
5. In the lower zone, half characters and vowels may be connected with the lower portion of the consonants, as in fig. 4(b). For these the HPPs are used and from bottom a sudden decrease in the values of HPPs, to the width of the line, provides the point for segregation. Character given in fig. 9 have two such points in their HPPs and as such are not wrongly segmented.

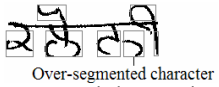
ਟ, ਡ, ਢ, ਣ, ਭ, ਦ, ਫ, ਭ, ਸ

Fig.8: Characters with two points of least counts in their HPPs

3.3 Phase-III: Handling over-segmentation

For over segmented characters, as in fig. 10, the property of Gurmukhi script, that all characters of middle zone are connected to headline, plays an important role.

0-7695-2521-0/06 $20.00



Over-segmented character

Fig.9: Over-segmented character in middle zone

1. There are two different cases of over-segmentation one for (ਸ and ਘ) and the other for all other characters.
2. If a middle zone character is not connected to head line then this is an over-segmented character, which should be merged with a character lying either on its left or on its right. The choice of candidate character with which the character is to be merged is made by comparing the distance of the sub-character with characters on both sides and the sub-character is merged with the nearest sub-character, which connects to the headline.

4. Results and Conclusion

For testing of the algorithm a set of 1907 handwritten words has been considered, from which 389 sets of connected characters were extracted after first phase of segmentation procedure. These 389 character sets belonged to 234 (20.39% of 1907) words, of these 389 connected character sets 3 sets each were present in 23 words, 2 sets each were present in 109 words and 102 words had only one set each of connected characters. Results of segmentation in each phase are given in table 1.

Table 1: Results of segmentation in each phase

Phase	**Words**	**Correctly Segmented**	**%age**
Phase I: Words without any overlapped, connected or merged characters	1673 (1)	1409 (2)	**84.22**
Phase II: Words with overlapped, connected or merged characters	234 (3)	189 (4)	**80.77**
Phase III: Over-segmented Words (1)-(2)	264	237	**89.77**
Overall Segmentation (1) + (3)	**1907**	**1835**	**96.22**

In the third phase, the remaining 264 words, with over-segmented characters are handled and of these 234 (95.12% of 246) words are properly segmented and the remaining (4.88%) error was primarily because of broken characters with gaps greater then the threshold value. The overall successful segmentation achieved through the proposed procedure is 96.22% (i.e. 1835 words out of 1907 total words). The errors of over-segmentation were unavoidable because of the gaps in the broken characters. Any readjustment of the threshold value leads to high degree of under-segmentation in the words and therefore is not recommended.

References

[1] Y. Lu, "Machine Printed Character Segmentation - an Overview", *Pattern Recognition,* vol. 28, No. 1, pp. 67-80, 1995.

[2] R. G. Casey, E. Lecolinet, "A Survey of Methods and Strategies in Character Segmentation", IEEE Tran. on PAMI, vol. 18 No. 7, pp. 690-706, July 1996.

[3] C. E. Dunn, P. S. P. Wang, "Character Segmentation Techniques for Handwritten Text-A Survey", Proc. 11th Int. Conf. on Recognition Methodology and Systems, vol. II, pp. 577-580, 30 Aug.-3 Sept. 1992.

[4] S. Tsujimoto, H. Asada, "Resolving Ambiguity in Segmenting Touching Characters", Proc. 1st ICDAR, pp. 701-709, 1991.

[5] S. Liang, M. Sridhar, M. Ahmadi, "Segmentation of Touching Characters in Printed Document Recognition", Pattern Recognition, vol. 27, pp. 825-840, 1992.

[6] T. Bayer, U. Kresel, "Cut Classification for Segmentation", Proc. IEEE vol. 80, pp. 1133-1149, July, 1992.

[7] M. C. Jung, Y. C. Shin and S. N. Srihari, "Machine Printed Character Segmentation Method using Side Profiles", Proc. Int. Conf. on Systems, Man, and Cybernetics, IEEE SMC '99, vol. 6, pp. 863-867, 12-15 Oct. 1999.

[8] C. Chen, J. DeCurtins, "Word Recognition in a Segmentation-Free Approach to OCR", Proc. 2nd ICDAR, pp. 573-576, Oct. 1993.

[9] X. Wang, V. Govindaraju, S. N. Srihari, "Holistic Recognition of Handwritten Character Pairs", Pattern Recognition, vol. 33, pp. 1967-1973, 2000.

[10] B. Verma, "A Contour Code Feature Based Segmentation for Handwriting Recognition"", Proc. 7th ICDAR, pp. 1203 - 1207, 3-6 Aug. 2003.

[11] Y. Ariki, Y. Mot, "Segmentation and Recognition of Handwritten Characters using Subspace Method", Proc. 3rd ICDAR, Vol. 1, pp. 120-123, 14-16 Aug. 1995.

[12] V. Bansal, R.M.K. Sinha, "Segmentation of Touching and Fused Devanagari Characters", Pattern Recognition, vol. 35, pp. 875-893, 2002.

[13] U. Garain, B. B. Chaudhuri, "Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts using Fuzzy Multifactorial Analysis", Proc. 6th ICDAR, pp. 805-809, 10-13 Sept. 2001.

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

[15] U. Pal, S. Datta, "Segmentation of Bangla Unconstrained Handwritten Text", Proc. 7th ICDAR, pp. 1128-1132, 3-6 Aug. 2003.

[16] A. K. Goyal, G. S. Lehal, S. S. Deol, "Segmentation of Machine Printed Gurmukhi Script", Proc. 9th Int. Graphonomics Society Conf., Singapore, pp. 293-297, 1999.

[17] G. S. Lehal, C. Singh, "Text Segmentation of Machine Printed Gurmukhi Script", Proc., of SPIE, vol. 4307, San Jose, USA, 2000.

[18] G. S. Lehal, C. Singh, "A Gurmukhi Script Recognition System", Proc. 15th ICPR, vol. 2, pp 557-560, Barcelona, Spain, 2000.