International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:4, 2007

A Study of Touching Characters in Degraded Gurmukhi Text

M. K. Jindal, G. S. Lehal, R. K. Sharma

***Abstract*—**Character segmentation is an important preprocessing step for text recognition. In degraded documents, existence of touching characters decreases recognition rate drastically, for any optical character recognition (OCR) system. In this paper a study of touching Gurmukhi characters is carried out and these characters have been divided into various categories after a careful analysis. Structural properties of the Gurmukhi characters are used for defining the categories. New algorithms have been proposed to segment the touching characters in middle zone. These algorithms have shown a reasonable improvement in segmenting the touching characters in degraded Gurmukhi script. The algorithms proposed in this paper are applicable only to machine printed text.

***Keywords*—**Character Segmentation, Middle Zone, Touching Characters.

I. Introduction

As part of the optical character recognition (OCR), character segmentation techniques are applied to word images before individual characters are recognized. The simplest way to segment the characters is to use inter-character gap as segmentation points. This technique does not work well if the text to be segmented contains touching characters.

The motivation behind writing this paper is that in a poor quality text page, degradation causes many problems such as: adjacent characters can touch one another; a character may be broken into several pieces; random noise or ink smears may make a character distorted. With the presence of such problems, for many word images, it is difficult to correctly determine their identities. Therefore, many recognition errors and uncertainties remain unresolved if the text image is highly degraded. The degraded texts mostly occur in xeroxed pages, fax massages, typewriter-printed pages, dot matrix printed pages, noisy images, images with blur or skew etc. Touching character is also one kind of degradation that may decrease the recognition results drastically.

Many algorithms have been proposed in the past [1-3] for segmenting clean English script. There is also much reported work available for segmenting fine printed Gurmukhi [4, 5], Devanagari [6, 7] and Bangla [7, 8] scripts. But very less work has been found on segmenting the touching characters of any Indian language. T. Hong [9] has utilized visual inter-word constraint available in a text image to split word images into pieces for segmenting degraded English language.

In this paper, we have proposed algorithms to segment touching Gurmukhi script characters. At the outset, a database has been prepared after scanning a number of poor quality printed documents containing 20-30% touching characters. Then all the touching characters were carefully analyzed and various categories are identified based on the structural properties of the Gurmukhi characters. After that, algorithms have been developed to segment the touching characters in middle, upper and lower zone.

M. K. Jindal is with the Panjab University Regional Centre, Muktsar (Punjab) India. (e-mail: mk1_jindal@ yahoo.co.in).

G. S. Lehal is working as Professor, in the Department of Computer Science & Engineering, in Punjabi University, Patiala (Punjab) India. (e-mail: gslehal@lycos.com).

R.. K. Sharma is Professor in School of Mathematics and Computer Applications, at Thapar Institute of Engineering & Technology , Patiala(Punjab) India(e-mail: rksharma@tiet.ac.in).

II. Characteristic of Gurmukhi Script

A. Gurmukhi characters

Gurmukhi script alphabets consist of 41 consonants and 12 vowels as shown in fig 1. Besides these, some of the characters in form of half characters are present in the feet of characters. Writing style is from left to right. The concept of upper/lower-case characters is absent in Gurmukhi. A line of Gurmukhi script may be partitioned into three horizontal zones, the middle zone being the busiest one. These zones are shown in fig. 2. The upper and lower zones may contain parts of vowel modifiers and diacritical markers.

In Gurmukhi Script, most of the characters, as shown in fig.1, contain a horizontal line at the top of the middle zone. This line is called the headline. The characters in a word are connected through the headline along with some symbols as ਾ, ਿ, ੀ etc. The headline helps in recognition of script line positions and character segmentation. The segmentation problem for Gurmukhi script is entirely different from scripts of other common languages such as English, Chinese, and Urdu etc. In Roman script, windows enclosing each character composing a word do not share the same pixel values in horizontal direction. But in Gurmukhi script, as shown in fig. 2, two or more characters/symbols of same word may share the same pixel values in horizontal direction. This adds to the complication of segmentation problem in Gurmukhi script. Because of these differences in the physical structure of Gurmukhi characters from those of Roman, Chinese, Japanese and Arabic scripts, the existing algorithms for character segmentation of these scripts may not work efficiently for Gurmukhi script.

Consonants				
ੳ	ਅ	ੲ	ਸ	ਹ
ਕ	ਖ	ਗ	ਘ	ਙ
ਚ	ਛ	ਜ	ਝ	ਞ
ਟ	ਠ	ਡ	ਢ	ਣ
ਤ	ਥ	ਦ	ਧ	ਨ
ਪ	ਫ	ਬ	ਭ	ਮ
ਯ	ਰ	ਲ	ਵ	ੜ
ਸ਼	ਜ਼	ਖ਼	ਫ਼	ਗ਼
Vowels in Upper zone				
Vowels in Upper and Middle zone				
Vowels in Middle zone				
Vowels in Lower zone				
Half characters in Lower zone				

Fig. 1 Gurmukhi script characters and symbols

Fig. 2 Three zones of Gurmukhi script characters

(a) Upper zone: from line no 1 to 2
(b) Middle Zone: from line no 3 to 4
(c) lower zone: from line no 4 to 5
(d) line no 2 start of head line
(e) line no 3 end of headline

B. Composition of Characters and Symbols for writing words in Gurmukhi Script

A horizontal line is drawn on top of all characters of a word, which is referred as headline. A character is usually written such that it is vertically separate from its neighbors. It is convenient to visualize a Gurmukhi word in terms of three zones as shown in fig. 2. The top and bottom zones may be empty for some words but only the vowels/half characters will be present in these zones.

III. Identification of Touching Characters

A. Data Collection

Data collection is a time consuming and difficult task. We selected true degraded documents containing touching characters from various books and magazines as well as normal documents, faxed them, copied them and scanned them at 300 dpi resolutions. About 500 such documents were scanned which contain almost 6000 touching characters, thus a sufficiently large database of touching characters is created. Fig 3 shows one paragraph taken from this database. This paragraph contains touching characters in middle, upper and lower zone

B. Categories of the touching characters in middle zone

After carefully analyzing the database of touching characters in middle zone, it is found that on the basis of

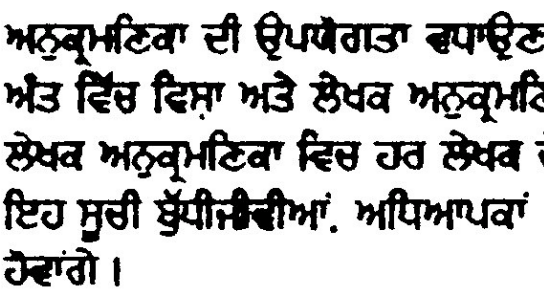

Fig. 3 Gurmukhi paragraph containing touching characters

structural properties of the Gurmukhi script, various touching characters can be classified among few categories. Some characters also fall in multiple categories. For each pair of touching characters, these categories are defined on the basis of left character of the pair. These categories are now briefly described.

Fig. 4 Words containing touching characters in middle zone

1) Category 1: Touching characters containing sidebars at right end

By carefully analyzing, it is found that 54% of the total pair of touching characters contains these characters at left side, which have sidebars at their right end. There are total 12 Consonants and one vowel in Gurmukhi script containing sidebars at right end, as mentioned below.

ਅ, ਸ, ਖ, ਗ, ਘ, ਜ, ਬ, ਧ, ਪ, ਬ, ਮ, ਯ, ।

For example, in fig 4 touching characters 1, 3, 5 and 6 are from this category.

2) Category 2: Touching Characters containing curved shape at right end

It has been revealed from the analysis that approximately 15% characters of the total touching characters fall in this category. Here, the touching character contains curved shape at right extreme end. There are 10 consonants in Gurmukhi script, namely, ਙ, ਞ, ਠ, ਡ, ਤ, ਲ, ਭ, ੜ, ੳ, ਨ fall in this category. Fig 4 contains touching character 8 from this category.

3) Category 3: Touching Characters containing little sidebar at right end

It has been found that approximately 10% characters of the total touching characters fall in this category. In this category, the characters contain a little sidebar at right side of the

character. Approximate size of the sidebar is half of the total length of the character. There are 7 consonants and one vowel in Gurmukhi script falling in this category and these are: ਏ, ਟ, ਦ, ਵ, ਛ, ਵ, ਫ, ਾ . Fig 4 contains touching characters 16 and 20 from this category.

4) Category 4 : Touching characters containing partial sidebar at right end

There are four consonants in Gurmukhi script falling in middle zone, that do not have full sidebar at their extreme right end but it contains 75-85% of the full sidebar. It has been seen that approximately 14% characters of the total touching characters fall in this category. These characters are: ਗ, ਰ, ਹ, ਚ . Fig 4 contains touching characters 2, 4 and 15 from this category.

C. Categories of the touching characters in upper zone

Following three categories are being proposed in the upper zone for touching characters.

Fig. 5 Gurmukhi words containing touching characters in upper zone (touching characters have been marked with circles)

1) Category 1: Bindi (ਂ) touching with other vowels

By carefully analyzing, it is found that 85% of the total pair of touching characters in upper zone fall in this category. In this category, vowel "*Bindi*" (dot shaped) touches with other vowels present in upper zone either form left or right side. Fig. 5.a contains words from Gurmukhi script in which *Bindi* touches with other vowels in upper zone.

2) Category 2: Adhak (ੱ) touching with other vowels

Approximately 10% touching characters of the total touching characters in upper zone fall in this category. In this category, a*dhak* vowel touches with other vowels present in upper zone. Fig 5.b contains some examples of *adhak* touching with other vowels in upper zone.

3) Category 3: Tippi (ੰ) touching with other vowels

It has been seen that 5% touching characters of the total touching characters in upper zone fall in this category. In this category, *tippi* vowel touches with other vowels present in upper zone. Further, it has been revealed from the analysis that the vowel *tippi* always touches with upper zone segment of the vowels ਿ, ੀ. Fig 5.c contains examples of *tippi* touching with upper zone segment of ਿ, ੀ in upper zone.

D. Categories of the touching characters in lower zone

Based on the analysis, we consider the following two categories in lower zone

Fig. 6. Touching characters in lower zone (touching characters have been marked with circles)

1) Category 1: Lower zone vowels touching with middle zone characters

Approximately 60-70 % of the total lower zone vowels always touch with middle zone characters. Even in non-degraded text this happens. Fig 6.a shows some example of this kind of touching characters.

2) Category 2: Lower zone vowels touching with each other

There is also a possibility of lower zone vowels touching with each other. Approximately 0.7% of the total lower zone vowels touch each other. Fig. 6.b shows this kind of touching pattern of characters in lower zone.

IV. Segmentation in Middle Zone

The above mentioned categories of touching characters are treated individually for segmentation, as detailed below. We have devised algorithms to segment the touching characters falling in above mentioned categories.

A. Algorithm for segmenting touching characters falling in first category

For segmenting the characters of a word having touching characters of the first category, we have developed the following algorithm.

Algorithm 1

Step 1	Recognize the headline for individual words by taking horizontal projection.
Step 2	Mark the start of the headline row and end of the headline row.
Step 3	Take the vertical projection between the word boundary.
Step 4	Note down all the positions of the columns having number of pixels equal to the height of the characters and touching the base line also. Identify these as sidebar columns.
Step 5	Starting from left of the word to right side whenever a continuos run of sidebar columns end change all the black pixels of the next column to white and that marks the segmentation of the touching character of this type.

Fig. 5 Horizontal & Vertical Projection of a touching word

Fig. 6 White dots showing start of headline, end of headline and possible positions of sidebar Columns

One can see horizontal and vertical projection of a word having touching characters in fig 5. Also, start of the headline and end of the headline in fig 6 have been marked by white marks in horizontal projection area. The possible positions of sidebar columns in fig 6 are marked by white marks in vertical projection area. Now we put a white line after these positions and segmentation is achieved as shown in fig 7.

Fig. 7 Touching characters segmented using first algorithm

This algorithm is based upon the structural property of Gurmukhi script that, in all the Gurmukhi characters if sidebar exists, it is always present at extreme right end of the character, in contrary with Devanagari and Bangla script, where it may be in the middle of the character. The advantage of this algorithm is that, we do not need to identify the candidate for segmentation. Also, more than two touching characters in a single word can be segmented using this algorithm and if the width of touching blob is greater than or equal to the width of the stroke, even then, this algorithm works.

B. Algorithm for segmenting touching characters falling in third and fourth category

A challenging task in segmenting the touching characters falling in this category is how to identify the little sidebar, which is approximately half of the total height of the character. We have developed following algorithm for this type of degradation (first three steps are same as that of algorithm 1)

1) Algorithm 2

Step 4 Note down all the positions of the columns which start from headline and no of pixels in which are approximately 50-80% of the total height of the character. Call these columns as partial sidebar.

Step 5 Starting from left of the word to right side whenever a continuous run of partial sidebar columns ends, change all the black pixels of the next column to white and that marks the segmentation of the touching character of this type.

This algorithm works fine if the characters are from category 4 but in case of characters from category 3 sometimes over segmentation may occur, since there are some characters from category 2 which have little sidebar at their middle or at extreme left end. Those characters are ਭ, ਠ, ਨ . A solution for this problem has been implemented by considering the fact that whenever the partial sidebar columns of height less than 60% of the total height of the character are detected, we segment the characters using algorithm 2. After marking the segmentation column, we see whether the segments created by segmentation process are attached to the headline or not. If segments are not attached with headline, it is taken as improper segmentation column and we in such a case remove the segmentation column; otherwise segmentation has been achieved.

C. Algorithm for segmenting characters falling in second category

For segmenting the characters falling in these categories, we have used the segmenting objective function suggested by Kahan et al. [2].

V. Results and Discussions

It has been found that by applying the algorithm 1 on degraded documents, almost 55-60% of the total touching characters in the middle zone have been correctly identified and segmented. After that, second algorithm is applied on the output of first algorithm and it was found that approximately another 20-25% of the total touching characters of initial degraded documents, in the middle zone, have been correctly identified and segmented. When we apply the technique for second category, another approximately 10% touching characters are segmented. As such, on the basis of all these algorithms, 92-95% of the touching characters in the middle zone have been correctly segmented. A major advantage of these algorithms is that these are capable of segmenting more than two touching characters in a single word.

References

[1] Y. Lu, "Machine Printed Character Segmentation – an Overview", *Pattern Recognition,* vol. 29, no. 1, pp. 67-80, 1995

[2] S.Kahan, T.Pavlidis, and H.S.Baird, " on the recognition of printed characters of any fonts and sizes", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 9, no. 2, pp. 274-288, Mar. 1987

[3] S. Liang, M. Sridhar and M. Ahmadi, "Segmentation of Touching Characters in Printed Document Recognition," Pattern Recognition, vol. 27, no. 6, pp 825-840, June 1994

[4] G. S .Lehal and Chandan Singh, "Text segmentation of machine printed Gurmukhi script", *Document Recognition and Retrieval VIII, Proceedings SPIE, USA, vol. 4307*, pp. 223-231, 2001.

[5] G.S.Lehal and Chandan Singh, "A technique for segmentation of Gurmukhi script", *Computer Analysis of Images and Patterns, Proceedings CAIP 2001,* Warsaw, Poland, Lecture Notes in Computer Science, vol. 2127 Springer-Verlag, pp. 191-200, 2001.

[6] Veena Bansal and R.M.K. Sinha , "Segmentation of touching characters in Devanagari," in Indian Conference on Computer Vision, Graphics and Image Processing, New Delhi: pp 377-380(1998)

[7] U. Garain, B.B. Chaudhuri, "Segmentation of touching characters in printed Devanagari and Bangla scripts using fuzzy multifactorial analysis", IEEE Trans. Systems Man Cybern. Part C-32 (2002) 449–459.

[8] U. Garain, B.B. Chaudhuri, "On recognition of touching characters in printed Bangla Documents", Proceedings of the Fourth International Conference on Document Analysis and Recognition, 1997, pp. 1011–1016.

[9] Tao Hong, **"**Degraded text recognition using visual and linguistic context**",** *a dissertation submitted to the faculty of the graduate school of the State University of New York at Buffalo,* 1995.