SEGMENTATION OF BROKEN CHARACTERS OF HANDWRITTEN GURMUKHI SCRIPT

Bharti Mehta
Department of Computer Engineering
Yadavindra college of Engineering
Talwandi Sabo (Bathinda)
bhartimehta13@gmail.com

Simpel Rani
Department of Computer Engineering
Yadavindra college of Engineering
Talwandi Sabo (Bathinda)
simpel_jindal@rediffmail.com

Abstract: *Character Segmentation of Handwritten Documents has been an active area of research and due to its diverse applicable environment; it continues to be a challenging research topic. The desire to edit scanned text document forces the researchers to think about the optical character recognition (OCR). OCR is the process of recognizing a segmented part of the scanned image as a character. OCR process consists of three major sub processes - pre processing, segmentation and then recognition. Out of these three, the segmentation process is the most important phase of the overall OCR process. Different problems in the characters segmentation of handwritten text is due to the different writing style of different people because the size and shape is not fixed while we write any text. In this work, we formulate an algorithm to segment the scanned document image as a character. According to proposed algorithm, broken characters in Gurmukhi script, we used the segmentation of these characters that can become easily identify how many characters are in one word. To develop the algorithm to segment the characters from a word we are using combinations of two approaches which are Horizontal Profile Projection and Vertical Profile Projection. And get the accuracy is 93%.*

Keywords: *Gurmukhi Script, OCR, Segmentation, Handwritten Document, Horizontal Profile Projection, Vertical Profile Projection.*

1. Introduction

Optical character recognition (OCR) is the prominent area of research in the world. OCR is the translation of scanned images of handwritten, typewritten or printed document into machine encoded form. This machine Optical character recognition is the most important area of research in the world. OCR is the translation encoded form is editable text and compact in size. Character recognition for handwritten numeral is more complex due to varying writing styles of people. Most commercial Optical character recognition systems are designed for well-formed business documents. Recognizing older typewritten document is also more challenging.

1.1 Characteristics of Gurmukhi Script

Gurmukhi script is mainly used for Punjabi language, i.e. the world's 14th most widely spoken language. The Gurmukhi script is derived from the Punjabi old word "Guramukhi", that means "from the mouth of

the Guru". In Gurmukhi script nine vowel symbols called laga or matras, forty-one consonants called vianjans, two symbols for nasal sounds (ਂ, ਁ), one symbol for reduplication of sound of any consonant (ੱ) and three half characters (੍ਹ, ੍ਰ , ੍ਵ) that lies at the feet of a consonants. Following are the properties of Gurmukhi Script are:

i. Writing style is from left to right.

Consonants and Vowel Carriers

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 ਖ਼ ਗ਼ ਜ਼ ਫ਼ ਲ਼

Vowels

ਾ ਿ ੀ ੇ ੈ ੋ ੌ ੁ ੂ

Additional Symbols

ਂ ਃ ੱ

Half Characters

੍ਹ ੍ਰ ੍ਵ

ii. No concept of upper and lower case characters.

iii. Gurmukhi script is cursive.

Gurmukhi Script has following challenges:

i. Variability of writing style, both between different writers and between separate examples from the same writer overtime.

ii. Low quality of text images

iii. Unavoidable presence of background noise and various kinds of distortions.

The character set of Gurmukhi script is shown in figure 1

Figure 1: Characteristics of Gurmukhi Script

In Gurmukhi script three zones are available i.e. upper, lower and middle zone. In figure 2 the upper zone of a word is above the headline where the vowels are located, the area below headline is middle zone where consonants and sub part of vowels are present and the lower zone is below the middle zone in which some half characters and some vowels are lie at the feet of consonants. The characters present in the middle zone touch the headline.

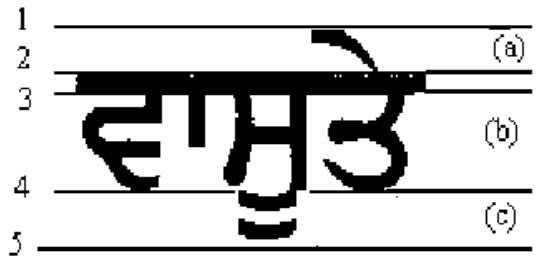


Figure 2: Gurmukhi Script Word, the line number 1 is called the start line, line number 2 is start of the headline and the 3 line number is end

of the headline Line number 4 is base line and 5 are called the end line [6].

2. Broken Characters

The broken characters in Gurumukhi script is mainly found in the middle zone, less in upper zone and rarely in lower zone [3]. It is very difficult to segment the characters. In handwritten Gurumukhi script the text may be missing and the broken characters contain horizontally and vertically [8]. As shown in Figure 3(a) and 3(b)

Figure 3: (a) horizontally broken characters

Figure 3: (b) vertically broken character

The characters can be broken by three phase:

- The characters can be broken from the headline.
- The characters can be broken horizontally.
- The characters can be broken vertically.

3. Segmentation

Segmentation is the main part of this paper; it is used to decomposition of images into sub parts [12].This process consisting of two different stages. In the first stage, the beginning of segmentation is performed that will execute text line segmentation. That leads to the separation of sub images related to each text line of the complete text. These sub-images contain more than one word. In the second stage, we will perform word segmentation which results into the sub-images which consisting of each word as a separate image. Segmentation is the technique which partitions the handwritten words into the individual characters [1]. Segmentation is the initial step of optical character recognition (OCR) for any language. Word segmentation is one of the core problems of OCR of handwritten text, which has long been an active area of research. Since recognition heavily relies on isolated characters, segmentation is a very critical step for character recognition [7]. The segmentation process follows the following pattern [12]:

- Identify the text lines in the page.
- Identify the words in individual lines.
- Identify the individual character in each word.

The segmentation has mainly three different types:

3.1 Line Segmentation

Line segmentation is to the very important part of segmentation of the character recognition process that extracts text lines and words out of the document images. The whole text document is of not for use due to the limitation of feature

extraction or classification phases. So we need to extract each word out of the document images for the use of feature extraction phase [1].

3.2 Word Segmentation

It is similar to the line segmentation, in word segmentation we find the horizontal plot of density of pixels seeking for minima's as inter word space. We crop the image using these minima's and the final output is an isolated word [1]. To segmenting the text line image into words, compute vertical projection profiles. Segmented line from first stage is taken as input for second stage that is word segmentation.

3.3 Character Segmentation

Character segmentation is done when the individual words are identified. To extract character from word removal of headline is essential. For this first horizontal projection of individual word is computed and the rows having highest projection is consider as a headline and removed for further character segmentation. After removal vertical projections of individual word is computed. Using these profiles separation of the base characters is done

4. Techniques used for segmentation

4.1 Horizontal Projection (HP)

In horizontal projection, for any given binary image of size P × Q where P is the height and Q is the width of image, the horizontal projection is defined by [8] as:

HP (i), i = 1, 2, 3,….., P, where HP (i) is the total number of black pixels in ith row [3].

4.2 Vertical Projection (VP)

In vertical projection, for any given binary image of size P × Q where P is the height and Q is the width of image, the vertical projection is defined by [8] as:

VP (j), i = 1, 2, 3, ..., Q, where VP (j) is the total number of black pixels in jth row [3].

4.3 Water Reservoir

The water reservoir method of segmentation mainly for touching characters to analyze the large space, if we pour the water from top and bottom of the numerals and characters, the cavity regions of characters and numerals where water will b stored are considered as reservoirs [10, 8]. As in Fig. 3

Fig. 3: A reservoir obtained from water flow from top is marked by dots [8].

5. Problems under Character Segmentation

There are some problems occur in the character segmentation. These are because of the different writing styles etc.

i. Problem of Broken Characters

Reason: - Sometimes the characters are broken due to improper functioning of the writing element [12] and also the pen may be stop while we write any text.

ii. Problem of Touching Characters

Reason: - The problem of touching characters is come because while we write a text then one character may touch the other character in any zone i.e. middle, upper and lower zone.

iii. Problem of Overlapping Characters

Reason: - This problem arises when one character is written above the other character.

iv. Problem of Skewed Characters

Reason: - In this problem the words are not written in the straight line i.e. horizontal line, the words are written in upward or downward. Then the segmentation of any character is quiet difficult.

v. Problem of Irregular intensity within the Characters

Reason: - This problem occurs due to the bad quality of writing material that leads to the different intensities of pixels.

vi. Problem of Gagga

Reason: - In this problem while we segment the character of Gagga (ਗ), then this character is treated as two different characters i.e. Rarra (ਰ) and Kanna (ਾ). So in this problem we can't segment the word Gagga(ਗ).

6. Related Work

Singh et. al. (2011), the word and line segmentation on Devnagari Script in parallel and sequential based on Graphics Processing Unit (GPU). The main goal of this research work is to make segmentation faster for processing of document images [1].

Sharma and Lehal (2006), the phases of segmentation in handwritten Gurmukhi script are described and also the proper segmentation of text into lines, words and then individual characters or sub-characters is elaborated. [2].

Bansal and Sharma (2010), In this paper the problems of Gurmukhi Script and there segmentation techniques of handwritten words, in these different types of categories are used like, overlapping and touching characters in all three zones. A method for segmenting overlapping characters in middle zone has been proposed [3].

Lehal (2009) presented a robust and font independent Gurmukhi OCR system, which performs on old documents. In this the OCR is based on four classifiers operating in serial and parallel mode. The problem of broken characters, which frequently appear in old documents, has also been tackled using a structural feature based algorithm [4].

Jindal et. al. (2007), in this paper we study about the touching of Gurmukhi characters that are divided into various categories. In this segmentation of middle zone of touching characters is described. The algorithm is more

capable to segment more than two touching characters in a single word [5].

Jindal et. al. (2009), in this paper author presented the segmentation technique of touching characters in upper zone, the technique based on the structural properties of Gurmukhi Script [6].

Kamble and Kamble (2011), in this paper the author represent towards the segmentation of Hindi handwritten Devnagari Text. This system deals with segmentation of modifiers (matras) and fused characters in handwritten Devnagari word. It also describe segmentation in hierarchical order i.e. firstly we identify the header line and segmented it. Segmentation of modifiers consists of top as well as bottom modifiers segmentation [7].

Kumar et .al. (2014), explained generalized information about hand written character recognition system. In order to detect and segment characters in scanned word of handwritten Gurmukhi script documents, they used vertical projection profile technique and water reservoir base area point's technique. These techniques have been applied on the documents of three different categories [8].

Kumar and Singh (2010), in this paper an approach to segment the scanned document image. In this initially consider the whole image as one large window, then the window is broken into smaller parts and identify the lines to find a word present in a line and finally to characters. We proposed a concept a variable sized window. And the window size can adjust according to their need and finally got the good results [9].

Pal et. al. (2003), in this paper the author will discuss about the touching segmentation of numeral by the use of water reservoir method. A reservoir is a symbol to show the region where numerals are touch. It is used by poured the water from top and bottom [10].

Rani and Kumar (2013), In this paper they describes about the problems comes in character segmentation of only handwritten Devnagari script and also the characteristics of Devnagari script, different zones present in the devnagari [12].

Below that the comparison table no. 1 of different techniques is given:-

Table 1: Comparison between Various Techniques

Technique	**Information or Process Used**	**Advantages**	**Disadvantages**
Horizontal Projection (3,8)	Counts the total number of black pixels in each horizontal row.	Capable of segmenting more than two characters in a single word.	Lines are extracted from paragraph only further another technique used
Vertical Projection (3,8)	Counts the total number of black pixels in each vertical column.	Capable of segmenting more than two characters in a single word	It is not work on the paragraph segmentation
Water Reservoir (10,8)	Fill cavity region by pouring the water.	No need of normalization.	Method fails if there is a break point on the contour**.**

7. Methodology

The segmentation can be done on the basis of zoning, line segment from text, word segment from line and character segment from word. This can be done by the use of horizontal, vertical and water reservoir method. Firstly remove the headline of the text, binarization of the text and skeletonization of the text. In this the text is divided into three zones i.e. upper, middle and lower zone with the these zones, we can find where the text is broken from this we can segment the text by using horizontal and vertical projection because it is gives more accuracy.

Pre-Processing Algorithm & Character Segmentation Algorithm:

This algorithm segments the characters from the words to form broken characters which can be segmented into characters. This algorithm consists of these following steps:

Step 1: Input the image of variable size.

Step 2: Apply Global threshold technique Otsu method.

Step 3: Find the connected components and their area. Remove those connected components whose area is less than 70.

Step 4: Apply Morphological operation named Erosion with structure elements square type with value 4.

Step 5: Using horizontal Profile Projection Technique by counting the frequency of 1's along each line to find the header line along with its neighbour pixels.

Step 6: Maximum index is founded by above step. Now, 10 rows above and 10 rows below forms a rectangle which is removed.

Step 7: Disclosure the image.

Step 8: Using vertical profile projection technique finds the gap between the two characters by comparing its neighbour pixels along Horizontal direction to check whether the character is broken or not. If gap is less than or equal to 9 pixels then consider it as same character.

Step 9: Segment the Characters and draw lines between two characters

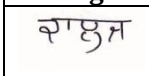
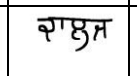
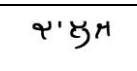
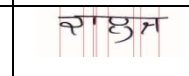
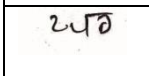
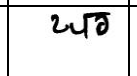
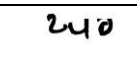
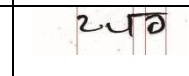
Step 10: Store the output in the file.

Step 11: Exit

8. Result and Discussion

We use Matlab 2010b as programming language to implement the algorithm we developed. The following are the snapshots which show various steps results performed during the algorithm on few database images. We have written the words on A4 paper sheets from which we want to segment the characters from the words and then scan to create a .Jpg image. We have tested our algorithm on more than 1000 handwritten words written in Gurmukhi script by different writers. The accuracy of the algorithm is 93%.

Table 2: Result of Original Image is discussed

Broken Word in Original Image	Eroded Image	Headline Removal	Correctly Segmentated Characters
			
			

In table 2, we take an original word image having broken characters, after doing binarization of the original image we apply morphological technique and get an eroded image and after that we remove its headline and at the end segmenting the original word and count the number of characters.

1. First problems is with letter 'ਗ' as after removal of headline it is considered as two characters 'ਰ' and 'ਾ'. Where the space between these two characters is less it considered as one character otherwise two characters as shown above some of the words are shown in table 3(a) and table 3(b).

Table 1(a): Considering the problem of 'ਗ'

wrongly segmented

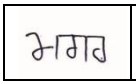
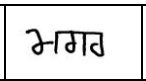
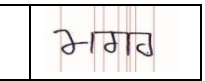
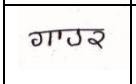
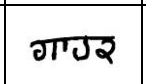
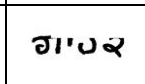
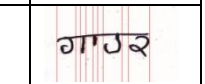
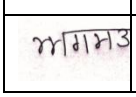
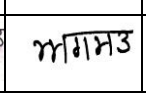
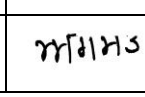
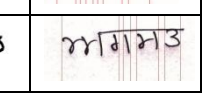
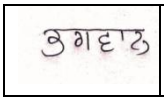
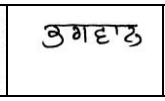
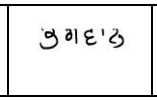
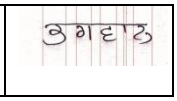
			
			
			

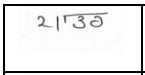
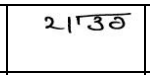
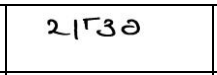
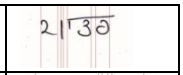
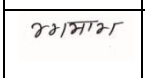
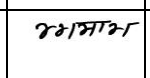
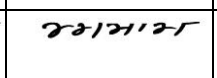
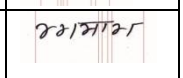
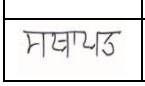
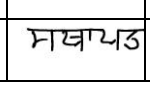
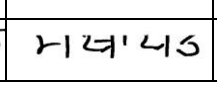
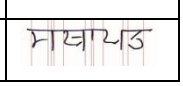
Table 3(b): Considering the problem of 'ਗ'

correctly segmented

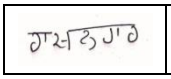
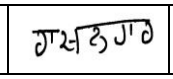
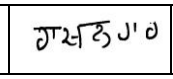
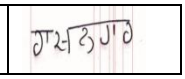
2. Second problem is due to gap between two characters as our algorithm works for 9 pixels difference. If we take less pixels or more some character counts get differs as shown in table 4.

Table 2: Considering the problem of pixel gap

3. Third problem is due to skewness in word. This problem has been solved for headline by removing the headline in a rectangular box of width 20 pixels as shown in Table 5.

Table 3: Considering the problem of Skewness

We have also applied the same database for pixel difference of 5 pixels and 9 pixels with two times erosion. Result in accuracy form we get 78% and 77% respectively as shown in table 6.

Table 4: Accuracy for Broken word Segmentation

Type of Words	Technique	Total Scanned Words	Segmented Words	Accuracy
Broken and Skewed	With single erosion and 9 pixel difference	1000	930	93%
Broken and Skewed	With double erosion and 5 pixel difference	1000	780	78%
Broken and Skewed	With double erosion and 9 pixel difference	1000	770	77%

9. Conclusion and Future Scope

Optical character recognition (OCR) is the process of converting scanned images of machine printed or handwritten text into a computer process able format. The practical importance of OCR applications, as well as the interesting nature of OCR problem, has led to great research interest and measurable advances in the field. But very limited research is reported on OCR of the scripts of Indian languages. There are a few research papers on Gurmukhi script on handwritten text documents for character segmentation reported till now. Here, in this work, a set of very simple and easy to compute features is used and a hybrid classification scheme is employed based on

Horizontal profile projection and vertical profile projection techniques and by using the neighbouring pixels method we are able to segment the broken characters. Firstly we extract the line from which we want to segment the word into characters by accepting input co-ordinates of corresponding line which is to be extracted from the user. After that we checked that whether the word is broken or not and finally we segment the character using these techniques and results are very good and accuracy is 93 %. This work reported on Gurmukhi language script may be extended in several directions. Here, we discussed only segmentation of the skewed, simple and broken characters in Gurmukhi script. But this work can be extended to segmentation of touching characters in a word, or overlapping characters in a word or for characters with unequal stroke intensity in Gurmukhi script also. Furthermore, the work can be extended to various fonts for Punjabi language and then recognition of the characters can be a next task in OCR process.

10. References

[1] Brijmohan Singh, Nitin Gupta, RashiTyagi, Ankush Mittal and DebashishGhosh, "Parallel Implementation of Devnagari Text line and Word Segmentation Approach on GPU", *International Journal of Computer Applications*, Vol. 24, No. 9, pp 7-14, 2011.

[2] Dharamveer Sharma and Gurpreet Singh Lehal, "An Iterative Algorithm for Segmentation of Isolated Handwritten Words in Gurmukhi Script", *In Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, Vol. 2, pp. 1022-1025. IEEE, 2006

[3] Galaxy Bansal and Dharamveer Sharma, "Isolated Handwritten Words Segmentation Techniques in Gurmukhi Script", *International Journal of Computer Applications*, Vol. 1, No. 24, pp 104-111, 2010.

[4] Gurpreet Singh Lehal, "Optical Character Recognition of Gurmukhi Script using Multiple Classifier", *In Proceedings of the International Workshop on Multilingual OCR*, P.T, ACM, 2009.

[5] M. K. Jindal, G.S. Lehal and R. K. Sharma, "A Study of Touching Characters in Degraded Gurmukhi Text", *International Journal of computer, Information Science and Engineering*, Vol. 1, No.4, pp. 870-873, 2007.

[6] M.K. Jindal, R.K.Sharma and G.S.Lehal, "Segmentation of Touching Characters in Upper Zone in Printed Gurmukhi Script", *In proceedings of the 2nd Bangalore Annual Compute Conference*, p.9, ACM, 2009.

[7] Mr. Sandip N. Kamble and Prof. Mrs. Megha Kamble, "Morphological Approach for Segmentation of Scanned Handwritten Devnagari Text", *International Journal of Emerging trends in Engineering and Development*, Issue 1, Vol. 3, pp 99-108, 2011.

[8] Munish Kumar, M. K. Jindal and R. K. Sharma, "Segmentation of Isolated and Touching Characters in Offline Handwritten Gurmukhi Script Recognition", *I.J. Information*

Technology and Computer Science, Vol. 2, pp. 58-63, 2014.

[9] Rajiv Kumar and Amardeep Singh, "Detection and Segmentation of Lines and Words in Gurmukhi Handwritten Text", *2nd International Advance Computing Conference* IEEE, pp. 353-356, 2010

[10] U. Pal, A. Belaid and Ch. Choisy, " Touching Numeral Segmentation using Water Reservoir Concept", *Pattern Recognition Letters Elsevier Science B.V.*, pp. 261-272, 2002.

[11] Vikas J Dongre and Vijay H Mankar, "Devnagari Document Segmentation Using Histogram Approach", *International Journal of Computer Science, Engineering and Information Technology*, Vol. 1, No. 3, pp. 46-53, 2011.

[12] Vneeta Rani and Pankaj Kumar, "Problems of Character Segmentation in Handwritten Text Documents written in Devnagari Script", *International Journal of Advanced Research in Computer Engineering & Technology*, Vol. 2, Issue 3, pp. 1026-1029, 2013.