Segmentation of Printed Text in Devanagari Script and Gurmukhi Script

Vijay Kumar
ECD, Deptt.
IIT, Roorkee

Pankaj K. Sengar
ECD, Deptt.
IIT, Roorkee

ABSTRACT

In this paper, we describe the line, word, character and top character segmentation for printed Hindi text in Devanagari script. And also describe the line and word segmentation for printed text in Gurmukhi script. A performance of 100% at line level, approximately 100% at word level, 99% at character level, and 97% at top character level for Devanagari script and performance of 100% at line level and 99% at word level for Gurmukhi script is obtained. Here we have observed the performance of segmentation with the help of five documents in devanagari script and five document in gurmukhi script.

Categories and Subject Descriptors

I.7.5 [**Document and text processing**]: Document Capture – *Document analysis, Graphics recognition and interpretation, Optical character recognition (OCR), Scanning.*

General Terms

Segmentation, Algorithms, Performance, accuracy and pixel classification.

Keywords

Optical character recognition, Character Segmentation, Middle Zone, Upper Zone, Lower Zone, line, word, character, top character.

1. INTRODUCTION

In optical character recognition (OCR), a perfect segmentation of characters is required before individual characters are recognized. A lot of research work has been investigated for character recognition of Indian scripts. For an OCR system, segmentation phase is an important phase and accuracy of any OCR heavily depends upon segmentation phase. Segmentation subdivides an image into its constituent regions or objects. Basically in segmentation, we try to extract basic constituent of the script, which are certainly characters. This is needed because the classifier recognizes these characters only [14]. Segmentation phase is also crucial in contributing to this error due to touching characters, which the classifier cannot properly tackle. Even in good quality documents, some adjacent characters touch each other due to inappropriate scanning resolution. So Incorrect segmentation leads to incorrect recognition. Segmentation phase includes line, word and character segmentation. Before word and character segmentation, line segmentation is performed to find the number of lines and boundaries of each line in any input document image. Incorrect line segmentation may result in decrease in recognition accuracy. In this paper, our work is related with segmentation of Devanagari and Gurumukhi script. And we obtained the better accuracy at line, word, character and top character level from previous result.

2. BACKGROUND

A survey on segmentation techniques, for machine printed text, can be found in references [1, 2]. Many works on Indian scripts OCR have been reported [3, 4, 5, 6, 7 and 11]. However, none of these works have considered real-life printed Hindi text in Devanagari consisting of character fusions and noisy environment. But very little work has been carried out for Indian scripts like Devnagari, Bengali, and Gurmukhi etc. Current research lags in the field of segmentation of machine printed. Devnagari [8, 9] and machine printed Gurmukhi [10-11] scripts.

3. CHARACTERISTICS

3.1 Characteristics of Devanagari Script

Devanagari is used in many Indian languages like Hindi, Nepali, Marathi, Sindhi etc. More than 300 million people around the world use Devanagari script. This script forms the foundation of Indian languages. So Devanagari script plays a very major role in the development of literature and manuscripts. Devanagari script has about 11 vowels and **33** consonants. We have illustrate the Characters and symbols of Devanagari script in Figure 2. And Devanagari word is written into the three strips namely: a core strip, a top strip, and a bottom strip as shown in figure 1. The core strip and top strip are differentiated by the header, while the lower modifier is attached to the core character.

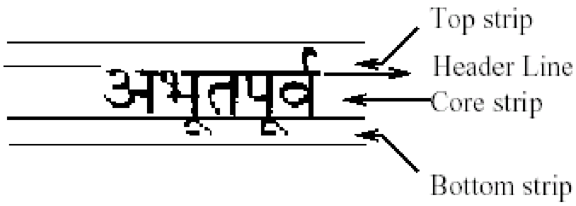


Figure 1: Three strips of Devanagari word

OCR for Devnagari script becomes even more difficult when compound character and modifier characteristics are combined in 'noisy' situations. The image below in Figure 3. Illustrates a Devanagari document with background noise. We can clearly see that compound characters and modifiers are difficult to detect in this image because the image background is not uniform in color, and marks are present that must be distinguished from characters.

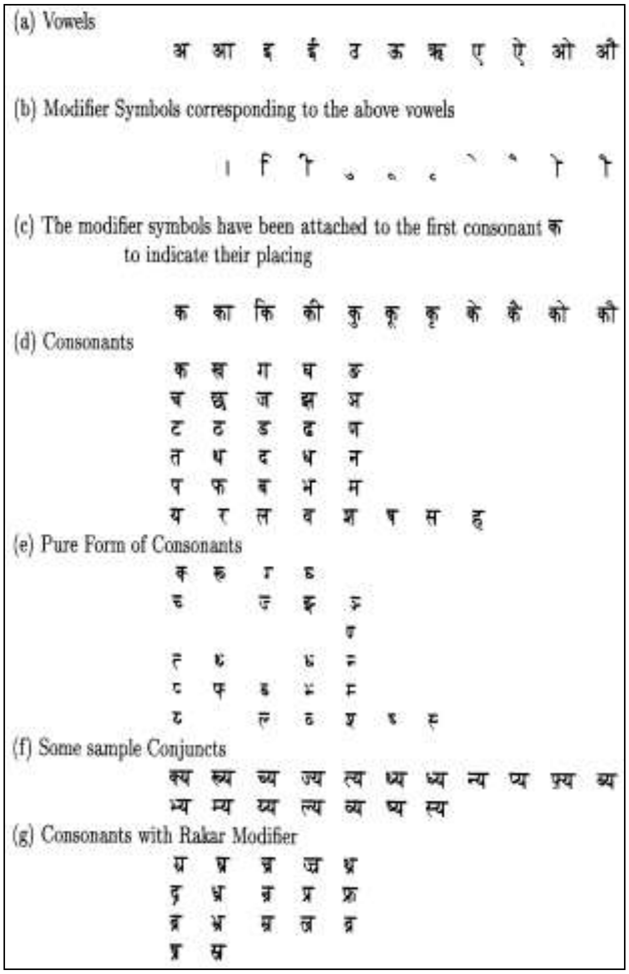

(a) Vowels
अ आ इ ई उ ऊ ऋ ए ऐ ओ औ
(b) Modifier Symbols corresponding to the above vowels
ा ि ी ु ू ृ े ै ो ौ
(c) The modifier symbols have been attached to the first consonant क to indicate their placing
क का कि की कु कू कृ के कै को कौ
(d) Consonants
क ख ग घ ङ
च छ ज झ ञ
ट ठ ड ढ ण
त थ द ध न
प फ ब भ म
य र ल व श ष स ह
(e) Pure Form of Consonants
(f) Some sample Conjuncts
क्य ख्य च्य ज्य त्य ध्य ध्य न्य प्य फ्य ब्य
भ्य म्य य्य ल्य व्य ष्य स्य
(g) Consonants with Rakar Modifier

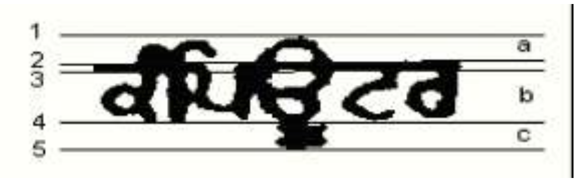
Figure 2: Characters and symbols of Devanagari script.

Figure 3: Images with background noise

3.2 Characteristics of Gurmukhi Script

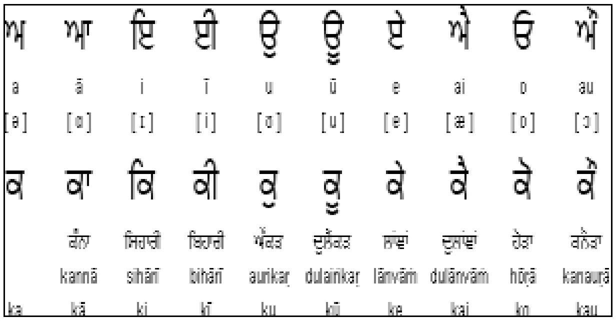
Gurmukhi script alphabet consists of 41 consonants and 12 vowels [12].Some characters in the form of half characters are present in the feet of characters. Writing style is from left to right. In Gurmukhi, There is no concept of upper or lowercase characters. A line of Gurmukhi script can be partitioned into three horizontal zones namely, upper zone, middle zone and lower zone. Consonants are generally present in the middle zone. These zones are shown in Figure 4. The upper and lower zones may contain parts of vowel modifiers and diacritical markers. [13]

Figure 4: a) Upper zone from line number 1 to 2, b) Middle Zone from line number 3 to 4, c) lower zone from line number 4 to 5.



In Gurmukhi Script, most of the characters, as shown in Figure 6, contain a horizontal line at the upper of the middle zone. This line is called the headline. The vowel and vowel diacritics, consonants and other symbols illustrate in Figure 5, 6, 7 respectively. The characters in a word are connected through the headline along with some symbols. The headline helps in the 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 Gurmukhi script, as shown in Figure 4, 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 does not work efficiently for printed Gurmukhi script.

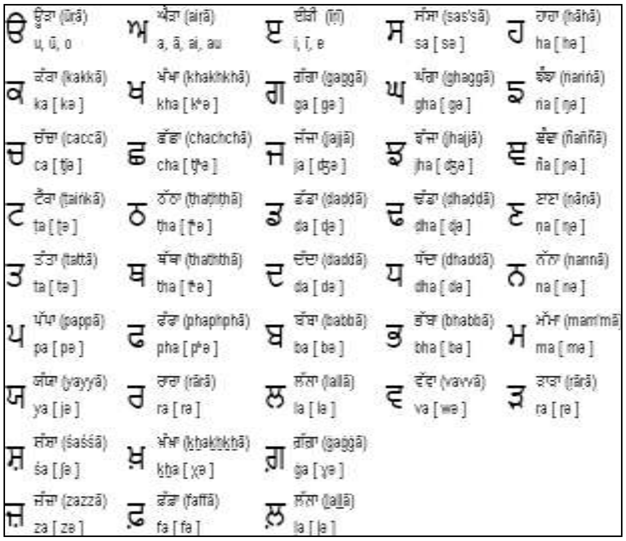
Vowels and Vowel diacritics (Laga Matra)



ਅ	ਆ	ਇ	ਈ	ਉ	ਊ	ਏ	ਐ	ਓ	ਔ
a	ā	i	ī	u	ū	e	ai	o	au
[ə]	[ɑ]	[ɪ]	[i]	[ʊ]	[u]	[e]	[æ]	[o]	[ɔ]
ਕ	ਕਾ	ਕਿ	ਕੀ	ਕੁ	ਕੂ	ਕੇ	ਕੈ	ਕੋ	ਕੌ
	ਕੰਨਾ	ਸਿਹਾਰੀ	ਬਿਹਾਰੀ	ਔਂਕੜ	ਦੁਲੈਂਕੜ	ਲਾਂਵਾਂ	ਦੁਲਾਂਵਾਂ	ਹੋੜਾ	ਕਨੌੜਾ
	kannā	sihārī	bihārī	aunkaṛ	dulainkaṛ	lānvām	dulānvām	hōṛā	kanauṛā
ka	kā	ki	kī	ku	kū	ke	kai	ko	kau

Figure 5: Vowels and Vowel diacritics (Laga Matra)

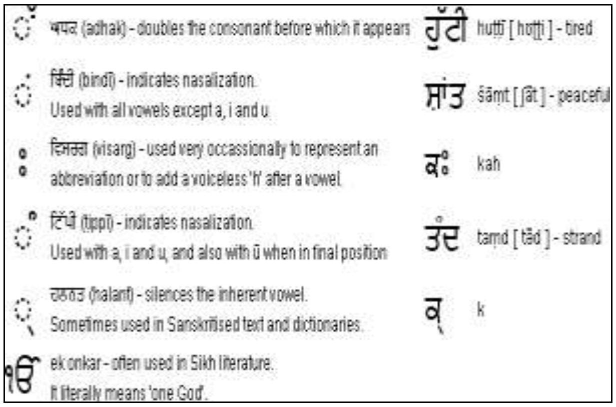
Consonants (Vyanjan)



ੳ ਊੜਾ (ūṛā) u, ū, o	ਅ ਐੜਾ (aiṛā) a, ā, ai, au	ੲ ਈੜੀ (īṛī) i, ī, e	ਸ ਸੱਸਾ (sas'sā) sa [sə]	ਹ ਹਾਹਾ (hāhā) ha [hə]
ਕ ਕੱਕਾ (kakkā) ka [kə]	ਖ ਖੱਖਾ (khakhkhā) kha [kʰə]	ਗ ਗੱਗਾ (gaggā) ga [gə]	ਘ ਘੱਗਾ (ghaggā) gha [gə]	ਙ ਙੰਙਾ (naṅṅā) ṅa [ŋə]
ਚ ਚੱਚਾ (caccā) ca [tʃə]	ਛ ਛੱਛਾ (chachchā) cha [tʃʰə]	ਜ ਜੱਜਾ (jajjā) ja [dʒə]	ਝ ਝੱਜਾ (jhajjā) jha [dʒə]	ਞ ਞੰਞਾ (ñaññā) ña [ɲə]
ਟ ਟੈਂਕਾ (ṭaiṅkā) ṭa [ʈə]	ਠ ਠੱਠਾ (ṭhaṭhṭhā) ṭha [ʈʰə]	ਡ ਡੱਡਾ (ḍaḍḍā) ḍa [ɖə]	ਢ ਢੱਡਾ (ḍhaḍḍā) ḍha [ɖə]	ਣ ਣਾਣਾ (ṇāṇā) ṇa [ɳə]
ਤ ਤੱਤਾ (tattā) ta [tə]	ਥ ਥੱਥਾ (thaththā) tha [tʰə]	ਦ ਦੱਦਾ (daddā) da [də]	ਧ ਧੱਦਾ (dhaddā) dha [də]	ਨ ਨੱਨਾ (nannā) na [nə]
ਪ ਪੱਪਾ (pappā) pa [pə]	ਫ ਫੱਫਾ (phaphphā) pha [pʰə]	ਬ ਬੱਬਾ (babbā) ba [bə]	ਭ ਭੱਬਾ (bhabbā) bha [bə]	ਮ ਮੱਮਾ (mam'mā) ma [mə]
ਯ ਯੱਯਾ (yayyā) ya [jə]	ਰ ਰਾਰਾ (rārā) ra [rə]	ਲ ਲੱਲਾ (lallā) la [lə]	ਵ ਵੱਵਾ (vavvā) va [wə]	ੜ ੜਾੜਾ (ṛāṛā) ṛa [ɽə]
ਸ਼ ਸ਼ੱਸ਼ਾ (śaśśā) śa [ʃə]	ਖ਼ ਖ਼ੱਖ਼ਾ (k̲h̲ak̲h̲k̲h̲ā) k̲h̲a [χə]	ਗ਼ ਗ਼ੱਗ਼ਾ (ġaġġā) ġa [ɣə]		
ਜ਼ ਜ਼ੱਜ਼ਾ (zazzā) za [zə]	ਫ਼ ਫ਼ੱਫ਼ਾ (faffā) fa [fə]	ਲ਼ ਲ਼ੱਲ਼ਾ (l̲al̲l̲ā) l̲a [ɭə]		

Figure 6: Consonants (Vyanjan)

Other symbols



◌ੱ	ਅਧਕ (adhak) - doubles the consonant before which it appears	ਹੁੱਟੀ	huṭṭī [hoʈʈi] - tired
◌ਂ	ਬਿੰਦੀ (bindī) - indicates nasalization. Used with all vowels except a, i and u	ਸ਼ਾਂਤ	śāṃt [ʃãt] - peaceful
ਃ	ਵਿਸਰਗ (visarg) - used very occassionally to represent an abbreviation or to add a voiceless 'h' after a vowel	ਕਃ	kah
◌ੰ	ਟਿੱਪੀ (ṭippī) - indicates nasalization. Used with a, i and u, and also with ū when in final position	ਤੰਦ	taṃd [tə̃d] - strand
◌੍	ਹਲੰਤ (halant) - silences the inherent vowel. Sometimes used in Sanskritised text and dictionaries.	ਕ੍	k
ੴ	ek onkar - often used in Sikh literature. It literally means 'one God'.		

Figure 7: Other symbols

4. IMAGE CATEGORIES AND PRE-PROCESSING

The various categories of the images that could be fed as an input which is in three categories. Binary level images, pseudo color and true color images. For a binary level image, the preprocessing required is minimal. There are two colors, a foreground and a background color. The text is usually represented in the foreground. So we would need to look for the foreground components and perform the analysis.

The next category of images is the pseudo color images. The best example of the pseudo color images are GIFs. It makes use of only 256 different colors. The True color images make use of 16M colors. In this paper we use GIF images, TIFF image, JPEG image and JPG images.

5. SEGMENTATION FOR DEVANAGARI SCRIPT AND GURMUKHI SCRIPT

Segmentation is a classifier which helps to fragment each character from a word present in a given image / page. The objective of the segmentation is to extract each character from the text present in the image. While this algorithm is proposed by Veena Bansal [8] but we have done some modification in algorithm of Preliminary segmentation of words. Hence we obtained better result for segmentation of Devanagari script and Gurmukhi script. Here we have considered bottom strip with core strip. The process of segmentation mainly follows the following pattern:

. First, it identifies the page layout,

. After that, it identifies the line from the page,

. Identifies the word from that line, and

. Finally, identifies the character from that word.

5.1 Proposed Algorithm for Segmentation of Devanagari Script and Gurmukhi Script

Step 1: Line Segmentation

The global horizontal projection method computes sum of all black pixels on every row and constructs corresponding histogram. Based on the peak/valley points of the histogram, individual lines are separated. In line segmentation our aim is to draw of one upper horizontal line and one lower horizontal line for each line of text image. The steps for line segmentation are as follow:

. Construct the Horizontal Histogram for the image

. Using the Histogram, find the points from which the line starts and ends.

. For a line of text, upper line is drawn at a point where we start finding black pixels and lower line is drawn where we start finding absence of black pixels. And the process continues for next line and so on.

Step 2: Word Segmentation

. Construct the vertical histogram for each segmented line

. Using the vertical Histogram, find the points from which the word starts and ends.

.Vertical lines are drawn at starting and ending points for each word.

Step 3: Character Segmentation

. Draw the horizontal histogram for each segmented line

. From the horizontal histogram, find the row which consists of maximum value.

. The row which consists of maximum value of black pixel for each line is actually the row which consists of Header line.

. Draw the vertical histogram for each segmented word in below of header line.

. Draw the vertical histogram for each segmented word in above of header line

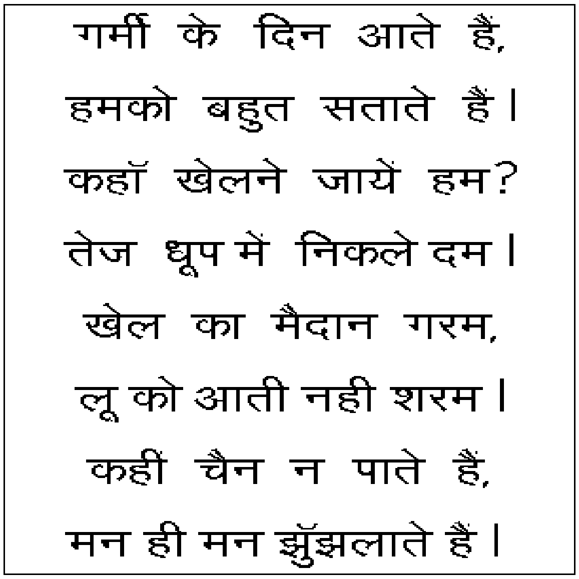
. Using the histogram, find the points from which the character starts and ends.

. Draw line according these coordinate.

Step 4: Maintain the data structure to feed the line, word and character boundaries such that the character boundary could be sufficiently extracted from the image which is required for the further training and recognition portion of the system.

6. EXPERIMENTAL RESULTS AND DISCUSSIONS

We have collected 10 printed documents, which is document-1 to document-5 in Devanagari Script and from document-6 to document-10 in Gurmukhi Script. We have show document-3 in figure-8 and we illustrate the number of lines and each line contains the number of words, characters and top characters with the help of Table-1.



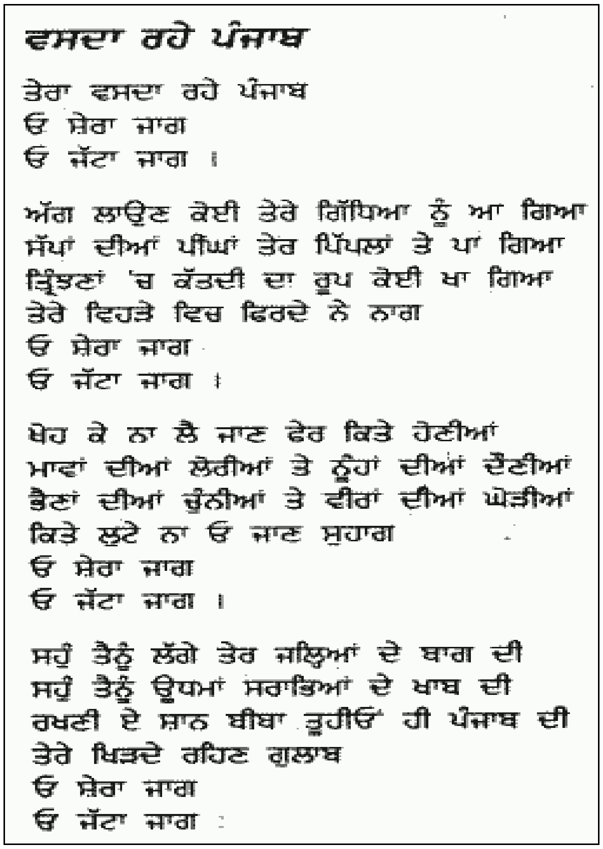
गर्मी के दिन आते हैं,
हमको बहुत सताते हैं।
कहॉ खेलने जायें हम?
तेज धूप में निकले दम।
खेल का मैदान गरम,
लू को आती नही शरम।
कहीं चैन न पाते हैं,
मन ही मन झुँझलाते हैं।

Figure 8: Document-3 in Devanagari Script.

We have prepared the Table-1, which is show the accuracy of word, character and top character segmentation for document-3 in devanagari script by Figure 8. Which is also illustrate the recognize words, characters, and top characters with respect of original words, original characters and original top characters respectively for each line of document-3.

Table 1. Result of word, Character and top character Segmentation of Document-3 in Devanagri script

Line no.	No. of words		No. of char.		No. of top char	
	Original words	recog words	Original char	Recog char	Original top char	Recog top char
0	5	5	13	12	5	5
1	5	5	13	13	4	4
2	5	5	12	12	5	5
3	6	6	12	12	6	5
4	4	4	13	13	2	2
5	6	6	15	15	4	4
6	5	5	11	11	4	4
7	6	6	13	13	5	5
Total	42	42	102	101	35	34
Accur	100%		99%		97%	



ਵਸਦਾ ਰਹੇ ਪੰਜਾਬ

ਤੇਰਾ ਵਸਦਾ ਰਹੇ ਪੰਜਾਬ
ਓ ਸ਼ੇਰਾ ਜਾਗ
ਓ ਜੱਟਾ ਜਾਗ ।

ਅੱਗ ਲਾਉਣ ਕੋਈ ਤੇਰੇ ਗਿੱਧਿਆ ਨੂੰ ਆ ਗਿਆ
ਸੱਪਾਂ ਦੀਆਂ ਪੀਘਾਂ ਤੇਰ ਪਿੱਪਲਾਂ ਤੇ ਪਾ ਗਿਆ
ਤ੍ਰਿੰਝਣਾਂ 'ਚ ਕੱਤਦੀ ਦਾ ਰੂਪ ਕੋਈ ਖਾ ਗਿਆ
ਤੇਰੇ ਵਿਹੜੇ ਵਿਚ ਫਿਰਦੇ ਨੇ ਨਾਗ
ਓ ਸ਼ੇਰਾ ਜਾਗ
ਓ ਜੱਟਾ ਜਾਗ ।

ਖੋਹ ਕੇ ਨਾ ਲੈ ਜਾਣ ਫੇਰ ਕਿਤੇ ਹੋਣੀਆਂ
ਮਾਵਾਂ ਦੀਆਂ ਲੋਰੀਆਂ ਤੇ ਨੂੰਹਾਂ ਦੀਆਂ ਦੌਣੀਆਂ
ਭੈਣਾਂ ਦੀਆਂ ਚੁੰਨੀਆਂ ਤੇ ਵੀਰਾਂ ਦੀਆਂ ਘੋੜੀਆਂ
ਕਿਤੇ ਲੁਟੇ ਨਾ ਓ ਜਾਣ ਸੁਹਾਗ
ਓ ਸ਼ੇਰਾ ਜਾਗ
ਓ ਜੱਟਾ ਜਾਗ ।

ਸਹੁੰ ਤੈਨੂੰ ਲੱਗੇ ਤੇਰ ਜਲ੍ਹਿਆਂ ਦੇ ਬਾਗ ਦੀ
ਸਹੁੰ ਤੈਨੂੰ ਊਧਮਾਂ ਸਰਾਭਿਆਂ ਦੇ ਖਾਬ ਦੀ
ਰਖਣੀ ਏ ਸ਼ਾਨ ਬੀਬਾ ਤੂਹੀਓਂ ਹੀ ਪੰਜਾਬ ਦੀ
ਤੇਰੇ ਖਿੜਦੇ ਰਹਿਣ ਗੁਲਾਬ
ਓ ਸ਼ੇਰਾ ਜਾਗ
ਓ ਜੱਟਾ ਜਾਗ :

Figure 9: Document-6 in Gurmukhi Script

Table 2. Result of Line Segmentation by Proposed

Document	No. of line	Correct Detected	Incorrect Segmentation	Accuracy
Document 1	13	13	0	100%
Document 2	9	9	0	100%
Document 3	8	8	0	100%
Document 4	15	15	0	100%
Document 5	12	12	0	100%
Document 6	22	22	0	100%
Document 7	24	24	0	100%
Document 8	20	20	0	100%
Document 9	17	17	0	100%
Document 10	16	16	0	100%

Table 3. Result of Word Segmentation by Proposed

Document	No. of words	Correct Detected	Incorrect segmentation	Accuracy
Document 1	68	68	0	100%
Document 2	118	117	2	99%
Document 3	42	42	0	100%
Document 4	90	90	0	100%
Document 5	87	87	0	100%
Document 6	120	120	0	100%
Document 7	104	104	0	100%
Document 8	98	97	1	99%
Document 9	56	56	0	100%
Document 10	103	102	1	99%

We have show the document-6 in Gurmukhi script by figure-9 and also we have illustrate the result of line segmentation and word segmentation for different document by Table-2 and Table-3 respectively for both Devanagari script and Gurmukhi script. And for Devanagari script we have illustrate the result of word, Characters and Top Characters Segmentation for five documents by Table-4. And obtained the accuracy 99.75% at word segmentation level, 98.89% at character segmentation level and 97.40% at top character segmentation level. This is better than previous results.

Table 4. Result of word, Characters and Top Characters Segmentation for Devanagari Script Document

Document	No. of words		No. of Characters		No. of Top Characters	
	No. of original	No. of recognize	No. of original	No. of recognize	No. of original Top	No. of recognize Top
Document-1	68	68	182	180	59	58
Document-2	118	117	262	259	102	99
Document-3	42	42	102	101	35	34
Document-4	90	90	215	213	79	77
Document-5	87	87	228	225	72	70
Total	405	404	989	978	347	338
Accuracy	99.75%		98.89%		97.40%	

7. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a modified algorithm for segmentation of line, word, character, top character for Devanagari Script and Segmentation of line, word for Gurmukhi Script. The overall successful segmentation achieved through the proposed algorithm is better than previous result. Since at few point segmentation was good but at few point it was not up to the expectations. This may be because of the shape of characters. All these issues can be dealt in the future for printed documents in Devanagari and Gurumukhi script by making few changes to proposed work.

8. 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] S. Antani and L. Agnihotri, *Gujrati Character Recognition, in* Proceedings of the international conference on Document Analysis and Recognition (ICDAR- 99), Bangalore, India, pp. 418-421, 1999.

[4] Veena Bansal and R.M.K. Sinha, *Partitioning and Searching Dictionary for Correction* of *Optically-Read Devanagari Character Strings,* in Proceedings - Fifth International Conference on Document Analysis and Recognition, IEEE Publication, held at Bangalore from Sep21-23, 1999, pp. 410-413.

[5] *S.* S. Marwah, S. K. Mullick and R. M. K. Sinha, “Recognition of Devanagari characters using a hierarchical binary decision tree classifier”, IEEE International Conference on Systems, Man and Cybernetics, October 1994.

[6] R.M.K.Sinha, “Rule based contextual post processing for Devanagari text recognition”, Pattern Recognition, 20(5), pp. 475-485, 1987.

[7] I. K. Sethi, “Machine recognition of constrained hand printed Devanagari”, Pattern Recognition, vol. 9, pp. 69-75, 1977.

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

[9] 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.

[10] 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.

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

[12] M. K. Jindal, G. S. Lehal and R. K. Sharma. “Segmentation Problems and Solutions in Printed Degraded Gurmuk hi Script”. IJSP, Vol 2(4), 2005: ISSN 1304-4494.

[13] Rajiv K. Sharma & Dr. Amardeep Singh" Segmentation of Handwritten Text in Gurmukhi Script", International Journal of Computer Science and Security, volume (2) issue (3), 2006

[14] Raghuraj Singh. S. Yadav and Prabhat Verma” Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network” , International Journal of Computer Science & Communication, Vol. 1, No. 1, January-June 2010, pp. 91-95