The Hazards in Segmentation of Handwritten Hindi Text

Naresh Kumar Garg
GZS Collage of Engineering & Tech.
Bathinda, Punjab, India

Lakhwinder Kaur
Dept. of Computer Engineering, UCOE, Punjabi University, Patialas, Punjab, India

M. K. Jindal
Panjab University Regional Centre, Muktsar, Punjab, India

ABSTRACT
Optical Character Recognition (OCR) is a process to recognize the handwritten or printed scanned text with the help of a computer. Segmentation is very important stage of any text recognition system. The problems in segmentation can lead to decrease in segmentation rate and hence recognition rate. A good segmentation technique can improve the recognition rate. This paper deals with the hazards that occur in segmentation of handwritten Hindi text. We also explained the main reasons for some of these problems.

Keywords
Optical Character Recognition, Segmentation, Hazards in segmentation

1. INTRODUCTION
Hindi is the official language of India. The different writing styles, different sizes of characters and different shapes of characters in texts written by different people makes the job of segmentation very challenging. The technique used to segment the printed characters cannot be applied to handwritten documents due to variability in text written by different people. The problems in segmentation depend upon the text written by a writer. A good or clearly written text has fewer problems in segmentation as compared to badly written text.

2. RELATED WORK
A good survey about OCR is given in [1]. To the best of knowledge, no commercial OCR for handwritten Hindi text is available yet. Some papers dealing with line segmentation are referenced as [2-5]. The papers dealing with segmentation of overlapping lines are referenced as [6-7]. Many algorithms have been developed for segmenting touching characters in Indian scripts, but most of them are on printed text. Bansal and Sinha [8] have segmented the conjuncts (type of touching characters) based on structural properties of the text in printed Devanagari script. They segmented the conjuncts with an accuracy of 84%. Jindal et al. [9, 10] have segmented the touching characters in middle zone and upper zone of printed Gurmukhi script using structural properties of the script. Chaudhuri et al. [11] have used the principal of water overflow from a reservoir to segment touching characters in Oriya script. Garain and Chaudhuri [12] have used a technique based on fuzzy multifactorial analysis to segment touching characters in printed Devnagari and Bangla scripts.

The work on line segmentation, consonant segmentation, upper modifier segmentation and lower modifier segmentation and Half character segmentation in Handwritten Hindi text are explained by us in [13, 14, 15]. In this paper, we have explained some of the irregularities that generally occurs in a handwritten Hindi text and create problems in segmentation.

3. DATABASE
We analyze the data written by different writers. Some writers were asked to write paragraph of 10-15 lines of same text and some writers were asked to write different text.

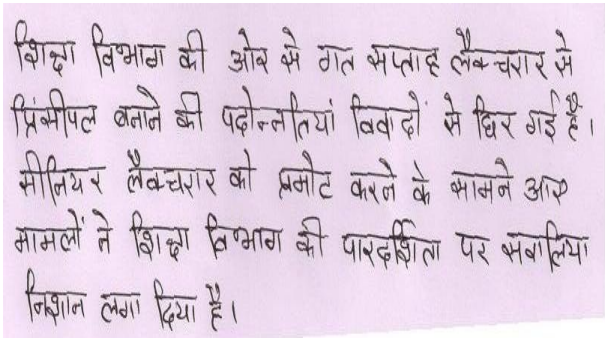

Fig 1a: Part of Database

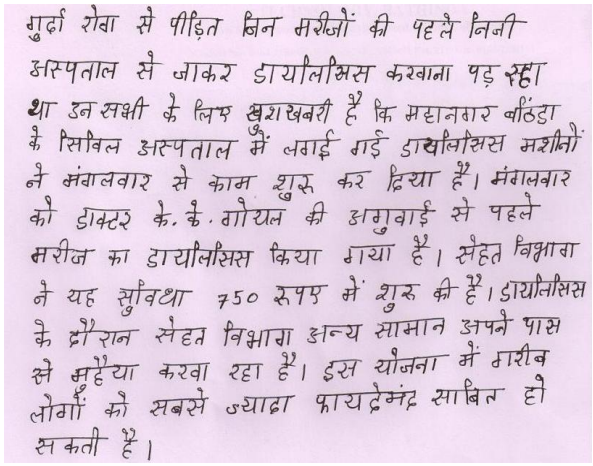

Fig 1b: Part of Database

4. CHARACTERISTICS OF HINDI LANGUAGE

Devanagari is the script for writing Hindi language. Hindi is written from left to right and there is no concept of upper or lower case. In Hindi language, most of the characters have a horizontal line at the upper part. When two or more characters are combined to form a word, the horizontal lines touch each other and generate a header line called *shirorekha*. The half characters may touch with full characters to make the characters called conjuncts. In each conjunct character, the right part is a full consonant, and the left part is always a half consonant. The vowels (modifiers) can be placed at the left, right (or both), top or bottom of the consonant. The Hindi word is divided into three regions-upper region, middle region and lower region. The upper and lower region includes vowels and middle region includes consonants.

5. PPROBLEMS IN SEGMENTATION

There are many types of problems in handwritten text. The badly written text can lead to decrease in segmentation rate and hence recognition rate.

The irregularities in handwritten text can be divided into two categories:

1) The irregularities that can be avoided
2) The irregularities which can not be avoided.

Some of the irregularities in the text can not be avoided due to writer's natural way of writing the text

The problems related with writer's natural handwriting i.e. the way of writing different characters creates irregularities in data which are difficult to overcome. This leads to decrease in recognition rate.

The irregularities that can be avoided occur due to bad quality of material, bad scanning and most important factor is speed of writing. If a writer uses the gel pen for writing the text then chances are more for touching characters as compared to thin tip pall point pen.

The bad quality of material like paper and pen create fewer irregularities as compared to irregularities created by speed by writing the text.

The major number irregularities in same text written by a single writer in different situations occur due to his natural handwriting and speed of writing. The irregularities due to speed of writing the text cane be avoided.

Now we discuss some of the practical problems.

Problems in handwritten text can be divided into three categories:

1) Problems in Line Segmentation
2) Problems in Word Segmentation
3) Problems in character Segmentation

5.1 Problems in Line Segmentation

The problems in line segmentation occur due to following reasons:

5.1.1 The lower modifier of one line overlaps with the upper modifiers of lower line

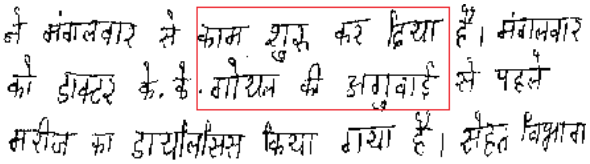

Fig 2: Overlapped lines

In figure 2, upper modifier of lower line overlaps with lower modifier of upper line. Due to overlapping of pixels of two lines it is not possible to segment the two lines with horizontal projection technique.

5.1.2 Slant in the lines of the text and slant in words of the same line. This creates curvature in the lines

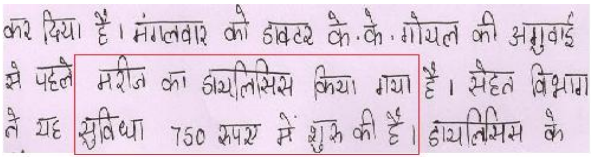

Fig 3: Curved lines

Due to curvature in the lines as shown in figure 3, it is very difficult to determine the proper header line. In such cases the segmentation of two lines is very challenging.

5.1.3 The presence of multiple header lines or non uniform header line

Due to thickness of pen tip or overwriting (cutting) by the writer the multiple header lines are formed over certain words as shown in figure 4, which creates problem in determination of proper header line.

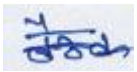

Fig 4: Multiple header lines

5.1.4 Absence of header line in the text

Some of the writer's do not form a header line on the words while writing the text as shown in figure 5. Due to absence of header line over the words it is very difficult to separate the lines.

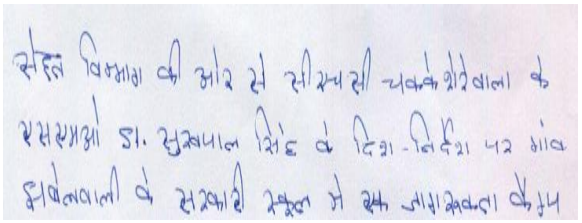

Fig 5: Absence of header line

5.2 Problems in Word Segmentation

The problems in word segmentation are very less. Some problems occur due to improper header line. Some times writer does not form a header line over some of the characters of a single word as shown in figure 6. So it leads to over segmentation of words.

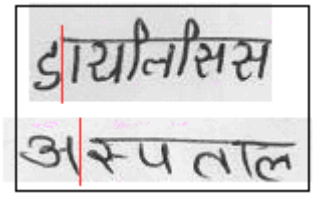

Fig 6: Absence of header line over certain characters of the word

5.3 Problems in Character Segmentation

The maximum number of problems occurs in character segmentation. The problems in character segmentation ca be further divided into following categories:

5.3.1 Problems in upper region
5.3.2 Problems in lower region
5.3.3 Problems in middle region

5.3.1 Problems in upper region

The problems in upper region can be further divided into two categories:

5.3.1.1 Problems due to unusual size of upper modifiers

Fig 7: large size of upper modifier

Due to large size of upper modifier as shown in figure 7, the determination of position of header line in a word is very difficult. It results in non segmentation of upper modifier from the consonant.

5.3.1.2 Touching of upper modifier like Bindi with header line or another upper modifier

In some words upper modifier merges with the header line or with another upper modifier. It is very difficult to segment these types of modifiers from the header line.

5.3.1.3 Touching of two or more modifiers

In figure 8, two modifiers in upper region touch each other. Such cases are very rare and are very difficult to segment..

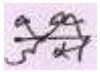

Fig 8: Touching of two modifiers

5.3.2 Problems in lower region

The problems in lower region can be further divided into following categories:

5.3.3.1 Determination of presence of lower modifier in a word

Due to variation in heights of different characters in a word it is very difficult to determine the presence of lower modifier in the word. To determine the presence of lower modifier in a word is very difficult task.

5.3.2.2 Problems due to unusual size of lower modifiers

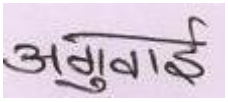

Fig 9: large size of lower modifier

Due to large size of lower modifier as shown in figure 9, the two consonants overlap.

5.3.2.3 Problems due to merging of lower modifier with consonant in middle region

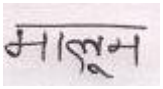

Fig 10: Merging of lower modifier with consonant

In figure 10, the lower modifier merges with character 'ल'. Due to merging of lower modifier with the character it is very difficult to determine the presence of lower modifier in a word.

5.3.2.4 Presence of lower modifier like features in some characters

\

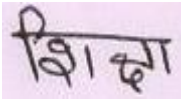

Fig 11: Characters with lower modifier like features

In figure 11, the character 'sh' and the character 'ksh' have lower modifier like features. They have loop in lower part which is similar to lower modifier.

5.3.3 Problems in middle region

The problems in middle region can be divided into following categories:

5.3.3.1 Touching of characters in middle region
5.3.3.2 Overlapping of characters in middle region
5.3.3.3 Broken Characters
5.3.3.4 Header line merges with consonants

5.3.3.1 The problem of touching characters can be further divided into three parts:

i) Touching of modifier with consonants in middle region.

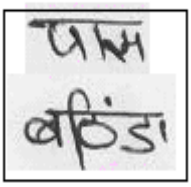

Fig 12: Modifier touches with consonant in middle region

The problem of touching the left modifier with the consonant generally occurs in many of the written documents. In figure 12, left modifier '◌ा matra' touches with character 'स' and left modifier also touches with character 'ड'.

ii) Touching of two or more consonants in middle region.

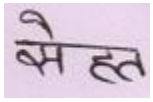

Fig 13: Two consonants touch each other in middle region

In figure 13, two consonants touch each other ie. Character 'ह' touches with character 'त'. But it is very difficult to determine the presence of two or more touching consonants in a word.

iii) Touching of half character with full character (conjuncts).

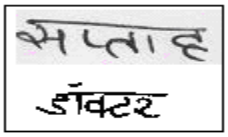

Fig 14: Half character touches the full character in middle region

The presence of half character touching full character makes the problem of segmentation of handwritten Hindi text very complex. In figure 14, half character 'प' touches the full character 'त' and half character 'क' touches the full character 'ट'. The above problem can be solved easily if we are able to determine the presence of conjunct in a word. The determination of presence of conjunct in a word is very challenging task

5.3.3.2 In figure 15, character 'ह' overlaps with half character 'न' and character 'ह' also overlaps with character 'त'. These types of characters are difficult to segment by vertical projection. This type of problem mostly occurs with no vertical bar characters.

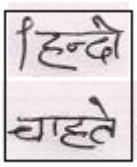

Fig 15: Overlapped characters

5.3.3.3 Some characters are difficult to write completely without lifting the hand at least once like characters 'अ','स' etc. In such cases sometimes space left with in a character ie. some pixels are missing which divides the character into two or more parts In fig 16, character 'अ' and character 'स' have some missing pixels which breaks the character into two parts.. This is very common problem in handwritten documents and it is very difficult to solve. It is a over segmentation problem and not segmentation problem. It can be solved during recognition.

Fig 16: Broken characters

5.3.3.4 Sometimes header line merges with characters in some words. In figure 17, header line merges with all the characters in a word. It is very difficult to determine the header line and hence difficult to segment.

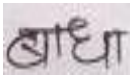

Fig 17: Header line merges with consonants

6. DISCUSSION

The problems explained above are very useful for complete segmentation of handwritten Hindi text. Some problems can be removed if writer uses the better material and write patiently. To solve the problems related with writer's natural handwriting efficient algorithms are to be designed to segment the text. The study may be carried out in future in the following direction:

1. The efforts should be made to solve the above problems. It is very difficult to determine the presence of lower modifier and to determine the presence of conjucts in middle region of the word.
2. The writer's can be given instructions to write patiently to minimize some of the problems of segmentation like touching characters.
3. The algorithms used in other Indian scripts for similar problems can be tried on handwritten Hindi text.

7. REFERENCES

[1] Mori, S., Suen, C. Y. and Yamamoto, K. 1992. Historical review of OCR Research and development. In Proceedings of the IEEE, Vol. 80, No. 7, pp. 1029-1058.

[2] Zahour, A., Taconet, B., Mercy, P., and Ramdane, S. 2001. Arabic Hand-written Text-line Extraction. In Proceedings of the Sixth International Conference on Document Analysis and Recognition, ICDAR, pp. 281–285.

[3] Tripathy, N. and Pal.,U. 2004. Handwriting Segmentation of unconstrained Oriya Text. In International Workshop on Frontiers in Handwriting Recognition, pp. 306–311.

[4] Louloudis, G., Gatos, B., Pratikakis I. and Halatsis K. 2006. A Block Based Hough Transform Mapping for Text Line Detection in Handwritten Documents. In Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition, pp.515-520.

[5] Li, Y., Zheng, Y., Doermann, D. and Jaeger, S. 2006. A new algorithm for detecting text line in handwritten documents. In Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition, pp. 35–40.

[6] Zahour, A., Taconet, B., Likforman-Sulem, L. and Wafa Boussellaa. 2008. Overlapping and multi-touching text line segmentation by Block Covering analysis. Pattern Analysis and Applications, Vol. 12, pp. 335-351,.

[7] Jindal, M. K., Sharma, R. K. and Lehal, G. S. 2007. Segmentation of Horizontally Overlapping Lines in Printed Indian Scripts. In International Journal of Computational Intelligence Research (IJCIR), Research India Publications, Vol. 3, No. 4, pp. 277-286.

[8] Bansal, Veena. 1999. Integrating knowledge sources in Devanagari text recognition. Ph.D. thesis, IIT Kanpur, INDIA.

[9] Jindal, M. K., Lehal, G. S. and Sharma, R. K. 2009. On Segmentation of touching characters and overlapping lines in degraded printed Gurmukhi script. In International Journal of Image and Graphics (IJIG), World Scientific Publishing Company, Vol. 9, No. 3, pp. 321-353.

[10] Jindal, M. K., Sharma, R. K. and Lehal, G. S. 2009. Segmentation of Touching Characters in Upper Zone in printed Gurmukhi Script. In Proceedings of the 2nd Bangalore Annual Compute Conference, Banglore, ACM, No. 9.

[11] Chaudhuri, B. B., Pal, U. and Mitra, M. 2001. Automatic recognition of printed oriya Script. In International Conference on Document Analysis and Recognition, pp. 795–799.

[12] Garain U. and Chaudhuri, B. B. 2002. Segmentation of touching characters in printed Devnagari and Bangla scripts using fuzzy Multifactorial Analysis. IEEE Transactions on Systems, Man and Cybernetics. *Part C, Vol. 4, No.* 32, pp. 449–459.

[13] Garg, Naresh Kumar, Kaur, Lakhwinder and Jindal, M. K. 2010. Segmentation of Handwritten Hindi Text. In International Journal of Computer Applications (IJCA), Vol. 1, No. 4, pp. 22-26.

[14] Garg, Naresh Kumar, Kaur, Lakhwinder and Jindal, M. K. 2010. A new method for line segmentation of Handwritten Hindi Text. In Proceedings of the IEEE 7th International Conference on Information Technology: New Generations (ITNG 2010), pp.392-397.

[15] Garg, Naresh Kumar, Kaur, Lakhwinder and Jindal, M. K. 2011. Half character segmentation of Handwritten Hindi Text. In Proceedings of ICISIL2011, pp.48-53.