ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 3, March 2013

Problems of character segmentation in Handwritten Text Documents written in Devnagari Script

Vneeta Rani, Pankaj Kumar

Abstract— **Character segmentation is a process of dividing a word from a text document. Document from which the words are to be used may be handwritten or printed text. In this paper, prime focus is on the problems which may occur during the segmentation of the character. This paper is divided into 5 sections. Section 3 will be focus on the major problems that may occur during the character segmentation process.**

***Index Terms*— Devnagari Script, Character segmentation**

I. INTRODUCTION

Character segmentation is very useful in many fields like automatic conversion of handwritten text into its equivalent machine format so that it can be edited by using a text editor. Character segmentation is very important part of OCR. After segmentation of the character features can be extracted from the segmented character. Accuracy in extracting the features is highly depends upon the segmented character. If character whose features are to extracted is not segment properly, can't be recognized accurately by the feature extraction algorithm. Character segmentation of the handwritten text documents written in the Devnagari script is dependent on the writing style of individual. Segmentation of characters is quite easy in case of printed documents as compared to the handwritten documents. Characters can be segmented written in the Devnagri script by the proposed technique as described in the following steps:

Step 1: calculate the frequency of the number of 1s in each row.

Step 2: With the help of frequencies of the 1's in each line horizontally determine a header line of the word

Step 3: remove that header line from the word

Step 4: segment the word by using vertical profile projection technique.

The horizontal line is present at upper part in Hindi language i.e. called Shirorekha. A Hindi is written as left to write. All these variations make devnagari Optical character recognition is a challenging problem. A Hindi text is divided into various zones as following as:

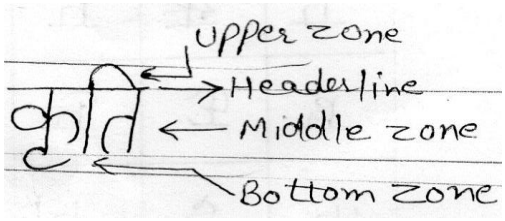

Fig1: Three zones of devnagari word

The header line is the most visible and the header line we can obtain the middle zone part of a word. We have separated the header line we segment the upper modifiers and lower modifiers of a word. After then we can check the touching and overlapped character or not. If it is a touching character then we segment that word. The above proposed method suffered from various problems line problem of broken characters, problem of overlapping characters, problem of touching characters, problem of skewed characters, problem of irregular intensity with the character, problem of detecting the header line. These are the various problems which may occur while segmenting a character from a handwritten text document. These problems will be explained in section 3 later in this paper.

***Vneeta Rani**, Guru Kashi University, Talwandi Sabo (Bathinda) Bathinda, India, Mobile No: 9463436124,*
***Pankaj Kumar**, Guru Kashi University, Talwandi Sabo (Bathinda) Bathinda, India, Mobile No: 9463436124,*

II. LITERATURE REVIEW

A good research about problems of segmentation is given in [1]. The best processes on segmentation of touching character in devnagari script are referenced as [2]. The main object of this paper the new methods for line segmentation and character segmentation of overlapping characters of Handwritten Hindi text. Then find the Algorithm of header lines and base lines by estimating the average line height is referenced as [3]. There are segmented of handwritten word of four different scripts namely; Bangla, Devanagri, Gurmukhi and Syloti are considered in [4]. The system deals with segmentation of modifier and fused characters in handwritten Devnagari word. Segmentation of modifiers consists of segmentation of top as well as bottom modifiers are given in [5]. This paper deals with a new set of features for recognition of Devnagari characters as shown as [6]. Some problems of segmentation is compounded by the possible presence of modifiers (matras) on all sides of the basic characters and due to the uncertainty introduced in the character shapes by way of different writing styles. The segmentation accuracy has been found to be around 75% are referenced as [7]. This paper deals with the new segmentation technique based on structure approach for

Handwritten devnagari script. The errors in segmentation propagate to recognition. The performance is evaluated on handwritten data of 1380 words of 200 lines written by 15 different writers shown as [8]. In this paper, we have explained some problems of character segmentation that occurs in Handwritten Hindi text.

III. CHARACTER SEGMENTATION PROBLEMS

There are various problem can be occur in character segmentation because all characters has not fixed size & shapes in handwritten document. The problems in character segmentation can be divided into a variety of categories as following as:

A. Problem of broken characters
B. Problem of overlapped characters
C. Problem of Touching characters
D. Problem of Skewed characters
E. Problem of irregular intensity with the character

A. PROBLEM OF BROKEN CHARACTER

Broken character problem may arise due to improper functioning of writing element e.g. some times while writing, the pen stops working properly in between the words. This leads to the formation of broken character Image is as shown below:-

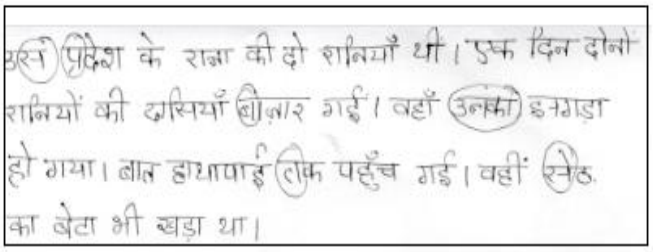

Fig 2: part of database

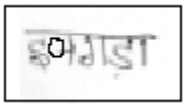

Fig3: Broken character

B. PROBLEM OF OVERLAPPED CHARACTER

This problem arises due to different writing styles of different people. In this problem one character is written above on the other characters mistakenly. This is termed as overlapping of characters.

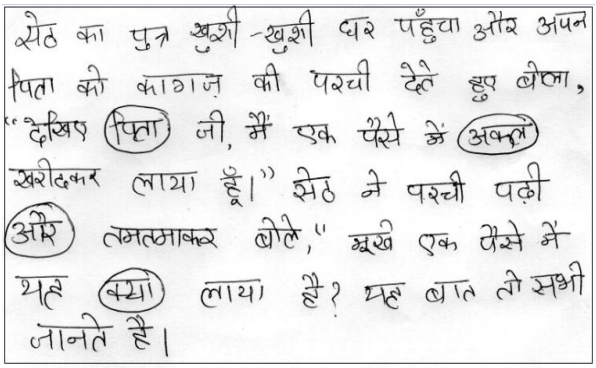

Fig 4: part of database

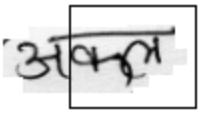

Fig 5: Overlapped character

C. PROBLEM OF TOUCHING CHARACTER

This problem also arises due to different writing styles adopted by different people. While writing, if one character touches the other character then it becomes difficult to recognize both characters . The following image shows this problem

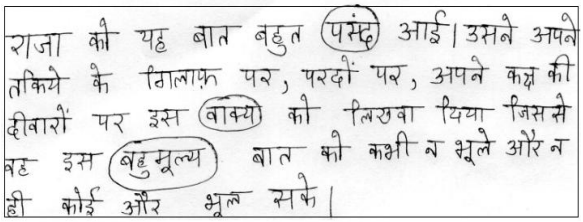

Fig 6: part of database

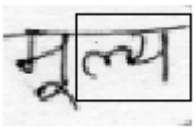

Fig 7: Touching character

D. PROBLEM OF SKEWED CHARACTER

In this problem words in a line are not wriiten into straight horizontal line but the word inclined either upward or downward which causes difficulty to detect the header line for that word

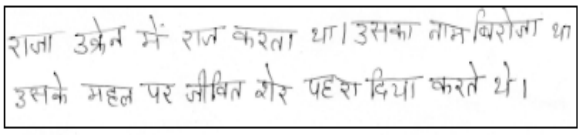

Fig 8: Skewed character

E. PROBLEM OF IRREGULAR INTENSITY WITHIN THE CHARACTER

This occurs due to bad quality of writing material which leads to different intensities of pixel. Then following image shows this problem

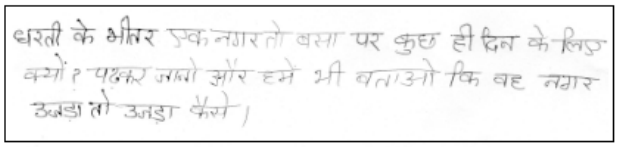

Fig 9: irregular intensity with in character

IV. CONCLUSION

The above problems are created in only handwritten documents not in printed documents. There is a difficult task to remove these segmentation problems. The problems are created due to different handwriting so we find an algorithm to remove such these problems.

V. REFERENCE

[1] Garg , Naresh kumar, kaur, Lakwinder and jindal, M.K. 2011. The hazards in segmentation of handwritten Hindi text. In international journal of computer applications (0975-8887) volume 29-No.2.

[2] Mr. Dipak V. Koshti, Mrs. Sharvari Govilkar. The segmentation of touching characters in handwritten devnagari script. In IJACEE Volume 2: Issue 2 [ISSN 2250 - 3765].

[3] Saiprakash Palakollu, Renu Dhir, Rajneesh Rani 2012. Handwritten Hindi text segmentation techniques for lines and characters. In Proceedings of the World Congress on Engineering and Computer Science 2012 Volume IWCECS 2012, San Francisco, USA.

[4] Ram Sarkar, Samir Malakar, Nibaran Das, Subhadip Basu, Mita Nasipuri 2010. A Script Independent Technique for Extraction of Characters from Handwritten Word Images. In International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 23.

[5] Mr.Sandip N.Kamble, Prof.Mrs. Megha Kamble 2011. Morphological Approach for Segmentation of Scanned Handwritten Devnagari Text. In International Journal of Emerging trends in Engineering and Development ISSN 2249-6149 Issue1,Vol.3.

[6] aarti desai, latesh malik, rashmi welekar 2011. a new methodology for devnagari character recognition. in jmijit ,volume -1 issue 1 ©jm academy issn: print 2229-6115.

[7] Segmentation of Handwritten Hindi Text: A Structural Approach M. Hanmandlu and Pooja Agrawal.

[8] Garg , Naresh kumar, kaur, Lakwinder and jindal, M.K. 2010. A Segmentation of Handwritten Hindi text. In International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 4

Vneeta Rani received her b.tech degree in computer science & Engineering from Lovely professional university (jalandhar) in 2010, Pursuing M.Tech from Guru Kashi University, Talwandi Sabo (bathinda). My research area is Segmentation in optical character recognition.

Pankaj Kumar received her B.tech degree in computer science & Engineering from yadavindra college of engineering, Talwandi Sabo in 2009, Pursuing M.Tech from Guru Kashi University, Talwandi Sabo (bathinda). My research area is Natural language processing.