ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A REVIEW ON VARIOUS TECHNIQUES FOR CHARACTER SEGMENTATION OF HANDWRITTEN TEXT DOCUMENTS

Ankita Chanana*, Chandana Jain

* Student of computer science and engeenring, Jan Nayak Chaudhary devilal memorial college, Sirsa, Haryana, India

Assistant professor of cse department, Jan Nayak Chaudhary devilal memorial college, Sirsa, Haryana, India

ABSTRACT

Character Segmentation is the process of separating characters from words. Character Segmentation of handwritten text is a challengenging task in O.C.R because of its features and varied writing styles of different writers. Handwritten text is also prone to the problems of overlapped characters, touching characters, skewed characters, broken characters which makes the segmentation process more complicated. Accuracy of character segmentation depend on upto which extent these problems are tacled and character is segmented. In this paper we provide a review of various techniques used for character segmentation and also discuss existing problems of segmentation. Correct segmentation is necessary for correct recognition of characters.

KEYWORDS: Segmentation, Classification, Feature Extraction Recognition

INTRODUCTION

O.C.R(optical character recognition) has been one of the most challenging research areas in the fields of image processing. The main aim of O.C.R. is to convert the scanned documents into editable format. O.C.R helps us to read and recognize the scanned documents. The main process of O.C.R is shown here :-

O.C.R consists of following stages-

#pre-processing ;- In the pre- processing step O.C.R removes the noise, garbage and optimizes the image to give best results.

#Segmentation :- In this step segmentation is carried out. Segmenting of lines from a scanned document, then separating words from lines and then characters from words.

#Feature Extraction:- In feature extraction stage text segment is analysed and a set of features are selected that can be used to uniquely identify the text segment.

#classification and Recognition :- The classification stage uses the features extracted in in the previous stage to identify the text segment according to the present rule.

The whole process of Segmentation is devided into following types:-

1) **Line Segmentation** :- In Line Segmentation a text block is taken as input and scanned horizontally. Frequency of black pixels are detected in order to construct a row.

2) **Word Segmentation** :- When a line has been detected, then each line has to scan vertically. Number of black pixels in each column is calculated to make a column. When there is no black pixel is found in vertical scan then there is a space between two words.
3) **Character Segmentation ;-** In Character Segmentation a word is separated into characters, each individual character and composite characters.

In character Segmentation the basic step is zoning level that is devide the word into different zones.
HEADER LINE- Header line is that portion of the word which has maximum intensity of black pixels.

 # BASE LINE- Base line is that portion of the word which contains minimum intensity of pixels.

The next basic step is devinding a word into different zones

 # UPPER ZONE - Upper part of header line is known as upper zone.

#LOWER ZONE- Lower part after base line is known as lower zone

 # MIDDLE ZONE – Middle zone consists of the data resides between header line and base line. This zone contains most of the data as compared to upper and lower zone.

All the three zones and header and base lines are shown in the fig. below.

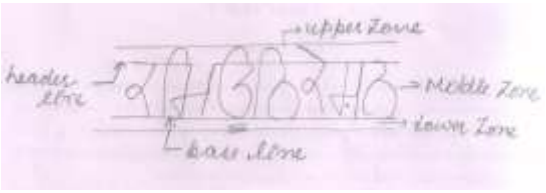

Figure 1: Classification of words into zones

MATERIALS AND METHODS

In today's Research Senarios, there are different techniques which have been discussed for character segmentation. Segmentation on handwritten text is difficult process then segmentation of characters on printed document.

An end detection algorithm of segmentation of broken and touching characters in Handwritten Gurumukhi word" A survey done by Parika Mangla and Harleen Kaur(**IEEE).** The paper provides a new segmentation technique based on neighbouring pixels for touching and broken characters of Handwritten Punjabi text that is the Gurumukhi Script.[1]. "Fragmentation of Handwritten Touching Characters in Devnagari Script" A survey by Shuchi Kapoor and Vivek Verma. In which a technique is developed to provide a solution to the touching characters[2]

"To Extract Feature of Handwritten Devnagari Script" A feature Extraction technique is used by Ajay Garg and Simpel Jindal to recognise handwritten Devnagari Script document[3].

"Text Line Detection and Segmentation in Handwritten Gurumukhi Scripts" An effective method is proposed by Namisha Modi and Khushneet jindal for text line segmentation in Handwritten Punjabi document that deals with the problems like overlapped and connected components[4].

"Handwritten Hindi Text Segmentation Techniques for Line and Characters" a survey done by Saiprakash palakollu, Renu Dhir and Rajneesh Rani. This paper deals with the various methods of line and character segmentation.. In this paper the basic technique which is followed is that header lines are detected and converted as straight lines. After that, each word is divided into upper modifier then consonant and then into lower part, to make character segmentation easy. Algorithm is based on finding header and base lines by estimating the average line height. This technique efficiently segments lines with accuracy upto 93%, segment words with accuracy upto 96% and characters with

accuracy upto 89%. This method for line segment is working efficiently in the cases of different text sizes and different resolution.[5]

"The Hazards in segmentation of Handwritten Hindi Text" In this paper Naresh Kumar Garg, Lakhwinder Kaur and M.K. Jindal done a very good job, This paper provides an over all view of the problems which are currently exist in Handwritten text[6].

"Segmentation of Handwritten Hindi Text" this survey is also done by Naresh Kumar Garg, Lakhwinder Kaur and M.K. Jindal also work on Segmentation of Handwritten Hindi Text. In this paper new segmentation technique based on structural approach is provided [7].

"Isolated Handwritten Words Segmentation Techniques in Gurumukhi Script" In this paper Galaxy Bansal and Dharamveer Sharma worked on Segmentation of Isolated words in Gurumukhi Script. The main objective of this paper was to discuss a compele solution for character segmentation phase of Gurumukhi Script[8].

"Segmentation of Printed Text in Devanagri Script and Gurumukhi Script" A survey is done by vijay kumar and Pankaj K. Sengar on printed text in Devanagri and Gurumukhi Script. This paper deals with the line, word, character and top character segmentation for printed Hindi text in Devanagari Script and it it also describe the line and word segmentation for printed text in Gurumukhi Script. In this paper a single algorithm is proposed for segmentation of Devnagari Script and Gurmukhi Script. A performance of 100% at line level, approximately 100% at word level , 99% at character level and 97% top character level.[9]

"A Study of Different kinds of Degradation in printed Gurumukhi Script" M.K. Jindal, R.K. Sharma and G.S. Lehal studied different kinds of degradation in printed text and provide solution to some of them.[10].

PROBLEMS OCCUR IN SEGMENTATION OF CHARACTERS

1) **Touching Characters**
   In handwritteni text , adjacent characters touch each other and separation of such characters is a major task. Touching of characters decrese the efficiency of recognition of character. If no gap is present between characters then they are treated as a single character.

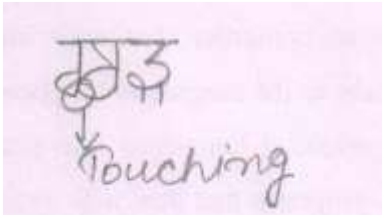

Figure 2: touching characters

2) **Overlapping Characters**
   Overlapping of characters is another major problem that exits in segmentation. Characters can overlapped in all the three regions –
   Middle zone characters with other middle zone characters.
   Middle zone characters overlap with lower zone characters.
   Middle zone characters overlap with upper zone characters.

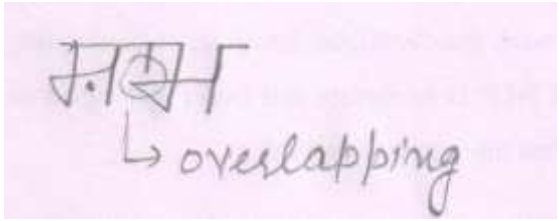

Figure 3: overlapping characters

3) **Broken characters**
Broken characters or missing characters are also one of the problems which can be seen in handwritten gurmukhi text. Broken characters leads to over segmentation. Broken characters mostly found in middle zone. The cracters can be broken horizontally or vertically. Shown in fig.

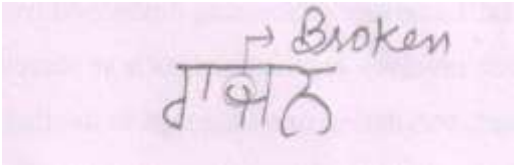

Figure 4: Broken characters

4) **Skewed characters.**
Characters can be skewed right or left due to variability of writing styles of different writers The main problem of skewed characters is due to its uneven header line, which creates problem in segmenting by vertical profile projection .

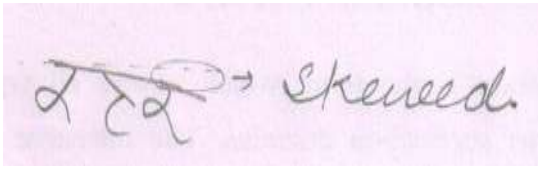

Figure 5: Skewed characters

VARIOUS TECHNIQUES USED IN CHARACTER SEGMENTATION

a) Horizontal Projection
For a given binary image of size X * Y where X is the width and Y is the height of the image, the horizontal projection is defined as:

HP(i), i= 1,2,3…..Y
Where HP(I) is the total number of black pixels in ith row.

H.P. is used to calculate the length of header line.

b) vertical projection
for a binary image of size X*Y, where X is the width and Y is the height of the image, the vertical projection is-
VP(j), j=1,2,3…….X

c) Water reservoir technique
Water reservoir technique is an efficient technique used for detecting whether characters are touching or not. Here in this technique, reservoir is an area to illustrate where characters touch.

Reservoir is obtained by considering accumulation of water poured from top or bottom of the characters. If the characters are touching then reservoir is formed, otherwise not.

d) Vertical Projection Technique
this method is used for character segmentation after header line detection.
After header line detection.a vertical projection is made from top to bottom of the word and columns with zero black pixels are treated as delimiter from upper modifier.

To separate lower modifier difference in height of characters is calculated. This technique provides good result for separation of characters in all the three zones. But cannot work for toucing or overlapping of characters.

e) End detection technique

This technique is very efficient for all the three isolated, touching and broken characters. In this technique mid point formula is used to solve the problem. But again it does not work on overlapping characters and is size dependent.

RESULTS AND DISCUSSION

The following table show the comparison of results obtained by existing techniques of character segmentation

Table 1. Comparitive study of existing work on different characters

researchers	Techniques used	Work done on overlapping characters	Work done on touching characters	Work done on broken characters	Work done on isolated characters
M.K. Jindal	Water reservoir	No	Yes	no	Yes
Naresh kumar garg	Vertical profile projection	No	No	No	Yes
Parika mangla	End detection	No	yes	Yes	Yes

CONCLUSION

From this paper we conclude that, work is done on touching characters, missing characters but there is no work done on overlapping characters according to the latest IEEE[1] paper. So a lot of work is to be done in character segmentation.

REFERENCES

[1] Parika Mangla,Harleen Kaur"An end detection algorithm of segmentation of broken and touching characters in Handwritten Gurumukhi word" in IEEE,2014.
[2] Shuchi Kapoor and Vivek Verma "Fragmentation of Handwritten Touching Characters in Devnagari Script"in IJITMC,2014 .
[3] Ajay Garg and Simpel Jindal "To Extract Feature of Handwritten Devnagari Script"in IJ of Advanced Research in Computer and Communication,2014.
[4] Namisha Modi and Khushneet Jindal "Text Line Detection and Segmentation in Handwritten Gurumukhi Scripts"in IJ of Advanced Research in Computer Science and Software Engeenring, 2013.
[5] Saiprakash palakollu, Renu Dhir and Rajneesh Rani "Handwritten Hindi Text Segmentation Techniques for Line and Characters" in WCECS 2012.
[6] Naresh Kumar Garg, Lakhwinder Kaur and M.K. Jindal "The Hazards in segmentation of Handwritten Hindi Text" in International Journal of Computer Applications,2011
[7] Naresh Kumar Garg, Lakhwinder Kaur and M.K. Jindal "Segmentation of Handwritten Hindi Text" in International Journal of Computer Applications,2010
[8] Galaxy bansal, Dharamveer Sharma "Isolated Handwritten Words Segmentation Techniques in Gurumukhi Script"in IJ of computer applications,2010
[9] Vijay kumar and Pankaj K. Senagar "Segmentation of Printed Text in Devanagri Script and Gurumukhi Script" in IJCA 2010.
[10] M.K. Jindal, R.K. Sharma and G.S. Lehal "A Study of Different kinds of Degradation in printed Gurumukhi Script" in IEEE, 2007.