ISSN 22773061

Segmentation of Handwritten Text Document Written in Devanagri Script for Simple character, skewed character and broken character

Vneeta Rani, Dr.Vijay laxmi
M.Tech. Final Year Student, deptt(CSE)
Gurukashi University Talwandi Sabo.
Vneeta_103@rediff.com
Associate Professor
Department of Computer Science
Gurukashi University Talwandi Sabo.
cse_Vijay2003@yahoo.co.in

Abstract

OCR (optical character recognition) is a technology that is commonly used for recognizing patterns artificial intelligence & computer machine. With the help of OCR we can convert scanned document into editable documents which can be further used in various research areas. In this paper, we are presenting a character segmentation technique that can segment simple characters, skewed characters as well as broken characters. Character segmentation is very important phase in any OCR process because output of this phase will be served as input to various other phase like character recognition phase etc. If there is some problem in character segmentation phase then recognition of the corresponding character is very difficult or nearly impossible.

Keywords: - OCR; Segmentation; Character segmentation; Broken character segmentation; Skewed character segmentation; Devnagari script.

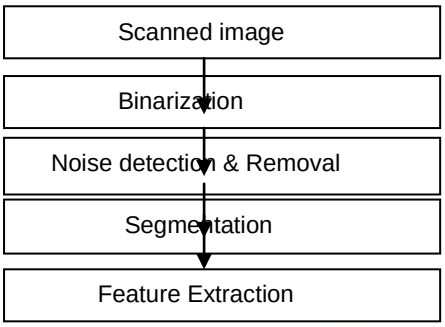
Council for Innovative Research
Peer Review Research Publishing System
Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY
Vol 8, No 1

editor@cirworld.com
www.cirworld.com, member.cirworld.com

ISSN 22773061

1. Introduction

OCR is a technology that enables us to convert different types of scanned document into editable documents. It is a part of electronic document Analysis system. It is used to extract text from scanned images of type written, handwritten or printed text. Process of OCR can be described as following:



Explanation of these processes is as follows:

a) **Scanning image**: In this step the document is converted into scanned image with the help of image scanner.

b) **Binarization**: In this step gray scale images are converted to binary image with the help of OCR Software.

c) **Noise detection & removal**: In this technique low pass filters are used for removing noise.

d) **Segmentation**: with this technique which partitions handwritten words into individual character. There are various types of segmentation which are paragraph segmentation, line segmentation, word segmentation and character segmentation.

e) **Feature Extraction**: After performing the segmentation process features can be extracted for corresponding characters by using various feature extracting techniques.

2. Character Segmentation

Character segmentation is a procedure in which from the word segmentation we take out only characters. Character segmentation is a critical step of OCR system. Character segmentation is an operation that seeks to decompose an image of a sequence of characters into sub images of individual symbols. It is depends on the script used in writing the document. A poor segmentation process produces misrecognition or rejection segmentation process carried after out only the pre processing of image.

Fig 2: Input script for word image

Fig 3: Segmented script output into word

3. Problems in Character segmentation

3.1 Broken Character

Character can be broken due to writer's pen or page quality used. The following figure shows the broken character

ISSN 22773061

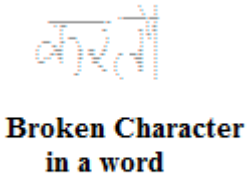


Fig 4: Broken Character in a Word

Segmentation of the broken character is quite difficult because vertical profile projection technique assumes the broken parts of the characters as individual characters.

The following diagram shows a wrong segmented word due to broken character problem:

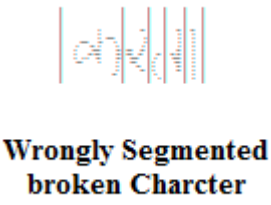


Fig 5: wrongly segmented word with broken character

Solution: Broken character can be segmented by scanning the neighboring pixels before segmenting the word into characters. Neighboring pixels on both left and right side are to be checked and if the black pixels are there then that represents the character is broken and not to be segmented but if there are white pixels in its neighbor then these pixels are treated as a gap and hence to be segmented.

The following figure shows correctly segmented broken character by using above solution

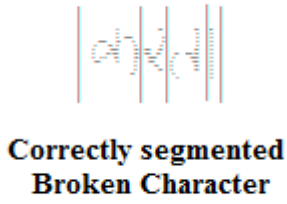


Fig6: Correctly segmented word with broken character

3.2 Skewed Character

Characters in a word may have slant either upward or downward which results in the skewed characters. Skewed characters are generated due to the writing skills of a person.The following figure shows a skewed character

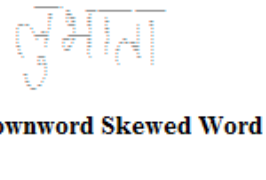


Fig7: A Skewed Word

While segmenting a skewed character, the problem arises in detection of the header line of the word which results in improper segmented word.

The following diagram shows a improper segmented word due to skewed character.

ISSN 22773061

Improper Segmented skewed Character

Fig8: wrongly segmented Skewed word

As shown in fig 8 skewed word is not segmented properly.

Solution: To solve this problem of skewed character header line of a word is detected by calculating the frequency of multiple neighboring rows.

The following figure shows correctly segmented skewed character by using above solution

Downword Skewed word

Correctly Segmented word

Fig9: Correctly segmented skewed character

As show in the fig. skewed word is segmented properly using our approach.

4. Our Approach

Our algorithm to segment the characters which may be skewed or broken have following steps:

Step1. Scan the document into image from which words are to be segmented into characters

Step 2: Binarize the scanned image

Step 3: Remove the noise from the binarized document

Step 4: Extract the line from which we want to segment the words

Step 5: Calculate the frequency of black pixels in each row along with neighbors using horizontal profile projection technique.

Step 6: Find the row with the highest numbers of black pixels and treat that row as header row.

Step 7: Remove that header row from the word for segmentation

Step 8: Using vertical profile projection technique parse the word column wise

Step 9: Check for each i^{th} column of the word if all the pixels are white and if so then check i-1 and i+1 number of pixels. If all three pixels are white then treat them as gap between two characters and then segment the word. But if

either of the two pixels (i-1, i+1) is black, than it represents the broken character and don't segment the word from the i^{th} pixel.

Step 10: Exit

5. Experiments and results

We have tested our algorithm on 30 documents of different writings. Our system shows accuracy of 96%. The results after applying the algorithm are as follows:

ISSN 22773061

Original Word Segmented Word

Result Fig. 1

Input Orignal Word

Correctly Segmented Word into Characters

Result Fig. 2

Accuracy of character Segmentation

Type of Word	Total No. of Words	Segmented Words	Accuracy
Simple	500	500	100%
Skewed	200	190	95%
Broken	200	192	96%

Result Table fig: 3

6. Conclusion

From the result table I we can say that the new system is giving very good results. Our System to segment the simple words shows the accuracy of 100 % and on skewed words System shows the accuracy of 95% while on the broken characters it shows accuracy of 96%. System can be extended to segment the words with overlapped and/or touching characters.

7. References

[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] 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.

[5] 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.

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

[7] 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

[8] Ashwin S Ramteke, Milind E Rane,"Offline Handwritten Devanagari Script Segmentation" in 2012

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.

Dr. Vijay Laxmi received her B-Tech degree in computer science & Engineering from SLIET Longowal in 2003, and Ph.D degree in Computer Science & Engineering in year 2012. Her research area is Grid Computing & OCR. She has published 25 research papers in various National/International Conferences and Journals. At present, she is engaged in Guru Kashi University, Talwandi Sabo, and Punjab as Dy. Dear Research and an Associate Professor in Computer Science Engineering & Information Technology department.