International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 11 6384 - 6387

Zone Segmentation and Thinning based Algorithm for Segmentation of Devnagari Text

Er.Japneet Kaur
Student,CSE Dept.,BFCET
Bathinda,Punjab
japneet_bhullar@yahoo.co.in

Er.Suppandeep Kaur
Asstt. Proff.,CSE Dept,BFCET
Bathinda,Punjab

Dr.Meenakshi Arya
HOD,CSE Dept,BFCET
Bathinda,Punjab

Abstract:- Character segmentation of handwritten documents is an challenging research topic due to its diverse application environment.OCR can be used for automated processing and handling of forms, old corrupted reports, bank cheques, postal codes and structures. Now Segmentation of a word into characters is one of the major challenge in optical character recognition. This is even more challenging when we segment characters in an offline handwritten document and the next hurdle is presence of broken ,touching and overlapped characters in devnagari script. So, in this paper we have introduced an algorithm that will segment both broken as well as touching characters in devnagari script. Now to segment these characters the algorithm uses both zone segmentation and thinning based techniques. We have used 85 words each for isolated, broken, touching and both broken as well as touching characters individually. Results achieved while segmentation of broken as well as touching are 96.2 % on an average.

Keywords:- *Segmentation , OCR , Zone segmentation, Thinning Based Technique, BW Morph algorithm*

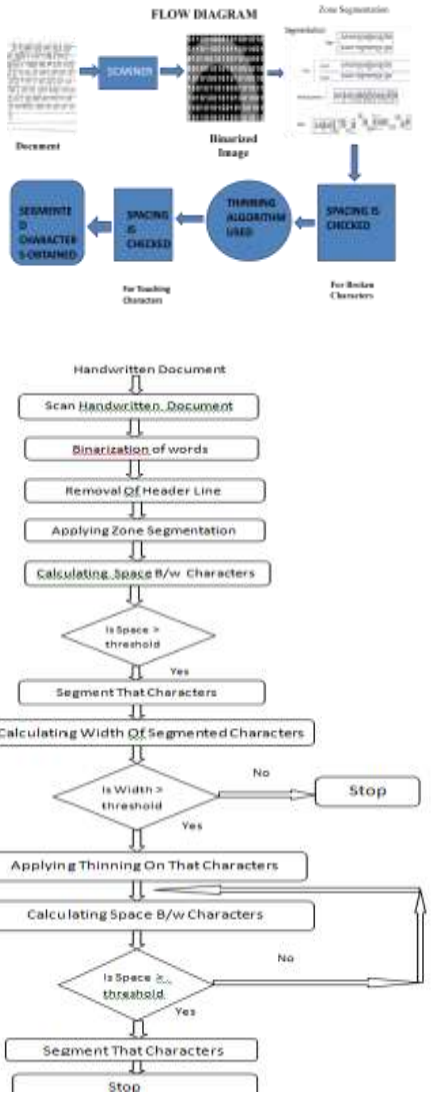
I. Introduction:

Segmentation is method that partitions the printed or handwritten content into individual lines, words or characters. Segmentation is one of the testing and challenging fields in OCR[1]. It is an operation that tries to divide an image into sub images of individual symbols. It is one of the decision procedures in OCR. In text report image examination, the significant step is extraction of text lines from documents, and afterward the text lines are divided into words and characters.

Now Segmentation of a word into characters is one of the important challenges in optical character recognition[22]. This is even more challenging when we segment characters in an offline handwritten document. In our thesis ,we formulate an algorithm to segment the handwritten devnagari text [20].

The next hurdle in this context is segmentation of handwritten text which contains broken, touching or overlapped characters[31]. Some work is done to segment these types of characters. But of work on both broken and partially overlapped characters touching characters is our main focus. Because in real life it is not possible to get text only with broken or overlapped or touching characters.

So in real life we can have text which contains both broken as well as overlapped characters. So there is a need to develop such a technique which can deal with presence of both broken and touching characters of Devnagari script. To deal with broken and touching characters zone segmentation and thinning algorithm is used respectively. In case of zone segmentation, image is divided in upper, middle and lower zones and in thinning algorithm or skeleton algorithm the characters are made thin[26]. To deal with broken and touching characters simultaneously mixing of both algorithms can be done so as to create a hybrid algorithm to segment broken as well as touching characters as shown below in flow chart:



flow chart of proposed work

 ISSN: 2321-8169
6384 - 6387

II. Methodology :

First of all , the document is converted into scanned image with the help of image scanner. The scanned image can be in any format for eg .jpg , .bmp , .tif.

Than after scanning document, binarization is done where the scanned images are converted into binary images with the help of OCR Software. Binarization is a process of converting image into binary form that means in the form of 0s and 1s[2].

After performing binarization of the scanned images, the header line of the word is removed to perform further segmentation. Header line is an important part of devnagari word that glues all the words together.

Firstly the row wise pixel density is checked, where 0 is represented for presence of pixel and 1 is represented for absence of pixels(white spaces) as follows:

```
Repeat for i from 1 to r.
Each row = binaryImg(i,:)  //  Finding row
                               with maximum number of
0
                                 (black pixels)
 nr(i) = Sum (eachrow()==0) //  array (sum up
                                    no of zeros)
End loop.
```

Now after calculating the maximum value of index(nr),than remove upper 13 and lower 13 rows. This was the header line is removed. After removing the header line the phase of character segment starts. In this phase, zone segmentation is done to segment characters into various zones namely upper, middle, lower zones. Now segmentation of zones is done in following 2 steps: Firstly we will remove the starting and ending white spaces Of the image Secondly ,now after removing starting and ending white spaces we, will segment the words into different zones. So,we will count characters vertically If pixel difference is less than onecount than count of no of columns having black pixels than it is considered as single character Otherwise broken characters. Similarly count horizontally characters by extracting rows instead of columns.

Now after applying zone segmentation on the word, it will detect the broken characters present in the word in different zones. Broken characters may be present in any zone but to detect the broken character in middle zones is more difficult. So, by applying zone segmentation we can detect broken characters present in the word Now after detecting the broken characters by applying zone segmentation we will calculate the width of characters to check whether the characters obtained are single character or multiple characters . we can calculate the width of characters as follows:

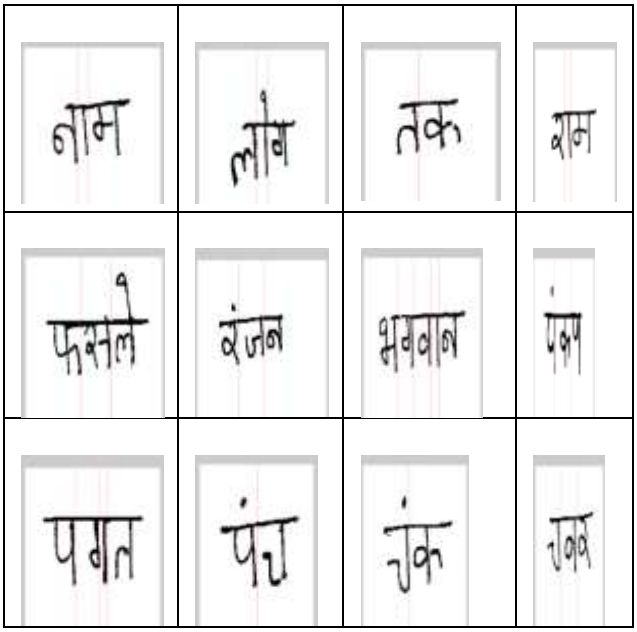
If ardiff ≥ threshold value // if character difference ≥ threshold value
i.e If width > threshold)

Than Consider it as multiple characters Otherwise consider it as single character and if we obtain multiple characters than go to next step. Now after calculating the width of characters, if we found any multiple characters or any touching characters than we will apply BW Morph algorithm on it to segment them[37].Now after detecting the touching characters it again calculate the characters vertically as well as horizontally to segment them.

III. Results :

Output For Isolated Characters

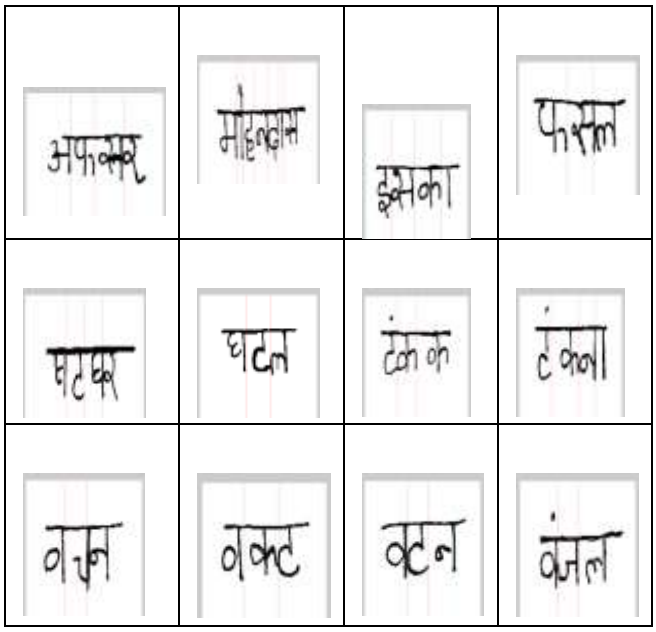
Output For Broken Characters:



 ISSN: 2321-8169
6384 - 6387

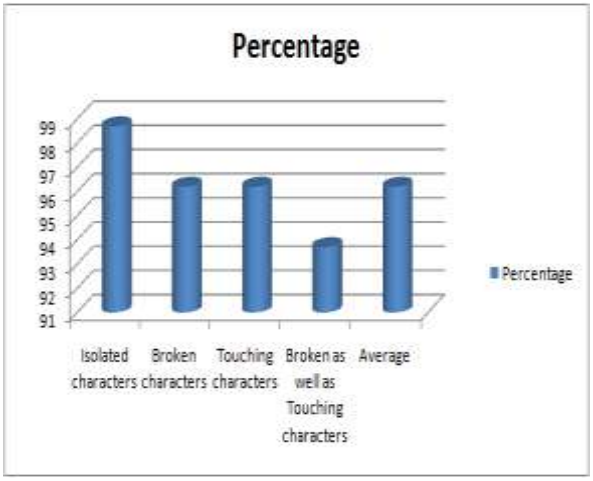
Output For Touching Characters

Outputs For Both Broken as well as Touching characters:



At the end after segmenting all the 85 characters individually for all. The results achieved are shown below in the table:

TYPE OF SEGMENTATION	PERCENTAGE
Isolated Characters	98.7
Broken Characters	96.2
Touching Characters	96.2
Broken as well as Touching Characters	93.7
Average	96.2



IV. Conclusion

The proposed algorithm work over test image database having isolated, broken, touching and broken as well as touching charcters. The proposed technique is easy for segmentation of devnagari text. The algorithm summarised that considerable amount of work has been carried out to segment words of machine printed Roman script, devnagari and there are varied and some well developed techniques used for segmentation of touching ,broken and overlapping characters, But very little work has been carried out for handwritten scripts like Devnagari etc. Segmentation of handwritten words in Devnagari script is a challenging task because of the structural properties of Devnagari character set and writing styles of individuals.

- Proposed algorithm solves our problem defination
- Proposed algorithm improves the accuracy as compared to existing techniques
- The technique not only gives the accurate results on isolated, touching, broken characters but also on broken as well as touching in single character.
- It gives accurate results on Handwritten Devnagari documents which is more difficult than printed document because of various writing styles.

V. Future Scope

The proposed method could be improved for better results with partially overlapped characters .More Research work can be done for fully overlapped characters. Accuracy of the proposed technique can be further improved. Thus we can futher improve the accuracy of character segmentation by overcoming challenges of character segmentation.

VI. References

[1] John S. Denker et .al in "Image Segmentation and Recognition" , SFI/CNLS,1-24,1992.

[2] S. Naoi , Y. Hotta et . al in "Global Interpolation in the Segmentation of handwritten Characters Overlapping a Border" , IEEE.0-8186-6950-0/94,149-153,1994.

[3] Zhongliang Fu et . al in "Algorithm for Fast Detection and Identification of Characters in Gray-level Images" , IAPRS. Vol. XXXIII, Part B3,305-311,2000.

[4] Utpal Garain et . al in " Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts Using Fuzzy Multifactorial Analysis", IEEE. 1094-6977/02,449-459,2002.

[5] Nafiz Arica et . al in " Optical Character Recognition for Cursive Handwriting" IEEE. 0162-8828/02,801-813,2002.

[6] Cheng-Lin Liu et .al in " Reading Lexicon-Driven Segmentation and Recognition of Handwritten Character Strings for Japanese Address Reading", IEEE,1-6,2002.

[7] Jiqiang Song et . al in "Recognition of Merged Characters Based on Forepart Prediction, Necessity-Sufficiency Matching, and Character-Adaptive Masking", IEEE. 1083-4419,2-11,2005.

[8] Keh-Shih Chuang et . al in "Fuzzy c-means clustering with spatial information for image segmentation", ELSEVIER. 0895-6111,9-15,2006.

[9] Dharam Veer Sharma et . al in "An Iterative Algorithm for Segmentation of Isolated Handwritten Words in Gurmukhi Script", IEEE. 0-7695-2521-0/06,1-4,2006.

[10] M Swamy Das et . al in "Segmentation Of Overlapping Text Lines, Characters In Printed Telugu Text Document Images", 2010.

[11] Rahul Kala et.al in "Offline Handwriting Recognition using Genetic Algorithm",IJCSI, Vol. 7, Issue 2, No 1,16-26,2010.

[12] Hanan Aljuaid et . al in "A Tool to Develop Arabic Handwriting Recognition System UsingGeneticApproach" , IJCS.1549-3636, 619-624 ,2010.

[13] Rajiv Kumar et .al in "Character Segmentation in Gurumukhi Handwritten Text using Hybrid Approach" , IJCTE.VOL.3,NO.4,499-501,2011.

[14] Trung Quy Phan et .al in "A Gradient Vector Flow-Based Method for Video Character Segmentation" IEEE.1520-5363/11,1042-1028,2011.

[15] Alex Graves et . al in "A Novel Connectionist System for Unconstrained Handwriting Recognition" EEE,1-6,2011.

[16] Chunming Li et .al in "A Level Set Method for Image Segmentation in the Presence of Intensity Inhomogeneities With Application to MRI" , IEEE. 1057-7149,2007-2016,2011.

[17] Mo Chen et . al in "Hidden-Markov Model-Based Segmentation Confidence Applied to Container Code Character Extraction" , IEEE. 1524-9050,1147-1156,2011.

[18] Manas Yetirajam et . al in "Recognition and Classification of BrokenCharacters using Feed Forward Neural Network to Enhance an OCR Solution", IJARCET.VOL.1,11-15,2012.

[19] Mohamed A. Ali et . al in "AN EFFICIENT THINNING ALGORITHM FOR ARABIC OCR SYSTEMS", SIPIJ.VOL.3,NO.3,31-38,2012.

[20] Rajeshwar Dass et . al in "Image Segmentation Techniques"IJECT, 2230-7109, Vol. 3,66-70,2012.

[21] Dr. Jenila Livingston L.Min et .al "Text Detection From Documented Image Using Image Segmentation", IJTEEE.VOL.1,1-5,2013.

[22] Sukhpreet Singh in "Optical Character Recognition Techniques: A survey" (IJARCET).Volume 2, Issue 6, 2010-2015,2013.

[23] Reza Azad et . al in " A novel and robust method for automatic license plate recognition system based on pattern recognition" , ACSIJ , Vol. 2, Issue 3,64-70, 2013.

[24] R.Yogamangalam et . al in "Segmentation Techniques Comparison in Image processing" ISSN : 0975-4024. Vol 5. No 1,307-313,2013.

[25] Xiang-Dong Zhou, et . al in "Handwritten Chinese/Japanese Text Recognition Using Semi-Markov Conditional Random Fields" , IEEE. 0162-8828/13,2413-2426,2013.

[26] S Gomathi @ Rohini et . al in "Trimming Approach for Word Segmentation with focus on Overlapping Characters" IEEE.978-1-4673-2907-1/13,1-4,2013.

[27] Atallah et al in "Skelton extraction:comparison of five methods on the ARABIC IFN/ENIT database" , IEEE.978-1-4799-3999-2/14,50-59,2014.

[28] Sakkayaphop Pravesjit et .al in "Segmentation Of Touching Lanna Characters",IEEE,1-5,2014.

[29] Anna Tonazzini et . al in "Non-stationary modeling for the separation of overlapped texts in documents", IEEE.978-1-4799-4874-1/14.2314-2318,2014.

[30] Sujata Saini et . al in "A Study Analysis on the Different Image Segmentation Techniques" , IJICT. ISSN : 0975-4024 Vol 5 No 1. pp. 1445-1452,2014.

[31] Wichian Premchaiswadi et . al in "Broken Characters Identification for Thai Character Recognition Systems".

[32] Volkmar Frinken et . al in "BLSTM Neural Network basedWord Retrieval for Hindi Documents" , IEEE.1520-5363/11,83-87,2014.

[33] Naveen Sankaran T et .al in "Recognition of Printed Devanagari Text Using BLSTM Neural Network", IIIT/TR/2012/-1,1-4,2014.

[34] Naveen Sankaran et . al in "Devanagari Text Recognition: A Transcription Based Formulation", IJERCSSE.VOL.4,1030-1034,2014.

[35] Ms. Neha Sahu et . al in "An Efficient Handwritten Devnagari Character Recognition System Using Neural Network", IEEE.978-1-4673-5090-7/13,1-5,2014.

[36] ZhongHua Cao et . al "A New Drop-Falling Algorithms Segmentation Touching Character", IEEE.978-1-4244-6055-7/10,380-383,2014.

[37] Meihua Bao et.al "Optimization of the *bwmorph* Function in the MATLAB Image Processing Toolbox for Binary Skeleton Computation"IEEE.978-0-7695-3645-3/09,2009.