WWJMRD 2017; 3(11): 253-257
www.wwjmrd.com
International Journal
Peer Reviewed Journal
Refereed Journal
Indexed Journal
UGC Approved Journal
Impact Factor MJIF: 4.25
e-ISSN: 2454-6615

Manpreet Kaur
Research scholar, Yadavindra college of Engineering, Talwandi Sabo, Punjab, India

Ashok Kumar Bathla
Assistant Professor, Yadavindra College of Engineering, Talwandi Sabo, Punjab, India

Correspondence:
Manpreet Kaur
Research scholar, Yadavindra college of Engineering, Talwandi Sabo, Punjab, India

Segmentation of Characters of Devanagari Script Documents

Manpreet Kaur, Ashok Kumar Bathla

Abstract
Segmentation of Characters is the final objective of OCR. OCR proceeds through the segmentation procedure by having sequence of steps like input, preprocessing, segmentation. The procedure of segmentation is applied on both handwritten and machine written documents. When the text in any script is single font with single alignment and with single width will be very easy to segment. Because each character has single unique behavior. But the process will be difficult to handle when the text is handwritten. Because while writing using hand various complexities like overlapping, touching or skewness, broken, and non-alignment arises. Due to these reasons the success rate of the segmentation will be reduced. To overcome these issues various segmentation techniques like water reservoir, Collusion dilation algorithms are used. In proposed research segmentation of broken letters of devnagri scripts is taken place. The methodology used is Collusion dilation. This technique has been applied on 500 images dataset of handwritten Devanagari script. The result generated is 96%.

Keywords: OCR, Head line, Middle zone, upper zone, lower zone

Introduction
Various researchers have given focus on the segmentation of characters from Devanagari script. This Devanagari script is handwritten or machine written. It is very easy to segment the characters written using machine. This machine written text has uniform alignment and predictive writing. But the segmentation results reduce to substantially low for handwritten text. This handwritten text has non predictive nature. Each user writes the text is its own way. Also the text written using hand is not equally spaced.
Handwritten text recognition in devanagari script can be differentiated into two categories. One is online handwritten character recognition, and second is offline character recognition. In online character recognition the handwritten text is written in PDA. Sensor picks up the pen-tip movements as well as pen-up and pen-down strokes. In offline handwritten character recognition a dataset is prepared with scanning the paper documents. This handwritten text dataset is read by OCR. The main objective of reading Hindi script is to convert historical documents in digital format.
Recognition of offline handwritten Devanagari script characters is a goal of many research efforts in the pattern recognition field. Many techniques have been applied so far. Each technique has provided the results not 100% but near to 90%. That means there is a scope of further research. Through which characters can be recognized.

Character Segmentation in OCR
OCR is based on multiple steps. These steps have to be taken in sequence from one to other. Each step is taken in sequence. These steps are conversion, preprocessing, segmentation, representation, training and recognition and last is post processing. To simplify the process of segmentation there is a need of scanned convert the paper document to digital format. While in pre-processing various types of noises skewness and overlapping etc has to remove. So that while segmenting the line or word success rate can be increased. In segmentation the OCR system will segment the required portion of the image and will process this segment there on. In post processing the segmented image will be having various types of noise and

alignment and categorization features. So that based on these prepared classification various features are being extracted.

Introduction of Devanagari Script

After English and Chinese, Hindi is the third most widely used language and all over the world, there are approximately 500 billion people who write and speak Hindi. In India, Devanagari is the basic script of many languages like Sanskrit and Hindi. Hindi language have 11 swar (vowels) as shown in Fig. 1.2, 33 Vyanjan (consonants) [3] as shown in Fig. 1.3 and 12 modifiers (special symbols) as shown in Fig. 1.4. Almost all the words are combination of these two along with the modifiers placed on left, right, bottom and above. It is written from left to right.

Levels of Text Segmentation

The text read from the digital text document involves various types of sequential activities. The process of reading text from the text document involves three types

1. **Line segmentation:** Line segmentation is first step of segmenting the text from the text document. This line segmentation is to be taken place from left to right and from top to bottom. The lines are segmented based on base line or on the basis of head line.
2. **Word Segmentation:** Once lines are segmented words from the line will be extracted. These words are segmented based on space. This segmentation can be done using vertical lining on space lies between two words.
3. **Character Segmentation:** Character level segmentation is the final step for segmenting the text. From segmented words, characters will be segmented. These characters are

Devanagari Script Character Segmentation Problems

a. Touching characters

The characters are touching in both lower zone and middle zone.

Fig. 1: touching at middle and lower zone

Fig. 2: Touching at Middle Zone

b. Broken characters

The characters in the middle zone are broken.

Fig. 3: Broken Letters

c. Overlapping

Fig. 4: Overlapping characters

Proposed Methodology

In proposed research Devanagari script handwritten text is to be segment. The segmentation of the characters which are broken is specially undertaken in current research. The technique used for segmenting this character based on Collusion dilation technique. This technique is to identify the natural bonds and enforced bonds. Based on these enforced bonds identification dilation is enforced. This technique involves many steps.

1. Input the scanned image.
2. Pre-process the image.
3. Segment the image

Input the scanned image of Devanagari hand written script words.

In preprocessing the image is de-noise the image and align the image.

In segment the image first head line will be removed. Later on segment the characters. The broken letters are recognized using Collusion dilation technique.

Related Work

Abdelhak Boukharouba (2016) et. al:[1] this paper is based on skew detection on the basis of base line and Hough transformation. This paper has proposed a solution on the dataset of language based on Arabic language. This paper has done this skew detection on base line, because the header line is absent in the Arabic letters. This paper has proposed a solution for the success rate of 90%. This paper has proposed a algorithm in which first the skew is detected. This skew detection angle is detected using Hough transformation. This skew can be detected based on header line and base line. In case of Arabic there is no headline present so identification is based on base line. But for the Punjabi Gurumukhi the segmentation is based on headline is detected. This headline based on angle with the horizontal axis, unskew is performed. This unskew angle is either positive or negative. That means wither clock wise direction or anti-clock wise fashion.

Bishnu a. (2012) et. al:[3] this paper is based in cursive hand written text of the bangle language. This mechanism is based on identifying the overlapping characters and they are also having no head line of middle line. The algorithm provides in this paper has success rate of 94%. They have

taken a large dataset of bangle written text. This paper has proposed a technique named as recursive contour following on to the hand written text recognition. In this paper they have used segmentation on to the hand written text on bangla language. Based on the writing style of the language different zones across the height of the word is detected. These zones provide the structural features of the constituent characters Recursive contour following in one of the zones across the height of the word to find out the extents within which the main portion of the character lies. If the subsequent characters are not touching then this technique provides good results. This technique follows various steps like segmenting the line. Once the line is recognized words are separated. Then from the words characters are recognized.

Dharam Veer Sharma, et.al (2009) [5] As per the proposed technique, isolated words having straight headline are not viewed as skewed but rather when length of headline is not as much as a threshold value then the word might be skewed and gets to be focus for correction. The calculation can be effectively connected on Devanagari, Gujarati and Bangla script words as these have an indistinguishable structural properties from Gurumukhi Script. In this paper, a robust technique for skew detection and correction of isolated words of machine printed Gurumukhi documents. The results show that the accuracy of segmentation and character recognition are improved by skew correction.

G S Lehal (2003) et. al: [6] This paper has proposed a projection based technique. They put the vertical and horizontal projection between the angles 0 to 90. The skewing can be detected only when the angle of skew is -180 to 180. This paper has proposed a scheme which provides efficient results for the documents having skewing angle. The document can be portrait or landscape. This paper is based on determination of skew angle with vertical projection at different angles at fixed interval in the range of 0 to 90. Under such projection for image with having no skewing angle the pixel intensity is large. But when there is skewing then the intensity value is varying. This intensity value of lower and upper lines are more compared to the central line.

Flowchart

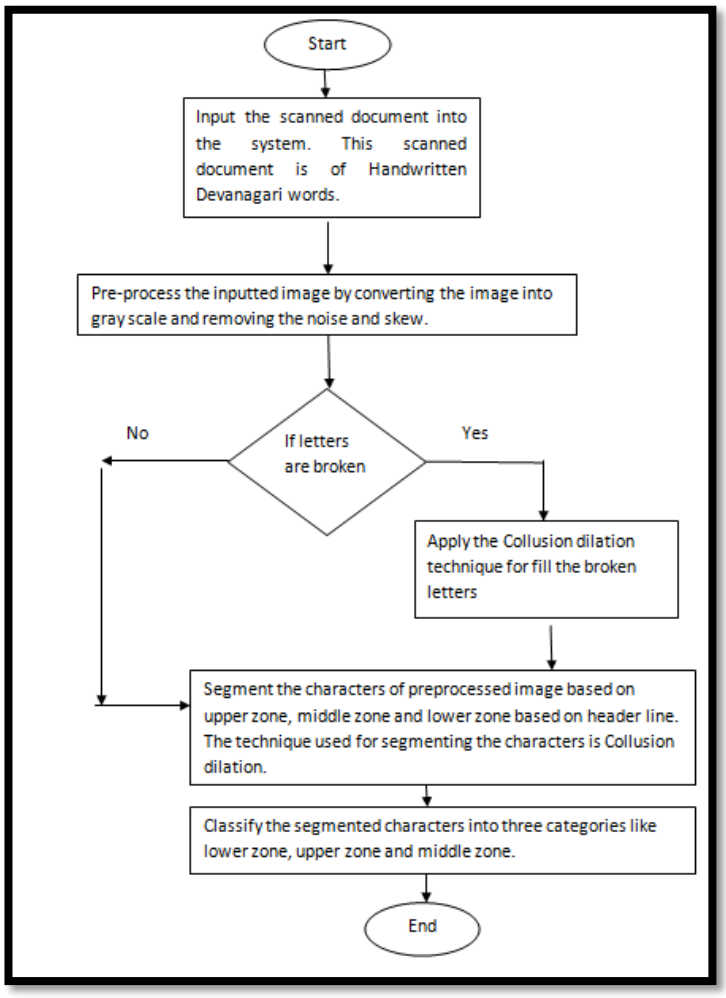


Fig. 5: Flowchart

Algorithm

Step1: Input the digitally scanned image of the text document containing Hindi Devanagari words.

Step2: Preprocess the digitally scanned image. This preprocessing includes resizing, Gray Scale conversion, skew removal etc.

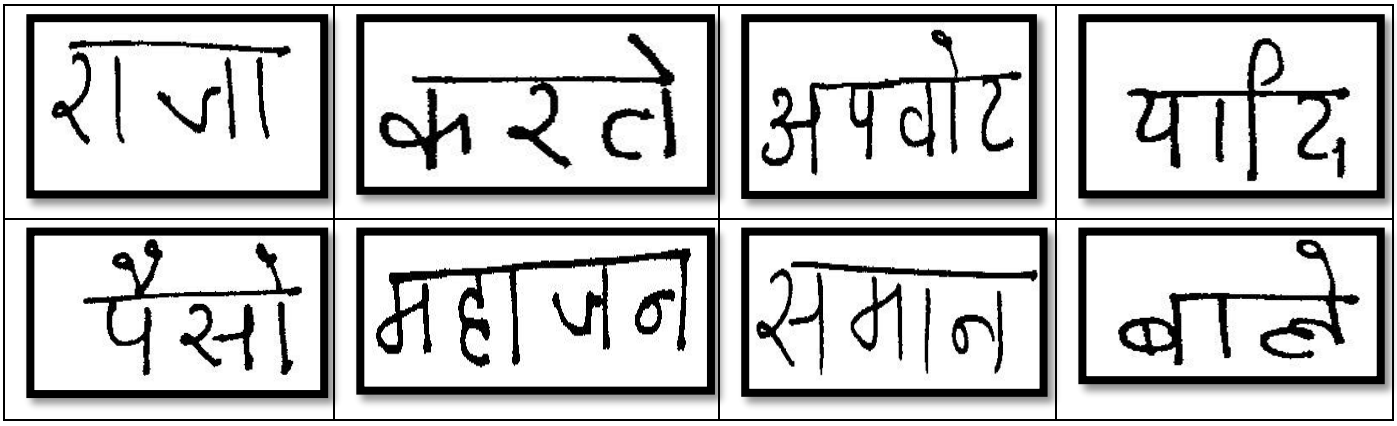
Step3: Check if the letters are broken. If broken go to step 4 else go to step 5

Step4: Fill the broken letters in the handwritten devanagari text.

Step5: Segment the characters in upper zone, middle zone and lower zone based on header line.

Step6: End.

Results Analysis
Dataset Taken



Steps
Step1: In Step1 Devanagari script handwritten document is inputted. This document is digitally scanned document.

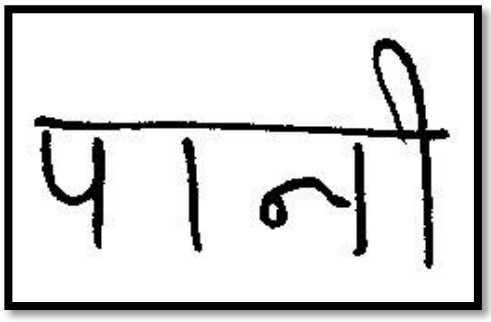


Fig. 6: Broken hand written text

Step2: After pre-process the image, head line is removed and based collision and dilation technique broken letter and identified and segmented.

Fig. 7: segment the character after removing brokenness

Discussions
From the above processing it is clear that the technique used for segmenting the characters from the Devanagari text using the Collusion dilation technique is giving satisfactory results

Accuracy
This algorithm has been applied on to 500 images. Each image contain scanner scanned document. The document is written with hand. The script used for writing text is Devanagari script. Each document that is image is undergoes through pre fixed steps. Based on header line the Hindi text is divided into three sections. One is upper zone, middle zone, and third is bottom zone. Using edge based and intensity variations and the technique of Collusion dilation the characters are recognized. 480 images have given right results. So the total accuracy is around 96%. Few images which do not follow the set criteria have failed to segment.

Comparison to Existing

Author, year	Script Language	Accuracy
Bishnu a.,2012	Bangla	94%
Abdelhak Boukharouba,2016	Arabic	90%
Proposed	Devanagari	96%

Conclusion
CR is optical character reader used for reading the digital scanned documents. These digitally scanned documents can contain the text of handwritten type in Devanagari language. The text in this is broken and skewed. The technique used for extracting the characters from the Devanagari handwritten text is Collusion dilation. The technique has been applied to large amount of dataset. Such that each provides satisfactory results. While performing the segmentation the broken and skewed text are also considered which has been identified at preprocessing phase. The results put up are satisfactory. Such that header line is removed and upper zone characters, middle zone characters and lower zone characters are being recognized.

Future Work
The character extraction from handwritten text of Devanagari script has given the better results for extracting the upper zone, middle zone and lower zone letters. This technique can be further improved when header line has noise. There is unequal width of header line. Also this can be further improved for isolation of middle zone characters with more precision.

References

1. Abdelhak Boukharouba," A new algorithm for skew correction and baseline detection based on the randomized Hough Transform", Journal of King Saud University – Computer and Information Sciences vol.29, 29–38, 2017.
2. Bag, S., Bhowmick, P., Harit, G., Biswas, A.," Character segmentation of handwritten Bangla text by vertex characterization of isothetic covers", National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, vol. 4, pp. 21–24, 2011

3. Bishnu, A., Chaudhuri, B.B.,” Segmentation of Bangla handwritten text into characters by recursive contour Following”, International Conference on Document Analysis and Recognition, vol. 3, pp. 236–239, 2011
4. Casey, R.G., Lecolinet, E.,” A survey of methods and strategies in character segmentation”, IEEE Trans. Pattern Anal., Mach. Intell., vol. 18, pp. 690–706,1996.
5. Dharam Veer Sharma, Gurpreet Singh Lehal,” A Fast Skew Detection and Correction Algorithm for Machine Printed Words in Gurmukhi Script”, ACM digital library, vol. 4, pp.45-55, 2011.
6. G S Lehal, Renu Dhir,” A Range Free Skew Detection Technique for Digitized Gurmukhi Script Documents”, 5th International Conference on Document Analysis and Recognition, vol. 3,pp. 47-152,2012.
7. Hanmandlu, M., Agrawal, P.,” A structural approach for segmentation of handwritten Hindi text”, International Conference on Cognition and Recognition, vol. 3, pp. 589–597, 2005.
8. Huang, L., Wan, G., Liu, C.,” An improved parallel thinning algorithm”, International Conference on Document Analysis and Recognition, vol. 8,pp. 780–783, 2003
9. Lei Huang, Genxun Wan, Changping Liu, "An Improved Parallel Thinning Algorithm", vol. 4,pp. 456-460,2003.
10. M. K. Jindal, R. K. Sharma, G. S. Lehal,” Segmentation of Touching Characters in Upper Zone in Printed Gurmukhi Script”, International Journal of Signal Processing, vol. 2(5),pp. 258-267,2009.
11. Munish Kumar, M. K. Jindal, R. K. Sharma,” Segmentation of Isolated and Touching Characters in Offline Handwritten Gurmukhi Script Recognition”, I.J. Information Technology and Computer Science, vol. 02, pp. 58-63,2014.
12. M. Hanmandlu and Pooja Agrawal, "A Structural Approach for Segmentation of Handwritten Hindi Text", vol. 4 pp.190-200, 2011.
13. Richard G. Casey, "Survey of Methods and Strategies in character segmentation, vol. 18, pp.34-44, 1996.
14. Soumen Bag and Ankit Krishna,” Character Segmentation of Hindi Unconstrained Handwritten Words”, vol. 4, pp.67-77, 2015,
15. R.G. Casey et.al. “A Survey of Methods and Strategies in Character Segmentation”, IEEE Trans. Pattern Analysis And Machine Intelligence, vol. 18, pp. 690-706, 1996.
16. Ravneet Kaur,Ashok Kumar Bathla,"A Review on Different Handwritten Line Segmentation Techniques",wwjmrd,vol.:3(10),pp:148-152,2017.
17. Manas Kaur,Ashok Kumar Bathla,"A Review on Different Character Segmentation Techniques for Handwritten Gurmukhi Scripts",wwjmrd,vol.:3(10),pp:162-166,2017.
18. Ashok Kumar Bathla,"Challenges in recognition of Devanagari Scripts due to segmentation of handwritten text", IEEE,6-18 March 2016