ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

A Review on Character Segmentation of Touching and Half Character in Handwritten Hindi Text

Preeti Sharma
M.Tech Scholar
Department of Computer Science
Sant Longowal Institute of Engineering & Technology
Sangrur(Punjab), 148106, India

Manoj Kumar Sachan
Associate Professor
Department of Computer Science
Sant Longowal Institute of Engineering & Technology
Sangrur(Punjab), 148106, India

Abstract: Segmentation is one of the challenging phases of OCR in case of handwritten text. It has been observed that different languages have different nature of writing the text that causes a number of issues in segmentation aspect. Hindi is one of the official languages of India having the complex structural features. The major problem in handwritten character segmentation is to deal with the touching characters such as conjunct, half-characters, overlapping characters, highly skewed, uneven header lines found in writing pattern of well known Hindi language[]. Segmentation is done on the basis of observed structural properties examined from the writing style of different individuals. This paper reviews different techniques available to dissect the simple, touching, overlapping Hindi text. Also, it discusses the existing techniques with their merits and demerits. It is observed on the basis of related work, few techniques are available for touching and half characters. Hybrid approach including projection profile combined along with the structural property such as width, height, pixel intensity etc is giving quite promising results but still, there exist some challenges due to the constraints that are taken in stated algorithms which need concerned efforts and immense research to improve the segmentation level.

Keywords: character segmentation, touching character, devnagari script, header line modifier, half character

I. INTRODUCTION

Natural Language Processing (NLP) is a field which deals with the interactions between natural languages and computer. This area of Artificial Intelligence is concerned with the processing of human (natural) languages like Hindi, Punjabi, English, Urdu etc. using computer programming. There are many applications developed in past few decades in NLP. Some of the NLP tasks are as script recognition [1, 2, 3], sentiment analysis [4, 5], speech recognition, information this topic to develop more practical and useful systems. Optical character recognition is basically used to digitalize the text data into computer process able format without invoking the extra time manually. The subject of OCR had received considerable attention in recent year. So, in this digital world, the need of OCR is very important to meet the requirement of document image processing and office automation. It will contribute immensely to the advancement of automation process and makes it possible to electronically editing, storing and searching in a document compactly. There are various phases of OCR as shown in fig. 1, i.e. Preprocessing, Segmentation, feature extraction, Classification, Recognition. Segmentation is one of the decision procedure and crucial phase of OCR. It is the process of splitting the different image of the sequence of characters into the individual symbols. Correct segmentation of individual characters decides the accuracy of the recognition system, so they can be recognized correctly.

Hindi is taken to be straight descendent of an early form of Sanskrit. It is written in Devanagari script and considered as official language of great number of states and some union territories of India. The content of Hindi is composed by the consonant, vowels, and modifiers. There are 33 consonant and 12 vowels as shown in fig. 2 and fig. 3 respectively. It is written from left to right and has conjuncts, which is the combination of the full consonant at rightmost and the half consonant at the leftmost which makes the structure of language complex.

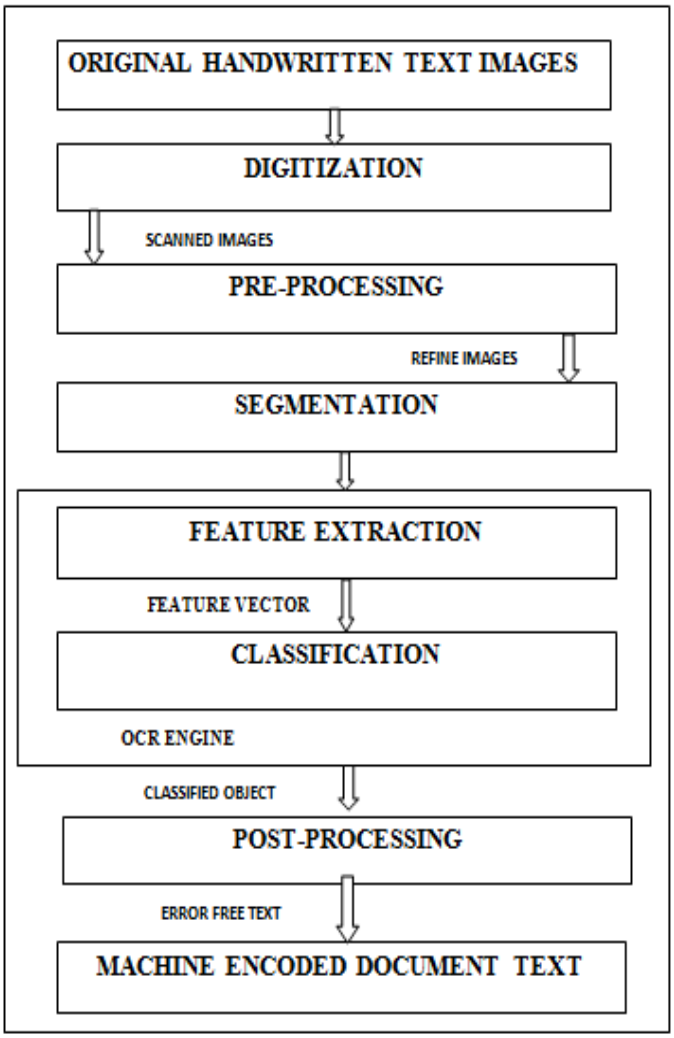


Fig 1: Optical Character Recognition System

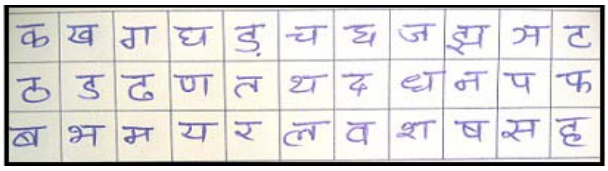


Figure 2: Consonants

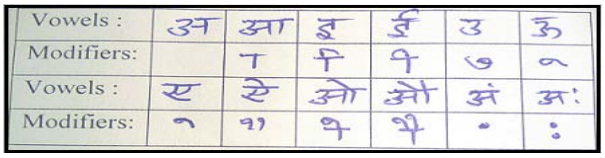


Figure 3: Vowels and corresponding modifiers

Pattern for writing Hindi words

The zone in the Hindi language is shown in fig. 4.The two or more characters joined together to form a word then the horizontal line connect each other and generates the header line called Sheorekha. Modifiers can be written with the consonant at the left, right, top and bottom. The vowel above the Sheorekha is called as ascenders or upper modifier and vowel below the core region is called descenders or lower modifiers. The overall with brief description and images are given by Raghuraj singh [6]

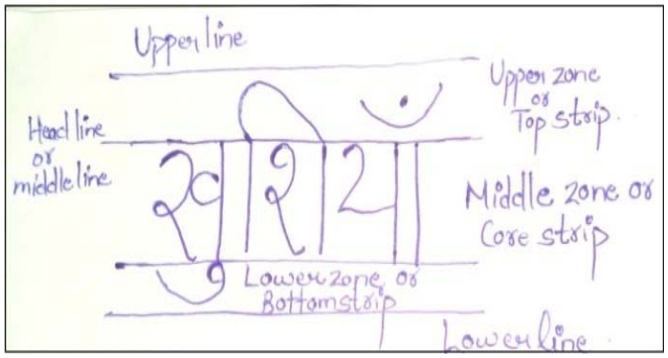


Figure 4: Strips with the header line in Hindi text

II. LITERATURE REVIEW

Srivastav et al. (2016) [7] this paper has Proposed hybrid segmentation scheme based on horizontal projection and a view of bounding box with MATLAB function i.e. region props and rectangle it include scanning the document, noise removed using medfilt2 followed by thinning process and further converted into grayscale and finally into binary image using Otsu's global thresholding method [8]. The stated method efficiently segments the isolated characters and modifiers but couldn't localize the connected component. Results for segmenting the isolated characters found to be approximately 90% accuracy.

Bag et al. (2015) [9] presented an efficient character segmentation method for handwritten Hindi words. The proposed method is divided into three stages where the 1st stage, extracts the header line and outline the upper-strip from the rest. In 2nd stage, statistical information about intermediate individual components is collected and the segmentation of upper modifier is performed. The 3rd stage utilizes this statistical information again to choose the components on which further segmentation needs to be attempted. This separates the lower modifiers from the middle zone components. This proposed method has performed much well at each level of segmentation to manage the large-scale problem of shape variation in the writing style of Hindi language. This method is tested on handwritten Hindi word images and the results were promising with an average accuracy rate of 96.93 %. But this method has some limitation in case of shadow characters occur in the handwritten text when one totally independent component occurs under some other part of component.

Garg et al. (2015) [10] Deals with the most common problem i.e. presence of touching of left modifier with the consonant in the middle region of the word. In Optical Character Recognition (OCR) system, the occurrence of touching character will decrease the recognition rate.

For upper modifier segmentation, position of header line was detected for each word. Vertical projection function is used to remove the lower and upper modifiers of the characters For segmentation of touching characters, structural properties of the text is observed to consider the constraints used for the segmentation. Based on the structural properties of the text, a modified projection profile algorithm is proposed to segment left modifier from the consonant in the middle region of the word. The results obtained with the proposed algorithm are very challenging. Some disadvantages are also related to this technique which occurs due to some remaining ligaments with the consonants, the character shape changes. The features cannot be extracted properly due to these extra pixels that remain with the characters. It is one of the major problems which hinder the recognition of consonant and hence it reduces the recognition rate.

Thakral et al. (2014) [11] Shows a method for segmenting the conjunct and for overlapping characters in Devanagri script. The proposed method applied the cluster detection technique to identify the pixel cluster for touching characters by locating the midpoint which is used to segment the isolated characters. This method gives the segmentation accuracy of 95% for touching and conjunct characters and 88% for overlapping characters. Given technique is further extended on upper and lower modifiers.

Kapoor et al. (2014) [12] Proposed a technique based on categories of characters named as middle, end and no bar character. Joint point's algorithm along with the bounding box which dissects the characters are used in this paper. Above approach will successfully find the joint points between the characters of a text and finally find the vertical and horizontal lines of the characters. The overall segmentation accuracy for handwritten document is 93% and for printed document is 100% but at some points, the handwritten characters are not fragmented properly due to the complexity of vertical bars and some constraints attempted to segment the character including the area, height, and width.

Bhujade et al. (2014) [13] Applied the algorithm based on contour tracing combined with the structural approach for segmenting complex matras and header line, counter is used to returns the position of connected point depends on the neighboring pixels to extract the upper and lower modifier, and to overcome the uncertainty issue, algorithm uses a concept of window to identify the character whether it is simple or joint. Mentioned method works efficiently in different text sizes and different resolution images.

Palakollu et al. (2012) [14] Present the novel procedure for straightening the header line and for overlapping characters, header line is computed based on the difference between actual and expected header line and lastly traced the header line. Above Method successively checked the pixel intensity pattern to find the path till the end of the row to segment the overlapping characters. The overall accuracy for consonant is 89.90%.

Garg et al. (2011) [15] Developed an algorithm by making a statistical assumption of the threshold value to compare height and width of characters to identify the presence of conjunct. This includes scanning successively no of pixels till 70% part of the characters considering some constraint such as three continuous column with one pixel, height of the column and two continuous columns with more than one pixel. Proposed algorithm gives 83% and 87.5% correctness for segmenting applied on both handwritten and printed text images. The main problem in half character separation is the overlapping of half characters.

III. CLASSIFICATION OF CHARACTER SEGMENTATION SYSTEM

Character Segmentation is one of the testing and key fields in Optical Character Recognition process. It is an operation that looks to divide an image of characters used as a final end product. It is one of the decision procedures in OCR which increase the immense research interest. And finally, the end product must be used for automatic processing of huge amount of data such as reading bills like payment slip, generating barcode using the address by reading on an envelope of the post card and for office automation [16]. Character segmentation is classified into two types as shown in fig. 5.

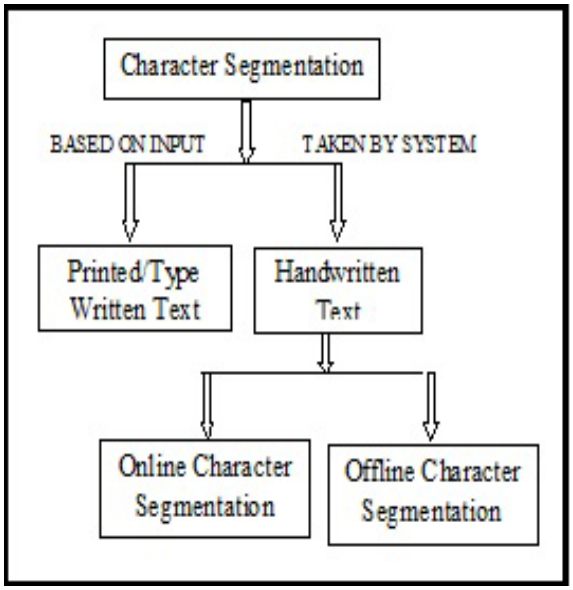


Figure 5: Organization of character segmentation

1. Online character segmentation

It is the system, which provides the interface between the pen and electronic devices through which the pen interact with the electronic surface and gives input as it moves over the devices such as PDA. More meaningful data can be grabbed through the input received by movement of pen such as the sequence of strokes used for character segmentation; therefore pen-based electronic surface has more information to be available to make the segmentation easier [17].

2. Offline character segmentation

It refers to the process of segmenting the words that have been taken as the input from the documents by scanning all images optically and storing in the form of gray scale format. Offline character segmentation is comparatively difficult as different because of the more no of strokes while writing by the pen on the paper. The input which has been taken from papers and documents may be printed and handwritten [17].

Dissection is done mainly on machine printed text and the handwritten text.

Machine Printed Content:

This can be collected from historical documents, newspaper, magazines, books and from the different images, writing posters, paragraph images and there are many complexities in printed character segmentation due to the wide variety of fonts, different text styles and some depends on the image quality as the degraded images. But on the other side, it is comparatively easy to extract the characters from machine printed words because of proper size and uniform alignment.

Handwritten Content:

It is manually written text in particular language. Handwritten can be categorized into two categories: cursive and hand printed script. Recognition of handwritten characters is a much more difficult problem. Characters are non uniform and can vary greatly in size and style. Even the text data written by the same person can vary considerably. In this, the location of characters is not predictable for any one and not even spacing between them [18].

IV. EXISTING DISSECTION TECHNIQUES

The different techniques will mainly used to segment the words into characters are given below

1. Horizontal Projection Profile

For a given binary image of size K*M, K represents the height of the binary image and M represents the width of a binary image. So the horizontal projection is defined as HP (i), i = 1, 2, 3, …, M. which counts the total number of black pixels in the ith horizontal row which is used to locate the Shirorekha[19]. But this method has drawback that some part of the character get cuts in removing the header line and second drawback is to locate the header line because some characters create confusion of largest projections which leads to improper segmentation. These limitations can be overcome by using morphological operation to locate and extract the Shirorekha.

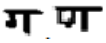
2. Vertical Projection Profile

For a given binary image K*M, K represents the height of the binary image and M represent the width of a binary image. So the vertical projection is defined as VP (j), i = 1, 2, 3, …, M is a function to find the total number of black pixels in the jth vertical column[20]. This will give the better results in the case of segmentation when white space is used as a delimiter, but sometimes it will add the features of other characters or modifiers while segmentation the overlapping characters

3. Joint Point approach using bounding box

The point at which the two characters meet is called joint points**.**

- This approach starts with the categories of characters that is end, middle and no bar characters.
- The method is performed in two phases' word processing phase and character segmentation phase.
- In the first stage, Joint points are identified by evaluating the sum of the 3x3 matrix over each pixel minus pixel towards the midway followed by the condition that if the sum exceeds by 3 then it is referred as joint points otherwise if it equals to 1 then it is considered as terminating point.
- Bounding box is formed in order to extract the character from the word.

- To determine the header line, small horizontal rectangles are formed throughout the word based on the ratio of width and height using joint point. The line on one side which is having the max no of pixels, that rectangle is identified as header line.
  - In the same way, vertical rectangles over the word are formed and identified by the region covered by the white pixels in the vertical rectangle if it exceeds 60% then the presence of a vertical bar is analyzed.
  - Second is the character processing phase, character are segmented through bounding box after removing the header line and vertical bar. In mid bar touching character segment at the leftmost joint point of the bounding box, in sidebar touching character segment after each identified vertical bar and lastly in no bar segment towards the right of a bounding box.
  - Constraints regarding height and width are the main concern in this algorithm.
  - And this will not work in given characters such as  and in the case of overlapping characters [12].

4. Pixel cluster detection techniques

Cluster brings some pixels together to make the heap of pixels. After removal of header line midpoint of the gap between the character of the word is calculated to segment the characters if the difference between the two midpoints exceed the expected value that is to be assumed by the study of experiments then it represent the presence of cluster and further that cluster needs to be identified by scanning vertically and horizontally .At the instant cluster is identified, then at that point of cluster segment the character by adding the vertical line. This algorithm makes an assumption that characters will have the consistent size [11].

5. Half character structural property based approach

In this approach presence of half character or conjunct are detected on the basis of threshold value. By observing an experimental value or making a statistical assumption that width is greater than 1.65times of height, where c is the no of the column, r is the no of row.

1. If in case two characters are found to be touching then scan 70% part of the character from r/7 to nth row and stored the values in matrix say 'v'.
2. From the leftmost pixel check the no of a pixel in two continuous columns whose column position less than c/5.if the no of the pixel is greater than 1, then continuously checking with the same column position till we get the single pixel. Likely conditions are checked for the column position greater than c/5.
3. Both of the conditions are satisfied and column with more than one pixel encountered again then column position is between c/4 and c/7. And that value is used for separation of half characters. To avoid over-segmentation following three conditions needs to be satisfied [15].

6. Projection profile with bounding box

For a binary image, this method is used, invert the binary image and scan and compute a sum of no of black pixels for each row and store values. To remove the header line threshold value is compared with the computed sum and the row found to have the value greater than the threshold then it is set to 0**.** And find the label connected component and measure the properties of the image region and finally bounding box is applied using the MATLAB function Region props and rectangle to extract the characters [7].

7. Skewed header line detection Method

The method outlines global and local density row. Find the span of words. Scan the upper 50% part of the word to get the highest density pixel marked as global density row and divide the words vertically into no of strips to find the local max density row of the following vertical strips of characters till 50% part of the word. Calculate the difference between the global max and local max to adjust the global max and finally remove the header line [9].

8. Contour tracing algorithm

This is used to locate the point of intersection the row at which the first transition occur, record the position of the pixel and passed on to the contour which returns the position of all the pixels connected to the input around its 3X3 neighbors which helps to detect and subsequently segment the modifier [13].

V. ISSUES ASSOCIATED WITH STRUCTURE OF HINDI LANGUAGE

As the different individual have the different way of writing the text so there are many irregularities due to the varied shapes and size of characters in a text which can produce many complexities and makes the job of segmentation very challenging. There are many issues occur in handwritten character segmentation [21]. Issues are:

A. Words having touching characters.
B. Irregular ascenders and descenders.
C. The presence of skewed and uneven header lines.
D. Overlapping characters in middle zone.
E. Problems of missing characters

A. Words containing touching characters

This problem occurs due to different writing pattern of different people. While writing the text one touch the other character as shown in fig6 that generates problem to recognize whether it is a single character or touched or conjunct.

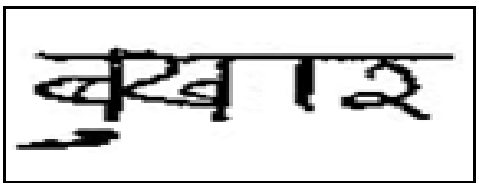


Figure 6: touching character

B. Irregular ascenders and descenders

Frequently writing style generates distinct shaped modifiers or matras and makes difficult to perform the proper segmentation. In fig7 Problems like the matras feature in some characters, upper modifier touches each other, large size of upper modifier. Extraction and localization of these mantras are the major problems.

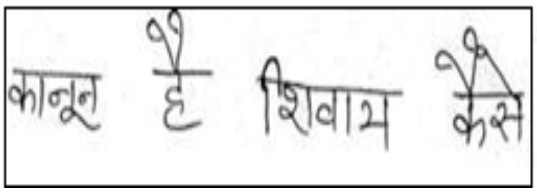


Figure 7: Size of lower and upper modifier and including features

C. Presence of skewed and uneven header lines

Header line is the most visible part of the text and continuous horizontal strip of black pixel lie at the top of the character but because of the cursive property in writing style, words are written non-uniformly and in skewed manner that makes the header line highly slanted upward or descending and create the issue in recognizing the header line as shown in fig. 8.

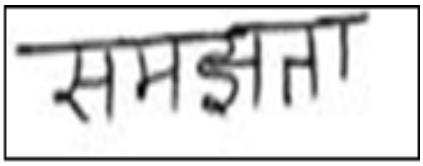


Figure 8: Skewed and uneven header line

D. Overlapping characters in middle zone

Because of the rushed handwriting of different individual sometimes one character pasted on other as shown in fig. 9, one character is written in such a way that it is having the feature of other character i.e. mixed with other character create a problem, segmented character consist of some parts of another character.

Figure 9: Overlapping characters

E. The problem of missing characters

Broken character in Hindi is mainly found in the middle zone as shown in fig. 10. A character can be broken due to writer's pens and page quality used. The vertical projection profile function found the character is broken then it will segment one character into more than one parts i.e. it will segment the broken part as the individual character which is the main problem while recognition.

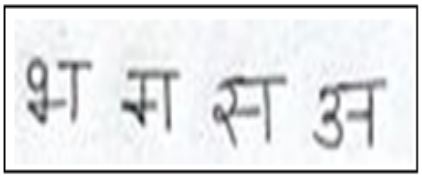


Figure 10: Broken characters

VI. CONCLUSION & FUTURE SCOPE

Handwritten character segmentation is a challenging job because of the variation in the structure of different characters in Hindi text. Already defined techniques are visualize in table1and have been summarized in this paper such as modified projection profile, joint point algorithm, structural based approach, pixel cluster detection technique, contour tracing method. These mentioned methods have some limitations because of the varying shape and size therefore they do not give the accurate result. By the study of related work, it has been observed that the above-mentioned techniques are not quite appropriate because of large variations in writings pattern of different people which needs to be handled in future.

In future, there is a wide scope to improve the existing techniques considering the constraints assumed on the basis of structural properties of Hindi text and little work has been done in a case of overlapping, half, broken and touching characters which can be extended further.

Table 1: Comparative study of different character segmentation techniques of Hindi text

Contributors	Approach /Method Used	Nature of Input Data	Limitation	Segmentation Accuracy
Srivastav et al. (2016) [1]	Projection profile with bounding box using two MATLAB function	Simple handwritten text	connected component cannot be segmented	90%
Thakral et al. (2014) [4]	Pixel cluster detection techniques	Isolated, touching, Conjunct, overlapping Words	Assumption of character having consistent size	94%
Kapoor et al. (2014) [5]	Joint point algorithm	Simple, touching Words categorized on middle, end and no bar both handwritten and printed	constraints like gap between parts of character ,height of vertical bars etc	71% for touching, 93% for non touching
Bhujade et al. (2014) [6]	Contour tracing methods	Handwritten Hindi text images		
Garg et al. (2011) [8]	Modified projection profile based algorithm	Words containing half characters both printed and handwritten	Overlapping of half character with the full character	83.02% for handwritten , 87% for printed

VII. REFERENCES

[1] Singh, Gurpreet, and Manoj Sachan. "Multi-layer perceptron (MLP) neural network technique for offline handwritten Gurmukhi character recognition." *Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on*. IEEE, 2014.

[2] Singh, Gurpreet, and Manoj Sachan. "Offline Gurmukhi script recognition using knowledge based approach & Multi-Layered Perceptron neural network." *Signal Processing, Computing and Control (ISPCC), 2015 International Conference on*. IEEE, 2015.

[3] Singh, Gurpreet, and Manoj Kumar Sachan. "Data capturing process for online Gurmukhi script recognition system." *Computational Intelligence and*

Computing Research (ICCIC), 2015 IEEE International Conference on. IEEE, 2015.

[4] Singh, Shailendra Kumar, Sanchita Paul, and Dhananjay Kumar. "Sentiment Analysis Approaches on Different Data set Domain: Survey." *International Journal of Database Theory and Application* 7.5 (2014): 39-50.

[5] Singh, Shailendra Kumar, and Sanchita Paul. "Sentiment Analysis of Social Issues and Sentiment Score Calculation of Negative Prefixes." *International Journal of Applied Engineering Research* 10.55: 2015.

[6] Singh, Raghuraj, et al. "Optical character recognition (OCR) for printed devnagari script using artificial neural network." *International Journal of Computer Science & Communication* 1.1 (2010): 91-95.

[7] Srivastav, Ankita, and Neha Sahu. "Segmentation of Devanagari Handwritten Characters." *International Journal of Computer Applications* 142.14 (2016).

[8] Otsu, Nobuyuki. "A threshold selection method from gray-level histograms." *Automatica* 11.285-296 (1975): 23-27

[9] Bag, Soumen, and Ankit Krishna. "Character Segmentation of Hindi Unconstrained Handwritten Words." *International Workshop on Combinatorial Image Analysis*. Springer International Publishing, 2015.

[10] Garg, Naresh Kumar, Lakhwinder Kaur, and M. K. Jindal. "Segmentation of touching modifiers and consonants in middle region of handwritten Hindi text." *Pattern Recognition and Image Analysis* 25.3 (2015): 413-417.

[11] Thakral, B. and Kumar, M., 2014, October. Devanagari handwritten text segmentation for overlapping and conjunct characters-A proficient technique. In *Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions), 2014 3rd International Conference on* (pp. 1-4). IEEE.

[12] Kapoor, Shuchi, and Vivek Verma. "Fragmentation of handwritten touching characters in devanagari script." *International Journal of Information Technology, Modeling and Computing (IJITMC) Vol* 2 (2014): 11-21.

[13] *Bhujade, Ms Vaishali G., and Ms Chhaya M. Meshram. "A Technique for Segmentation of Handwritten Hindi Text." International Journal of Engineering Research and Technology. Vol. 3. No. 2 (February-2014). ESRSA Publications, 2014.*

[14] Palakollu, Saiprakash, Renu Dhir, and Rajneesh Rani. "Handwritten Hindi text segmentation techniques for lines and characters." *Proceedings of the World Congress on Engineering and Computer Science*. Vol. 1. 2012.

[15] Garg, Naresh Kumar, Lakhwinder Kaur, and M. K. Jindal. "The segmentation of half characters in handwritten Hindi text." *Information Systems for Indian Languages*. Springer Berlin Heidelberg, 2011. 48-53.

[16] Syama, K., et al. "Performance study of active contour model based character segmentation with nonlinear diffusion." *Advances in Computing and Communications (ICACC), 2012 International Conference on*. IEEE, 2012.

[17] Dalbir and Singh S "review of oniline and offline character recognition" international journal of engineering and computer science vol. 4(2015):11729-11732.

[18] Kaur, Karamjeet, and Ashok Kumar Bathla. "A Review on Segmentation of Touching and Broken Characters for Handwritten Gurmukhi Script." *International Journal of Computer Applications* 120.18 (2015).

[19] Bansal, Veena, and R. M. K. Sinha. "Segmentation of touching and fused Devanagari characters." *Pattern recognition* 35.4 (2002): 875-893.

[20] Kumar, Munish, M. K. Jindal, and R. K. Sharma. "Segmentation of isolated and touching characters in offline handwritten gurmukhi script recognition." *International Journal of Information Technology and Computer Science (IJITCS)* 6.2 (2014): 58.

[21] Kaur, Lakhwinder. "The hazards in segmentation of handwritten hindi text." (2011).