WWJMRD2017; 3(10): 143-147
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, Yadawindra
College of Engineering,
Talwandi Sabo, Punjab, India

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

Correspondence:
Manpreet Kaur
Research Scholar, Yadawindra
College of Engineering,
Talwandi Sabo, Punjab, India

A Review on Different Techniques for Character Segmentation for HINDI Handwritten Words

Manpreet Kaur, Ashok Bathla

Abstract

The character level segmentation of printed or hand written text is a important step for OCR. It will be very difficult to segment the word to extract the characters when writing is cursive. There are various number of other problems that lies in the HINDI words. For example there is a problem of overlapping and cutting and upper line, lower line problems. The most emphasis point here is the segmentation of characters with different types of writing style of different persons under different moods. There are various methods and techniques are being followed to extract the text after the segmentation. No technique remains complete as far as exact result extraction is concerned. Upper line segmentation based technique is the better technique to extract the text while considering the overlapping and cutting of the text.

Keywords: Header Line detection, Character segmentation, Structural approach

Introduction

The separation of handwritten and machine written text is the most required issue as far as research for both type of text is concerned. Most difficult task for processing of the document is the programmed adding of the text contained in the document. Automatic programmed reading of optically scanned document is called as OCR. The purpose of OCR is to read the optically scanned document and convert that to the machine readable format. Till the time only success has been achieved is while getting into the single font where large amount of text is written with uniformed single font. There is less accuracy when the document is with multiple fonts. In OCR first the lines of Inline text is read from the document and separated. Once the lines are being segmented there requires the word segmentation and then it leads to the character segmentation. Various languages segmentation is at the advance stage of the working where the accuracy is up till 90%. This type of accuracy has been achieved in Roman letters. But subsequently the activity can be enhanced further. For machine written text this accuracy can be 100%. [1].

A traditional OCR is having various steps in common. The first and foremost step is to convert the paper written document in digital format. This digital format of the document is generated with the help of scanned convert the document. Later on ROI is identified where lines, words and then letters are being segmented. This process will go on till the final result can be obtained. Once the text characters are being recognized, it requires de-noising the text so that the exactness of the read characters is being ensured [2].

Today, many researchers can be focusing their researches in the field of Devnagari script character recognition. But it is very difficult to make an efficient artificial intelligent machine which can understand all the characters in case of free flow handwritten words. Various people write the words in their own ways. The previously done researches are more focused on this type of OCR with less efficiency.

Handwritten text recognition in Devanagri script can be differentiated into two categories. One is online handwritten character recognition, and second is offline character recognition. In online character recognition the hand written 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 Devnagari script characters is a goal of many research efforts in the pattern recognition field. Many techniques has 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.

OCR as optical character recognition includes series of steps. One single step will not be able to understand the all characters. These steps are like scanning the document, pre-processing the document, segmentation.

Neural network works as human brain, thus the architecture of neural network machine is same as the architecture of human brain. There are hundreds or even thousands of neurons in custom integrated circuits. In conjunction, the growing interest in learning machines, non-linear dynamics and parallel computation spurred renewed attention in artificial neural networks. There are too many applications, which can be solved easily using neural networks that are difficult to solve using traditional methods [2].

Overview of Devanagari Script

Devnagari script is employed to put in writing Hindi language in geographic area, a northern state of Asian country Devnagari script was standardized in eleventh century. Devnagari script consists of thirty six basic consonants, 6 special consonants, 12 vowals, 3 0.5characters and ten numerals. There's no idea of higher or minuscule characters and characters are unremarkably aligned below the road of writing. All Devanagari characters have a full/partial horizontal line (except numerals) at the highest referred to as headline [3].

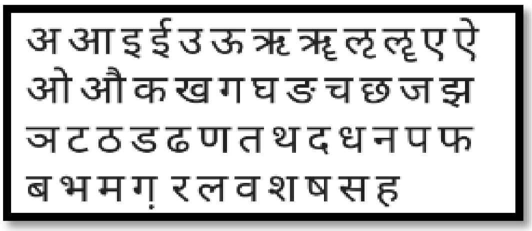
अ आ इ ई उ ऊ ऋ ॠ ऌ ॡ ए ऐ
ओ औ क ख ग घ ङ च छ ज झ
ञ ट ठ ड ढ ण त थ द ध न प फ
ब भ म ग़ र ल व श ष स ह

Fig. 1.1: Devanagari regular characters [1**]**

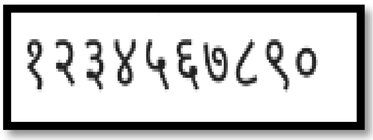
१२३४५६७८९०

Fig.1.2: Devanagari Numerals [1]

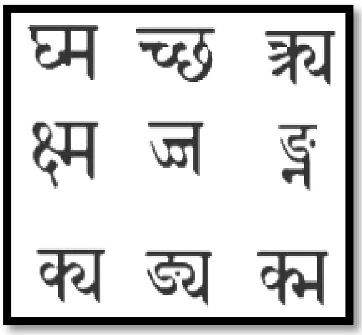
घ्म च्छ
क्ष्म
क्य क्म

Fig. 1.3: Combined Characters [2]

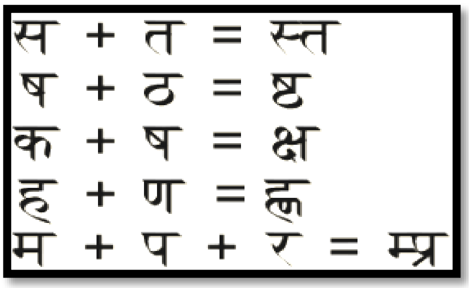
स + त = स्त
ष + ठ = ष्ठ
क + ष = क्ष
ह + ण = ह्ण
म + प + र = म्प्र

Fig. 1.4: additional characters [1]

Steps Followed

a. Optical Scanning: It is first step towards the reading process to be started. In this phase various text are being read using optically scanned documents. That means the documents which are paper written are to be read will be focused on to the reading process and scan convert the paper document using optical reader [3].

b. Local Segmentation: the process of segmentation is to read the document with the help of optically scanned document and identifying the region of interest. This will read that part of the text where exactly that text letters are being there. These segmented letters are being used in various real time applications like mail sorting and other purpose itself.

c. Pre-processing: for successful further steps there requires the preprocessing phase. This is required because while digitally scanned documents can have various types of noises. These noised are like skewness and blurriness. While preprocessing these problems are being identified and removed with those problems [4].

d. Feature Extraction: this phase involves the extraction of the features. These features are there required as the most important aspect as far as system is concerned. Now in our case this is to extract the characters from the text. These characters are devanagari text. The techniques for extraction of such features are regularly divided into three main groups, where the features are found from:
   - The distribution of points.
   - Transformations and series expansions.
   - Structural analysis.

e. Recognition Post-Processing: once all the characters are being recognized from ROI there requires classification of the characters. We can put the characters into subsequently three categories. These three categories are called classes. These methods are utilized when the description of the character can be numerically represented in a feature vector.

Levels of Text Segmentation

The digitally scanned document includes three stages of segmentation. In first segmentation phase lines are being identified from identified ROI. These lines are the text lines. Later on each line different words are being separated. These words are segmented using vertically scanning technique. From the words each character is recognized. This character is actually the required product. Whose categorization is to be performed?

i. Line Segmentation: Line segmentation is the first and a preliminary step for text-based image segmentation. It incorporates horizontal scanning of the image, pixel-push by pixel-push from left to right and top to bottom. At every pixel the intensity is tested. Contingent upon the values of the pixels we gather pixels into multiple regions from the entire image. The diverse region indicates distinctive content in the image file. Thusly the desired content can be extracted. Because of inaccuracies in the scanning process and writing style, the writing might be marginally tilted or inside the image. This can hurt the adequacy of later algorithms and, along these lines, ought to be detected and corrected.

ii. Word Segmentation: Word segmentation is the following level of segmentation. It incorporates vertical scanning of the image, pixel-push by pixel-push from left to right and top to bottom. At every pixel the intensity is tested. Contingent upon the values of the pixels we gather pixels into multiple regions from the entire image. The distinctive region indicates diverse content in the image file. Hence the desired content can be extracted. Slant angle estimation is utilized to perform skew correction for the extracted word in heavy noise. The skew correction can be performed by deciding the angle and rotating the image in the opposite direction.

iii. Character Segmentation: Character segmentation is the final level for text based image segmentation. It is like in operations as word segmentation. A couple of precautions ought to be taken after while performing character segmentation. The segments are not accurate, as "h" is extracted as "l" and "i". Such errors are undesirable. Another precaution is of ligatures. On the off chance that the text image contains a cursive sort textual style then while segmenting the ligature ought to be separated for better efficiency.

Classification

Image classification is a complex process and might be affected by many factors. The classification of document being processed is required for their efficient recognition as it reduces number of searches, simple recognition of document and furthermore reduces the possibility of error at various stages during processing. A classifier connects the document with class; labeling an observed document image as indicated by the class, region into which it falls. The classification stage identifies each input document image by considering the distinguished features like spatial arrangements with respect to each other, layout of document, size of the document, color of the paper, texture. The categorization (Indexing) of images significantly enhance the performance of document by filtering out the relevant document and the class to which it belongs. Classification of fundamental document is done initially followed by the sub-sections. A class prototype is put away in knowledge base.the approaching document is assigned to one of the classes, depending on the value of measure of closeness with the class prototype. This value is obtained by comparison of document under review and the class prototype. The document is assigned to the class with which highest value of the measurement is obtained.

Document classification is an important task in document Processing.

- Document classification will help in identify the text written direction. And also will be useful for extracting the required features. These features are in the form of type of document recognition. Like postal letter or money order. And send the appropriate document to appropriate department.
- Document classification helps in indexing of the library books. Type of book will helps in recognizing the department. And put the index based on the department. Later on is will be easy to locate the books department wise. It will be having efficient searching time.
- Document classification also helps in fetching the appropriate document whose demand is there. Generally those documents will be fetched which are close to the demand. First that paper is fetched where Table of contents if mentioned. Then later on appropriate category will be generated.

Literature Survey

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 unclock wise fachion.

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. This zone provides 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 Devnagari, 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.

Lei Huang (2003) et.al: [9] this paper has proposed an algorithm based on improved thinning algorithm. This algorithm proposed an algorithm based on contour and skeleton of the pattern. This paper has proposed a mechanism for saving the information loss. This paper has proposed the approach based on parallel thinning algorithm. According to this paper they have sub divided the total image into four segments. The 4-connected components of non-object pixels which contain the top and bottom rows and rightmost and leftmost columns of the image are the background. The skeleton leg is a leg in the limb of thickness. This pixel is not connected to any pixel. This pixel is called end pixel. It also put two pixel width processes. Each pixel width is determined based on their height and width.

Comparative Analysis

Paper Name	Author, year	Technique Followed	Constraints	Accuracy
Character Segmentation of Hindi Unconstrained Handwritten Words	Soumen Bag,2016	segmentation process extracts the header line and delineates the upper-strip from the rest	There are multiple cases which are still un- addressed. Which need to be addressed. It have not touched overlapped characters, broken letters etc.	93%
A Survey on Methods and Strategies on Touched Characters Segmentation	Tanzila Saba,2010	This paper has studied different techniques that can be used to extract the text letter which are of touching characters. They have studied the techniques for both numeric and English letters.	No technique is found complete which can extract the letters. But there can be scope if total linguistic of the language are available.	N.A.
Segmentation of Touching Characters in Upper Zone in Printed Gurmukhi Script	M. K. Jindal,2009	This paper has worked on the technique of segmenting the touching characters in upper zone in Punjabi gurmukhi script characters.	This technique of upper zone touching character provides the accuracy of 91%. Which can be further enhanced.	94%
Segmentation of Isolated and Touching Characters in Offline Handwritten Gurmukhi Script Recognition	M. Kumar,2014	It has put water reservoir based technique to run from the top. And fill the empty spaces with black dots.	This technique may not be suitable for text with broken letters. Because it will over flow the water in water reservoir technique.	92%
A Fast Skew Detection and Correction Algorithm for Machine Printed Words in Gurmukhi Script	Dharam Veer Sharma,2009	This paper has given the solution in two phases. In first phase the text is segmented and in second phase the text is skewed to normal.	The limitation of the algorithm is pepper noise in the document, which may lead to wrong calculation of height and width of the image or wrong detection of headline.	94%
Character Segmentation of Handwritten Bangla Text by Vertex Characterization of Isothetic Covers	Gaurav Harit,2011	It is based on the vertex characterization of outer isothetic polygonal covers so that each cover corresponds to a particular word or part of a word.	Further significant preprocessing step towards the development of an integrated handwritten OCR system.	92%.

Conclusion

In current review various papers are studied based on different techniques for text segmentations. They have applied different techniques on different languages. Each language has different dynamics and constraints. Each technique followed by the author does not provide the complete solution. Means each technique has constraints put and remain there while writing hand written Hindi text. Major two techniques are being used which provides better results upto 91%. One technique is water reservoir technique and other is upper line segmentation. Water reservoir technique is more suitable for those character sequence where there is overlapping but having least broken letters. Other technique based on upper line segmentation provides better results also solve various constraints like overlapping, broken letters etc.

References

1. Abdelhak Boukharouba,” A new algorithm for skew correction and baseline detection based on the randomized Hough Transform”, 2017, Journal of King Saud University – Computer and Information Sciences29, 29–38
2. Bag, S., Bhowmick, P., Harit, G., Biswas, A.,” Character segmentation of handwritten Bangla text by vertex characterization of isothetic covers”, 2011, National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, pp. 21–24. IEEE Press, New York
3. Bishnu, A., Chaudhuri, B.B.,” Segmentation of Bangla handwritten text into characters by recursive contour Following”, 1999, International Conference on Document Analysis and Recognition, pp. 236–239, IEEE Press

4. Casey, R.G., Lecolinet, E.," A survey of methods and strategies in character segmentation", 1996, IEEE Trans. Pattern Anal., Mach. Intell., 18, 690–706
5. Dharam Veer Sharma, Gurpreet Singh Lehal," A Fast Skew Detection and Correction Algorithm for Machine Printed Words in Gurmukhi Script", 2009, ACM digital library
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, 47-152
7. Hanmandlu, M., Agrawal, P.," A structural approach for segmentation of handwritten Hindi text", 2005, International Conference on Cognition and Recognition, pp. 589–597
8. Huang, L., Wan, G., Liu, C.," An improved parallel thinning algorithm", 2003, International Conference on Document Analysis and Recognition, pp. 780–783, IEEE Press
9. M. K. Jindal, R. K. Sharma, G. S. Lehal," Segmentation of Touching Characters in Upper Zone in Printed Gurmukhi Script", 2009, International Journal of Signal Processing, 2(5):258-267
10. Munish Kumar, M. K. Jindal, R. K. Sharma," Segmentation of Isolated and Touching Characters in Offline Handwritten Gurmukhi Script Recognition", 2014, I.J. Information Technology and Computer Science, 02, 58-63
11. Soumen Bag and Ankit Krishna," Character Segmentation of Hindi Unconstrained Handwritten Words", 2015, IWCIA
12. Richard G. Casey,"Srvey of Methods and Strategies in chraacter segmentation,vo. 18,1996.
13. M. Hanmandlu and Pooja Agrawal,"A Structural Approach for Segmentation of Handwritten Hindi Text",vol. 4 pp.190-200,2011.
14. Lei Huang, Genxun Wan, Changping Liu,"An Improved Parallel Thinning Algorithm", vol. 4, pp-. 456-460,2003.
15. G S Lehal, Renu Dhir,"A Range Free Skew Detection Technique for Digitized Gurmukhi Script Documents",vol. 3,pp-345-350,2004.