Optical Segmentation of Devanagari Script:A Scientific Analysis

Ipsita Pattnaik
MTech Computer Science
C-DAC
Noida, India

Tushar Patnaik
The Joint-Director
C-DAC
Noida, India

***Abstract*—In the subject Artificial Intelligence, OCR is of keen interest for the computer scientists and researchers. The reason being that the OCR helps in digitizing the printed or handwritten document into computer readable format. These digitizing of document have various futuristic use such as record of old document could be kept in track, multiple data storage of handwritten or printed document, etc. A large number of articles have been published in this area in various journals. A few works have been done in Devanagari Script. The Devanagari Script makes a full set on many other scripts like Hindi, Konkani, Marathi, Nepali, Sanskrit, Bodo, Dogri and Maithili. Challenging rate of OCR in Devanagari Script have not been solving in a better rate. Hence as effort have been generated to achieve a better rate of Segmentation in this Script. Objectives of this study are to find out the accuracy level of word and characters segmentation of Devanagari Script and then to analyze the Challenges faced during the process of word and character segmentation. The finding shows that the accuracy rate of character Segmentation is 89% and word segmentation is 91.70%.**

Keywords*— *Shirorekha; Devanagari; Printed; OCR; Conjunct Characters

I. INTRODUCTION

Machine simulation of human activities has been the most challenging research area since the evolution of digital computers. The main reason for such an effort was not only the challenges in simulating human reading, but also the possibility of methodical applications in which the information present on paper documents has to be transferred into machine editable form. OCR is a process of automatic computer recognition of characters and symbols in optically scanned and digitized pages of text [1]. Automatic recognition of information present on documents like cheques, envelopes, forms, and other manuscripts has a numerous practical and commercial applications in banks, post offices, library, publication houses, language processing, and forensic investigation. [2] Many works on Hindi Script have been reported due to the complications in Devanagari Scripts the result of OCR has been a challenging task. Hindi is the Official Language of 300 million people. Hindi uses Devanagari Script. Devanagari Script is an old one and evolved from the Brahmi script. Devanagari is used to write many languages such as Hindi, Konkani, Marathi, Nepali, Sanskrit, Bodo, Dogri and Maithili. [2] The Devanagari Script consist of many constraints which makes the OCR procedure challenging. The Optical Character Recognition undergoes many techniques and procedure to achieve a valuable result. The main step of OCR comprises of (i) Pre-processing in which the data is been cleaned up which includes Noise Removal, Skew Detection, Binarization and Grayscale Conversion. (ii) Segmentation in which various segmentation layers are undergone which is sub-diving into – Text Segmentation, Line Segmentation, Word Segmentation and finally to Character segmentation. This Segmentation is the most essential step as the Recognition rate are depended on them. (iii) Feature Extraction is the important step as after the Segmentation procedure the whole Feature Extraction of the particular data are depended on them as the particular feature on the character are extracted and then classified to get the recognized result.

II. LITERATURE SURVEY

The Recognition rate of OCR of Devanagari script depends on the Segmentation rate of Characters in a Script. Richard, Casey [3], proposed the basic segmentation strategies and algorithms based on four basic approaches- Classical, Recognition-based, Holistic and Hybrid approach. Veena Bansal, RMK Sinha [4] used the structural properties example height, width, shape and size of character addition with Collapsed horizontal projection to segment the conjunct characters that leads to the total 83% of accuracy. Garg, L. Kaur, M.K. Jindal [5] has given algorithm for segmenting lines, words and characters in Hindi language. They used projection profile approaches for segmenting characters and leads to 79.12% accuracy for simple characters; this algorithm doesn't give good results for half and touching characters. To which the Water reservoir principle for detection of touching bangle text along with projection profile methods are used in [6] given by U. Pal, S. Dutta for segmentation of lines, words and characters. The proposed algorithm gives 97% efficiency for non multi-touching points between characters. M.K. Jindal, R.K. Sharma, and G.S. Lehal in [7] for segmenting horizontally overlapping lines for printed text in Gurmukhi Script. They have used Horizontal projection profile for detecting and removing header line and average height of lines for segmenting overlapping characters, upper and lower modifiers. Dipak, Sharvari in [8] proposed joint point algorithm for segmenting touching characters of Marathi language. Bounding boxes are applied for on each horizontal and vertical line of a character to extract touching characters. Saiprakash, Renu, Rajneesh [9] proposed an algorithm based on based structural properties of Hindi language to segment lines and characters. 89.90% efficiency for characters and 93.6% for lines has been achieved.

III. OBJECTIVE

Devanagari Script being a Challenging and complicate language for OCR, the study has been conducted keeping following objectives in considerations:

(i) To find out the accuracy level of word and characters segmentation of Devanagari Script; and

(ii) To find out the challenging issues faced during the segmentation process of word and character segmentation of the Devanagari Script.

IV. FEATURES OF DEVNAGARI SCRIPT

Devanagari Script consists of 12 vowels and 34 constants. In Hindi, vowels are used in two ways: firstly, to produce their own sounds and secondly to modify the sound of a consonant. [3] The sound of the constant is being changed by applying modifiers in appropriate manner in constants of the script. The modifiers are of three types: the first one is the lower modifiers which is placed below the constants , second is upper modifiers which is placed above the constants and last one is core modifiers which are placed before or after the constants in order to deal with the dependencies of sounds .

In Devanagari Literature the constants are written in pure form and also at times the constants are merged into other characters making it touch or fused into other characters. This is called as conjuncts characters. The Shirorekha used is also called as header line. It is the main feature of Devanagari Script which makes it challenging to conduct OCR on the Devanagari Script. They are horizontal lines drawn on top of all characters of a word. The example is given in Figure-1.

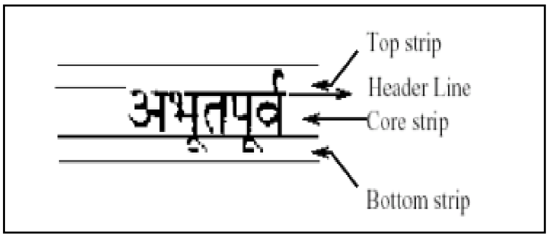


Fig.1. Zone of Devanagari Word

V. METHODOLOGY

The At the outset, document was Scanned at standard DPI format at 300 dpi form. After that Pre-processing procedure was applied to make it clean and easy for the application of further procedures. Segmentation procedure are applied with two procedure i.e. Contouring and Contouring with Horizontal Projection Profile. The procedure is explained below with a flow chart.

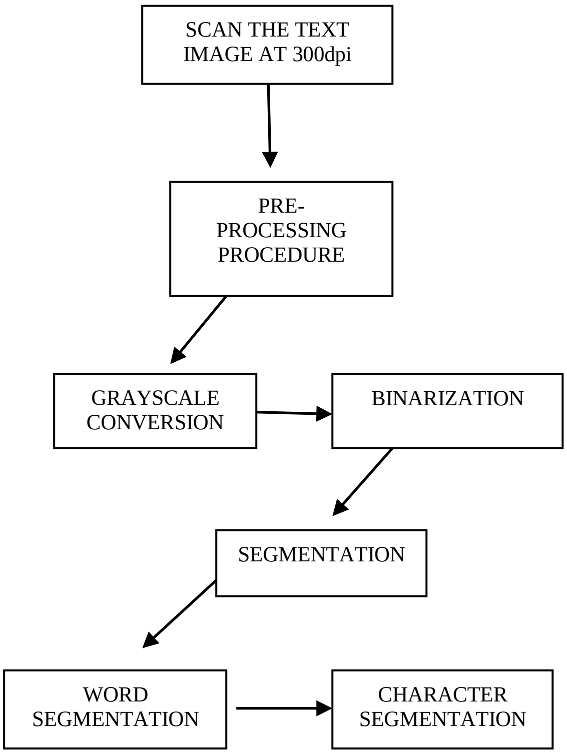
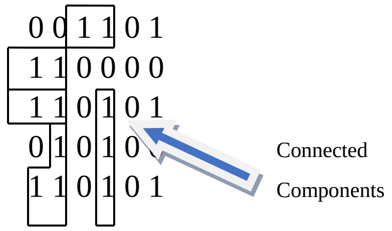


Fig.2. Proposed Structure of the System

A. PRE- PROCESSING – Pre-processing consist of sub-process that helps to clean the document for segmentation process. All phases of Pre-Processing have been shown in the diagram. Colored Image have been converted to Grayscale image. After which Binarization have been applied by fixing a particular threshold to bring the image at the scale of o and 1.

B. SEGMENTATION – Segmentation is a procedure which segment the document into line, word and characters. Word Segmentation and Character Segmentation have been applied with two techniques that are Connected Components and Horizontal Projection with Connected Components.

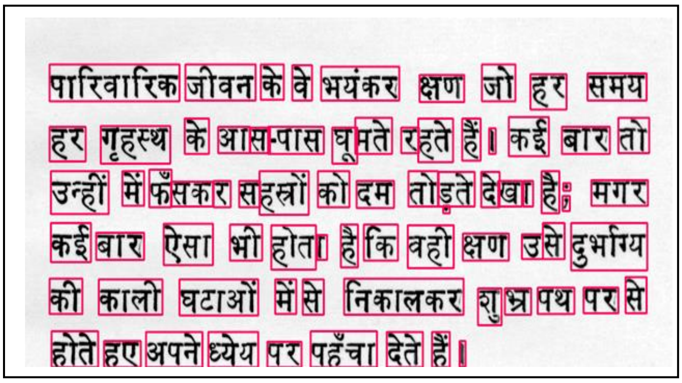
- CONNECTED COMPONENTS –A Connected Component in a Binary Image is a set of pixels that from a connected group. Connected Components helps to identify the connected component in a document and then assigning each identified component with unique labeling. After labeling the connected components in a document, the pixels are grouped according to its connectivity. The Morphological Process begins from the start of the document and spread on the connectivity of the pixel which have the same intensity values. When the groups are determined then the pixels are provided with a unique binary color to distinguish the connected component.



0 0 1 1 0 1
1 1 0 0 0 0
1 1 0 1 0 1
0 1 0 1 0
1 1 0 1 0 1

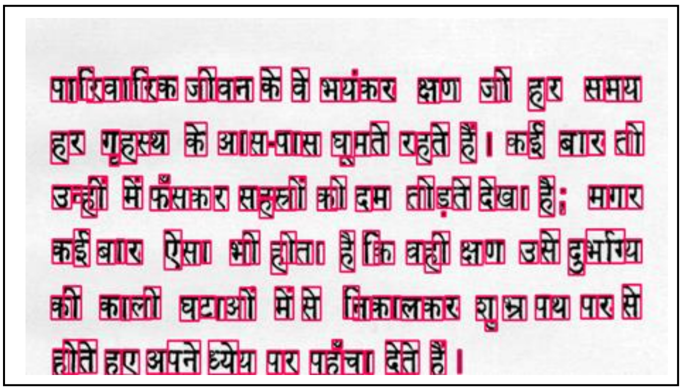
Fig.3.Connected Component

VI. RESULT



पारिवारिक जीवन के वे भयंकर क्षण जो हर समय
हर गृहस्थ के आस-पास घूमते रहते हैं। कई बार तो
उन्हीं में फँसकर सहस्त्रों को दम तोड़ते देखा है; मगर
कई बार ऐसा भी होता है कि वही क्षण उसे दुर्भाग्य
की काली घटाओं में से निकालकर शुभ्र पथ पर से
होते हुए अपने ध्येय पर पहुँचा देते हैं।

Fig.4. Word Segmentation on Devanagari Script



पारिवारिक जीवन के वे भयंकर क्षण जो हर समय
हर गृहस्थ के आस-पास घूमते रहते हैं। कई बार तो
उन्हीं में फँसकर सहस्त्रों को दम तोड़ते देखा है; मगर
कई बार ऐसा भी होता है कि वही क्षण उसे दुर्भाग्य
की काली घटाओं में से निकालकर शुभ्र पथ पर से
होते हुए अपने ध्येय पर पहुँचा देते हैं।

Fig..5. Character Segmentation on Devanagari Script

VII. ANALYSIS

TABLE I. WORD SEGMENTATION

Document	Number of Words in a Single document	Total Segmented words	Accuracy rate of Segmented Words
1	59	49	83.05%
2	175	160	91.42%
3	289	264	91.34%
4	87	84	96.55%
5	41	40	97.56%
Total	651	597	91.70%

TABLE II. CHARACTER SEGMENTATION

Document	Number of Words in a Single document	Total Segmented Characters	Accuracy rate of Segmented Characters
1	129	116	89.92%
2	423	391	92.34%
3	616	535	86.85%
4	191	165	86.38%
5	90	83	92.22%
Total	1449	1290	89.02%

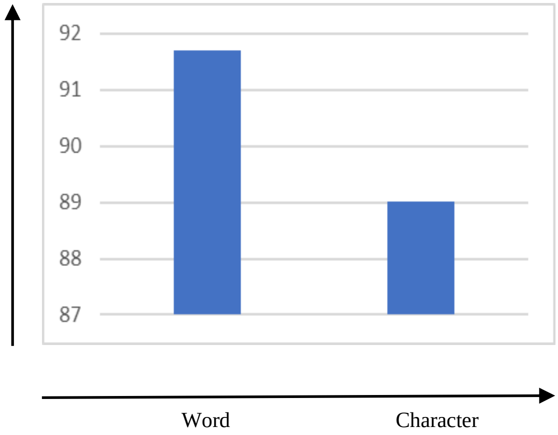


Fig.6. Segmentation Result Using Connected Components in Bar Representation of Word and Character

TABLE III. COMARISION OF ACCURACY OF PROPOSED SYSTEM AND OTHER TECHNIQUES

Ref. No.	*Techniques Used*	*Type of Input*	*Accuracy*
Venna et. al [4]	Structural Properties and Horizontal Projections	Devanagari Script conjunct Characters	83%
Dharam Veer et. al [11]	Horizontal and Vertical Projection Profile	Plain Gurumukhi text	96.22%
Naresh Kumar et. al [5]	Projection Profile Approaches	Plain Hindi text	97%
Proposed system approach	Connected Components	Degraded Hindi Text	89.02%

The Word Segmentation rate of Devanagari Document shows higher percentage of accuracy and a greater result on utilization of Connected Component techniques. On the other hand, in Character Segmentation, the result hasn’t shown a clean result and the percentage of accuracy is lower due to constraint present in Devanagari Script. It is pertinent to note here that the Character Segmentation is the most important and Critical step in Segmentation procedure of OCR as Feature Extraction Procedure is based on the Segmentation procedure. If the Segmentation of the Character is not been

Concentrated well then further step of OCR might not provide a better result.

- CHALLENGING ISSUES IN DEVANAGRI SCRIPTS – The Segmentation is the critical issue is Devanagari Script as the Recognition rate of OCR system depends on the rate of accuracy. Devanagari Script consists of Characters with the different Fonts and Sizes. However, addition of different modifiers makes the segmentation rate challenging. The most important aspects of the Devanagari Script segmentation are that the presence of the header bar which is called Shirorekha makes the segmentation process most challenging. To be more specific presence of upper modifier Shirorekha makes the character segmentation process more complex and challenging. One of such examples given in Fig-2. Here the segmentation of the character has been misled by the presence of the upper modifier and the shirorekh on the character.

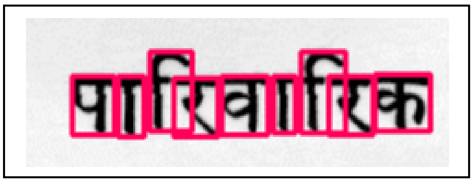


Fig 6. Character Segmentation of the word

The presence of Conjunct Character in Devanagari Script makes difficult in segmenting the character into a single character as the fusion of two or more constant in the middle of the core of the Devanagari Script makes confusion in analyzing the single segmented character. For example – in Fig.2. the Character Segmentation on the Conjunct Character at times Fails and Times shows the positive result.

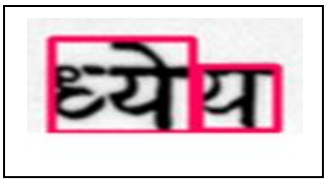


Fig 7. Conjunct Characters

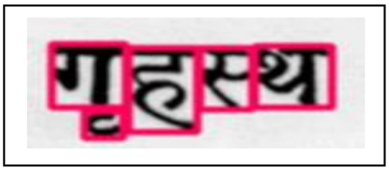


Fig 8. Conjunct Characters

VIII. CONCLUSION

Segmentation rate of the character in Devanagari script is very difficult and challenging due to the presence of different constraint such as Conjunct Characters, different type of modifiers, complex characters with different font and size, etc. Different Algorithm by different researchers have been applied but the Segmentation Rate of those Characters haven’t been achieved at better rate of accuracy. Our technique of Connected Component showed the 92% Accuracy Rate of Word Segmentation and 89% Accuracy Rate of Character Segmentation.

REFERENCES

[1] U. Pal and B. B. Chaudhuri, “Indian script character recognition: A survey”, Pattern Recognit., vol. 37, pp. 1887-1899, 2004.

[2] Rameshwar S. Mohite, Balaji R. Bombade,” Challenging Issue In Devanagari Script” , IJCTA , vol. 5,pp. 947-952 , 2014.

[3] Richard G. Casey, Eric Lecolinet, “A Survey of Methods and Strategies in Character Segmentation”, IEEE Transactions on Patterm Analysis and Machine Intelligence, Vol. 18, No.17, July 1996.

[4] Veena Bansal and R.M.K. Sinha, “Segmentation of Touching Characters in Devanagari”, report of IIT, Kanpur, India.

[5] Naresh Kumar Garg, Lakhwinder Kaur, M.K. Jindal, “Segmentation of Handwritten Hindi Text”, International Journal of Computer Applications, Vol. 1, No. 4, PP: 19-22, 2010.

[6] U. Pal, Sagarika Datta, “Segmentation of Bangla Unconstrained Handwritten Text”, Proceedings of the Seventh International Conference on Document Analysis and Recognition, IEEE 2003.

[7] M. K. Jindal, R. K. Sharma, G. S. Lehal, “Segmentation of Horizontally Overlapping Lines in Printed Gurmukhi Script”, IEEE 2006

[8] Dipak K. Koshti, Sharvari Govilkar, “Segmentation of Touching Characters in Handwritten Devanagari Script”, International Journal of Computer Science and its Applications, Vol. 2, Issue 2, PP: 83-87.

[9] Saiprakash Palakollu, Renu Dhir, Rajneesh Rani, “Handwritten Hindi Text Segmentation Techniques for Lines and Characters”, Proceedings of the World Congress on Engineering and Computer Science, Vol I, WCECS, Oct. 24-26, 2012.

[10] Bansal and Sinha ,- A Complete OCR for Printed Hindi Text in Devanagari Script, IEEE Publications , pp.800-804 , 2001.

[11] Dharam Veer Sharma, G.S. lehal, “An Iterative Algorithm for Segmentation of Isolated Handwritten Words in Gurumukhi Script”, IEEE 18th International Conference on Pattern Recognition, 2006.