International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-9, July 2020

Word and Chracter Segmentation in Devnagari and Odia Script – A Comparitive Analysis

Ipsita Pattnaik, Tushar Patnaik

Abstract: ***Optical Character Recognition has been an active research area in computer science for several years. Several research works undertaken on various languages in India. In this paper an attempt has been made to find out the percentage of accuracy in word and character segmentation of Hindi (National language of India) and Odia is one of the Regional Language mostly spoken in Odisha and a few Eastern India states. A comparative article has been published under this article. 10 sets of each printed Odia and Devanagari scripts with different word limits were used in this study. The documents were scanned at 300dpi before adopting pre-processing and segmentation procedure. The result shows that the percentage of accuracy both in word and character segmentation is higher in Odia language as compared to Hindi language. One of the reasons is the use of headers line in Hindi which makes the segmentation process cumbersome. Thus, it can be concluded that the accuracy level can vary from one language to the other and from word segmentation to that of the character segmentation.***

Keywords: ***Shirorekha, Pre-processing, Segmentation, Devanagari and Odia Scripts***

I. INTRODUCTION

Optical Character Recognition plays an important role in Digital Image Processing and Pattern Recognition. Segmentation has always been an important step in Recognition System. The Segmenting of Indian languages becomes difficult due to many constraints like modifiers, header liners, cursive style and also the degradation quality of the document. The present study deals with the character segmentation of Hindi and Odia language, the former is the National Language of India and the latter is one of Regional Language of Odisha, the state belong to the eastern part of India. Besides Odisha, Odia is also spoken by many people in the eastern Indian states of Jharkhand, West Bengal, Andhra Pradesh and Chhattisgarh There have been many research works conducted on Devanagari and Odia Scripts, however the accuracy percentage have not been achieved greatly.

Although many algorithm and formulae have been experimented on these scripts both on degraded and cleaned scripts, yet the result was not up to that mark.

Revised Manuscript Received on July 30, 2020.
***** Correspondence Author
Ipsita Pattnaik*, M.Tech Computer Science, C-DAC, Noida, India. ipsitapattnaik77@gmail.com
Tushar Patnaik, Research & Development, C-DAC, Noida, India. tusharpatnaik@cdac.in

Hindi operates the Devanagari script .In addition to Hindi, languages like Sanskrit, Marathi and Konkani also operates the Devanagari script, making it the most broadly used script in India.[1] The other considerations for selecting these two languages Hindi and Odia is that they produce two major classes of scripts in India – scripts with and without shirorekha(a head bar). Languages like Bangla, Gurumukhi etc. are in the first class that is with shirorekha while Malayalam, Kannada, Tamil etc. are in the second without shirorekha. [1]

II. PREVIOUS WORK DONE

Chaudhuri and Pal performed Segmentation of words by using the horizontal and vertical projection profiles of the scanned document image [2]. A horizontal center zone of the text line, corresponding to the vertical center parts of the characters, is used to generate a center-zone-only vertical projection profile. The center zone is determined using a horizontal projection profile, by locating the two major peaks of that profile and defining the two major peak positions as the upper and lower boundaries of the center zone. Spacing segments (white gaps) in the vertical projection profile are identified, and classified into two classes, namely character spacing (gap between characters with a word) and word spacing (gap between words). The word spacing was used to segment the text line into word segments. [3] Segmentation of touching characters in Devanagari and Odia script have been a challenging task.

Hindi language having shirorekha (Header Line) and Odia language without shirorekha (Header Line) becomes make it arduous in segmentation of characters in these two languages. Bansal and Sinha [4] mentioned a two-pass algorithm for separation including segmentation of Devanagari characters. Firstly, the words are segmented smoothly into compound characters for understanding the statistics about the height and width of each independent box which is used to find out the theory whether a character box is compound.

Secondly, the theory of whether a character box is compound is identified. The algorithm [5] proposed substantially used structural properties of the script. Shirorekha Removed does the segmentation of character from each Devanagari word. Graine and Chaudhari [6] idea of touching character which are initially identified and then segmented into basic ones are done by new fuzzy decision-making approach. It was developed after looking at the concern of touching character in Devanagari and Bangla Script.

Whereas Tripathi and Pal achieved the global horizontal projection method which is applicable for line segmentation of printed Odia documents,

it cannot be used in unconstrained handwritten Odia documents because the characters of two consecutive text-lines may touch or overlap. Therefore, the opted water reservoir method was used for segmenting the touching Odia characters.[7]

III. OBJECTIVE

The two important objectives of this study are: (i) to find out the accuracy rate of segmentation on Degraded Printed Odia and Devanagari Script and (ii) to differentiate the rate of accuracy of word and character segmentation between the printed Odia and Devanagari Script.

IV. HINDI-ODIA SCRIPTS- A BACKDROP

A. Hindi Scripts

Also known as Devanagari Script. Devanagari Script is an alphabetic Script which is an obvious configuration of its component symbols. They have the writing style from left to right in horizontal formation. The characters don't have the upper case as well as lower case also. It consists of 11 vowels and 33 simple constants, other which add up to the symbols in Devanagari are the set of vowels modifiers called Matras which can be placed to the left, right above or at the bottom of characters or conjuncts. And also, pure constant (also called half letter) which are combined with other constants makes conjuncts.

B. Odia Scripts

Odia script has been derived from the Brahmi script. Odia Script are the cluster of 12 vowels, 35 constants and 10 numerals which are called as basic characters of Odia Language. The have the writing style from left to right in which almost have the characters contain a straight line on the right side. In Odia Script some constants combined with other constants to from a new character. The special symbols are also present in this script known as Modifiers (Matras) which help in changing the meaning and sound of the characters.

C. Similarities in Hindi-Odia Scripts

Both the Hindi and Odia Script are partitioned into three zones. Upper Zone, Middle Zone sometimes also referred as Mean zone and lastly the Lower Zone these are the three zones in Scripts are partitioned into.

The Upper Zone are the one which lay between the mean line and the upper line. The middle zones are the one which lay in between the upper zone and the lower zone.

Whereas, the Lower Zone are the one which ly between the base line and the lower line.

The Modifiers which are also known as Matras are placed and present on the Upper and Lower Zone.

Sometimes the modifiers might have the connectivity at the three zones also. [8,9]

V. METHODOLOGY

10 sets of each printed Odia and Devanagari scripts with different word limits have been used in this study.

At the outset, the document was scanned at 300dpi before adopting pre-processing and segmentation procedure.

In pre-processing Grayscale conversion and Binarization is done by setting at fixed threshold. While, segmentation of word and character have been done by using contour-based procedure. The system work procedure is given in the form of a flow chart in Figure-1.

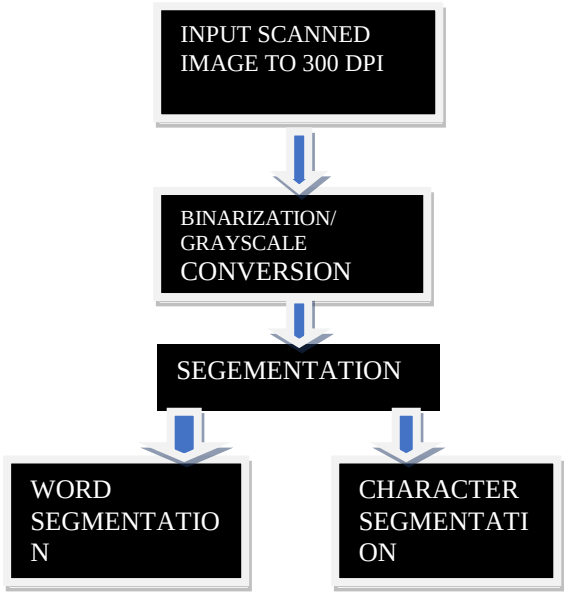


Fig 1. System Work Procedure

A. Input Scanned Degraded Image

In Optical Character Recognition procedure scanned input document has been the most important step as it provides the path for entering the printed or handwritten document into the computer after which the OCR procedure begins.

Besides, PDF files that are already stored are also be converted by the way digital camera captured image into computer readable form.

The images are defined in the configurations such as JPEG, BMT, BMP, TIF and TNG.

The initial input obtained may be in gray, color or binary tone [10]. The right scanning procedure have also been a great challenge due to large number of documents or due to non-serious scanning procedure .

In scanning procedure the DPI is used to illustrate the resolution of number of dots per inch in a digital print and the printing resolution of the hard copy print dot gain, which is the increase in the size of the halftone dots during printing. This is caused by the spreading of ink on the surface of the media.

Therefore, scanning a document in right dips' is a crucial task. 300 dpi have been used as the standard measure for scanning a document because lesser than 300 dpi scanning will result problems like the possible jagged lines, missing data and low-quality letter edges. Besides, scanning higher at higher than 300 dpi will lead to large image files making them inconvenient for transport and storage.

[11] Therefore, scanning at 300 dpi make it clear and easier to read. The scanned images of both Hindi and Odia script is given in Figure-2 and Figure 3 respectively.

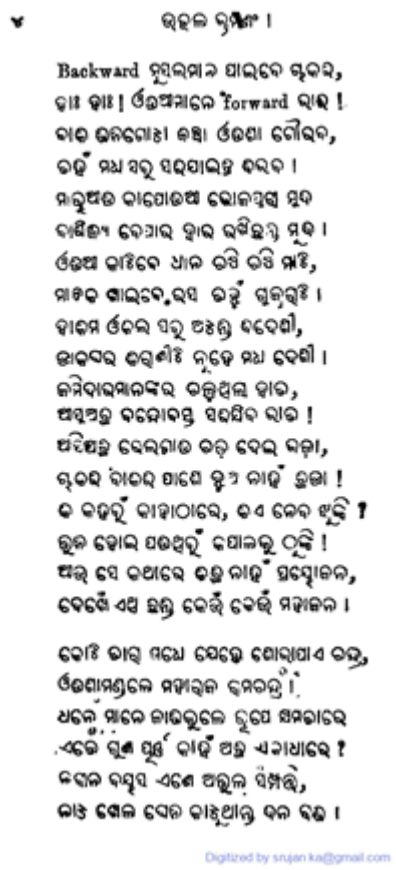
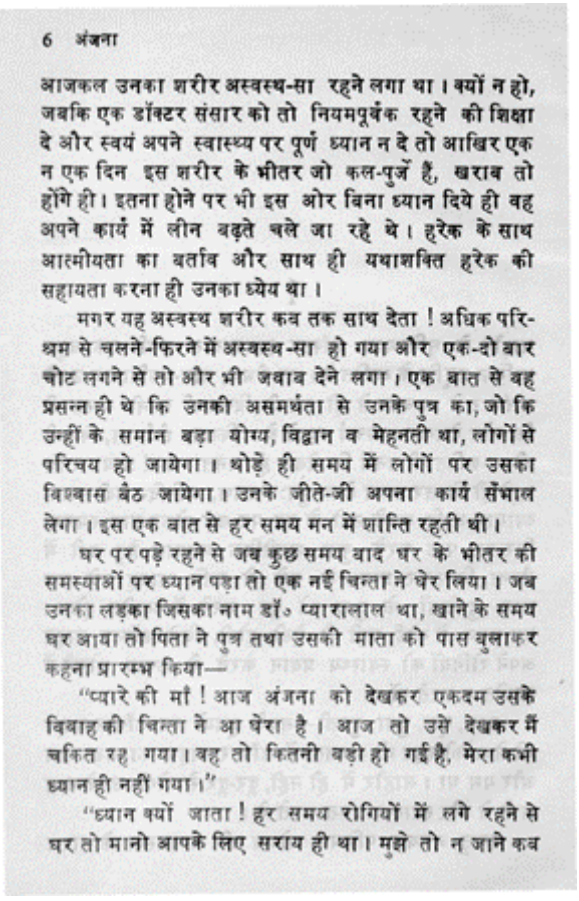


Fig 2. **Scanned Image of Odia Script**



6 अंजना

आजकल उनका शरीर अस्वस्थ-सा रहने लगा था। क्यों न हो, जबकि एक डॉक्टर संसार को तो नियमपूर्वक रहने की शिक्षा दे और स्वयं अपने स्वास्थ्य पर पूर्ण ध्यान न दे तो आखिर एक न एक दिन इस शरीर के भीतर जो कल-पुर्जे हैं, खराब तो होंगे ही। इतना होने पर भी इस ओर बिना ध्यान दिये ही वह अपने कार्य में लीन बढ़ते चले जा रहे थे। हरेक के साथ आत्मीयता का बर्ताव और साथ ही यथाशक्ति हरेक की सहायता करना ही उनका ध्येय था।

मगर यह अस्वस्थ शरीर कब तक साथ देता! अधिक परिश्रम से चलने-फिरने में अस्वस्थ-सा हो गया और एक-दो बार चोट लगने से तो और भी जवाब देने लगा। एक बात से वह प्रसन्न ही थे कि उनकी असमर्थता से उनके पुत्र का, जो कि उन्हीं के समान बड़ा योग्य, विद्वान व मेहनती था, लोगों से परिचय हो जायेगा। थोड़े ही समय में लोगों पर उसका विश्वास बैठ जायेगा। उनके जीते-जी अपना कार्य सँभाल लेगा। इस एक बात से हर समय मन में शान्ति रहती थी।

घर पर पड़े रहने से जब कुछ समय बाद घर के भीतर की समस्याओं पर ध्यान पड़ा तो एक नई चिन्ता ने घेर लिया। जब उनका लड़का जिसका नाम डॉ० प्यारालाल था, खाने के समय घर आया तो पिता ने पुत्र तथा उसकी माता को पास बुलाकर कहना प्रारम्भ किया—

"प्यारे की माँ! आज अंजना को देखकर एकदम उसके विवाह की चिन्ता ने आ घेरा है। आज तो उसे देखकर मैं चकित रह गया। वह तो कितनी बड़ी हो गई है, मेरा कभी ध्यान ही नहीं गया।"

"ध्यान क्यों जाता! हर समय रोगियों में लगे रहने से घर तो मानो आपके लिए सराय ही था। मुझे तो न जाने कब

Fig 3. **Scanned Image of Hindi Script**

B. Binarization/Grayscale Conversion

Grayscale Conversion are done in order to achieve any colored image. While, RGB image are required to convert into grayscale image. The grayscale image has value ranges from 0 to 255 and also the ranging can be in variation of color between black and white and in various shades of gray. Mathematical formulation of RGB is:

$$\frac{(R+G+B)}{3}$$

Binarization are the procedure where image is turned into 0 and 1 scale which is computer readable format. Binary means either 0 (black) or 1(white). Here the binarization are achieved by setting a threshold at fixed size.

The input image of grayscale pixels is treated as raster scan order and mentioned as $x_i \in [0, 1]$. The parallel output of binarization pixels are mentioned as $b_i \in \{0, 1\}$, where 0 refers to "black" and 1 refers to "white". The simplest binarization technique is the global fixed threshold technique where $b_i = 1$ if $x_i >= 0.5$ and $b_i = 0$ if $x_i < 0.5$.

C. Segmentations

This is the most important step in Optical Character Recognition procedure. This procedure is further sub-decided into line, word and character segmentation. Word segmentation is the most critical step in shirorekha languages as the segmenting the word with a shirorekha can change the meaning of a Word or a Character which can be a critical step and also at time of not proper segmenting can become difficult in recognition procedure.

D. Connected Component and Region Based Method

The Connected Component based methods finds out the words present in an image irrespective of the position of an image. When a word is detected from an image it is then processed. In Connected Component Region based method there is no specific difference between the parts where the text is present and the parts where the text is not present it fails when the Connectivity is lost within the Character. In Region based procedure [12].

An image is differentiated between two parts one defining the text region and other presenting the Non-Text region. They include a sub-part technique called Contouring which is used to identify contour of the words in the text line. They also track words from left to right. In Contouring the foreground pixels that have no foreground pixels on there right, top and bottom up to same threshold distance both vertical and horizontal are treated as Contour.

A simple technique was adopted to detect contour of the words in the text line. The contours were traced in the text line from right to left. The foreground pixel that have no foreground pixels on its right, top and bottom up to some threshold distance both vertical and horizontal are treated as contour. Contour-based segmentation technique gives a clear description of the word characters shape. This method facilitates determining the right segmentation points.

However, many methods have been tested to extract the contour of the abstracted word/sub word image.[13]

VI. DATA ANALYSIS

The analysis of data has been done in two parts: (i) word segmentation and (ii) character segmentation

A. The Word Segmentation

A dataset of 10 degraded documents of both Hindi and Odia Script have been taken with a set of different words and each document was tested including constants, special characters, modifiers, and digits from 0 to 9. The dataset fonts were in random and not of some similar font. Greater the number of dataset characters, greater is the system efficiency. The proposed segmented system has been implemented using Spyder version 4.1.2. The Connected component strategy gives 90 percent accuracy in segmenting Degraded Odia and Hindi Script Documents. The Table-1 and Table-2 shows the experimental results of word segmentation on Odia and Hindi Documents.

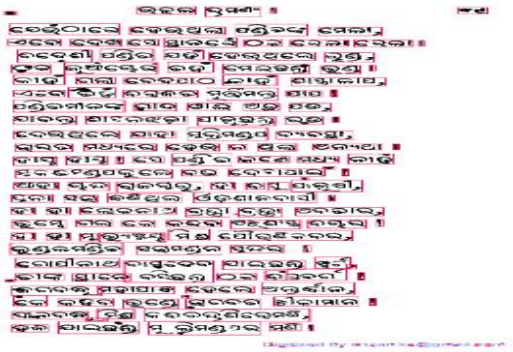


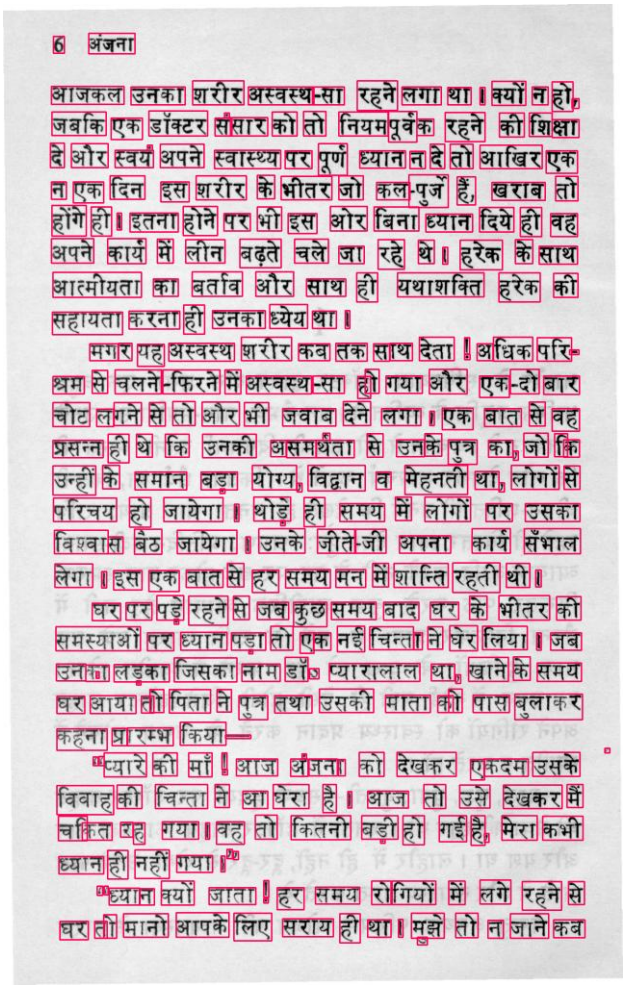
Fig 4. Experiments- Word Segmentation on Odia Scripts

TABLE-I: Hindi Document Analysis on Word Segmentation in Different Font and Style

DOCUMENT	TOTAL NUMBER OF INPUT WORDS	NO. OF CORRECTLY SEGMENTED WORDS	ACCURACY (%)
1	59	48	81.35%
2	172	158	91.86%
3	192	175	91.14%
4	87	83	95.40%
5	197	173	87.81%
AVERAGE	707	637	90.09%

TABLE-II: Odia Document Analysis on Word Segmentation in Different Font and Style

DOCUMENT	TOTAL NUMBER OF INPUT WORDS	NO. OF CORRECTLY SEGMENTED WORDS	ACCURACY (%)
1	187	177	94.65%
2.	97	83	85.56%
3	172	157	91.27%
4.	5	5	100%
5.	142	133	93.66%
AVERAGE	603	555	92.03%



6 अंजना

आजकल उनका शरीर अस्वस्थ-सा रहने लगा था। क्यों न हो, जबकि एक डॉक्टर संसार को तो नियमपूर्वक रहने की शिक्षा दे और स्वयं अपने स्वास्थ्य पर पूर्ण ध्यान न दे तो आखिर एक न एक दिन इस शरीर के भीतर जो कल-पुर्जे हैं, खराब तो होंगे ही। इतना होने पर भी इस ओर बिना ध्यान दिये ही वह अपने कार्य में लीन बढ़ते चले जा रहे थे। हरेक के साथ आत्मीयता का बर्ताव और साथ ही यथाशक्ति हरेक की सहायता करना ही उनका ध्येय था।

मगर यह अस्वस्थ शरीर कब तक साथ देता! अधिक परिश्रम से चलने-फिरने में अस्वस्थ-सा हो गया और एक-दो बार चोट लगने से तो और भी जवाब देने लगा। एक बात से वह प्रसन्न ही थे कि उनकी असमर्थता से उनके पुत्र का, जो कि उन्हीं के समान बड़ा योग्य, विद्वान व मेहनती था, लोगों से परिचय हो जायेगा। थोड़े ही समय में लोगों पर उसका विश्वास बैठ जायेगा। उनके जीते-जी अपना कार्य सँभाल लेगा। इस एक बात से हर समय मन में शान्ति रहती थी।

घर पर पड़े रहने से जब कुछ समय बाद घर के भीतर की समस्याओं पर ध्यान पड़ा तो एक नई चिन्ता ने घेर लिया। जब उनका लड़का जिसका नाम डॉ० प्यारालाल था, खाने के समय घर आया तो पिता ने पुत्र तथा उसकी माता को पास बुलाकर कहना प्रारम्भ किया—

"प्यारे की माँ! आज अंजना को देखकर एकदम उसके विवाह की चिन्ता ने आ घेरा है। आज तो उसे देखकर मैं चकित रह गया। वह तो कितनी बड़ी हो गई है, मेरा कभी ध्यान ही नहीं गया।"

"ध्यान क्यों जाता! हर समय रोगियों में लगे रहने से घर तो मानो आपके लिए सराय ही था। मुझे तो न जाने कब

Fig 5. Experiments- Word Segmentation on Hindi Script

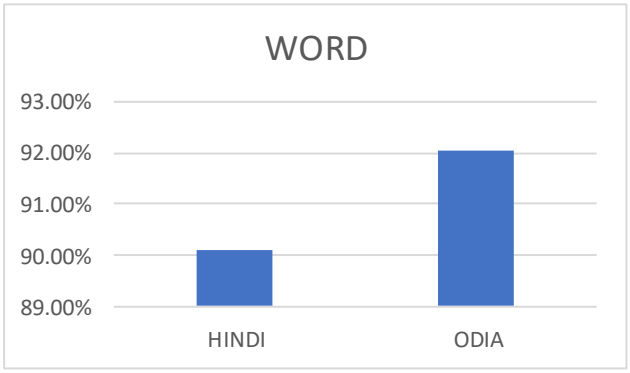


Fig 6. Analysis Of Word Segmentation On Different Document

Word Segmentation showed almost same accuracy as word have good spacing which helps connected components to count the nearest closed pixel. The header line that is shirorekha in Devanagari Script were also not a big concern in segmentation of words as the meaning of the word would change if the shirorekha was made to remove. However, the mean percentage of accuracy is slightly higher in Odia script 92.03 percent as compared to Hindi script 90.09 percent. This may be due to more use of Matras in Hindi language as compared to Odia language.

B. The Character Segmentation

Dataset of 10 degraded documents of both Hindi and Odia Script have been taken with a set of different words and each document was tested including constants, special characters, modifiers, and digits from 0 to 9. The dataset fonts are in random and not of same similar font. Greater the number of dataset characters, greater is the system efficiency.

The proposed segmented system has been implemented using Spyder version 4.1.2. The Connected component strategy gives different results in the accuracy of character segmentation of the Degraded Odia and Hindi Script Documents. It gives 70% accuracy in Hindi whereas 90% accuracy in Odia Scripts. This clearly shows that there is greater degree of accuracy in character segmentation of in Odia language as compared to the Hindi language. The Table-3 and Table-4 show the experimental results of character segmentation on Odia and Hindi Documents.

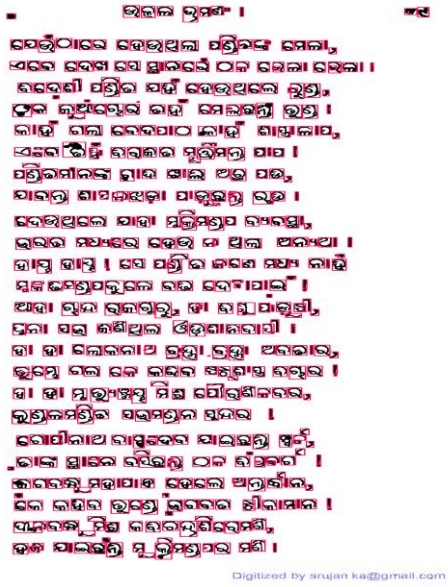


Fig 7. Experiments- Character Segmentation on Odia Scripts

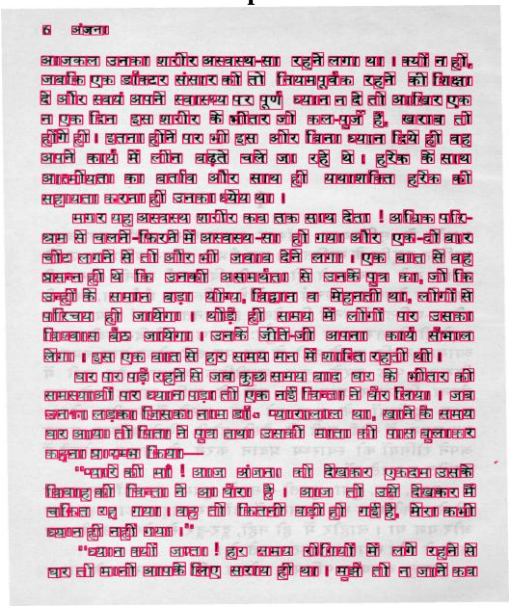


Fig 8. Experiments- Character Segmentation on Hindi Script

TABLE-III: Hindi Document Analysis on Character Segmentation in Different Font and Style

DOCUMENT	TOTAL NUMBER OF INPUT CHARACTER IN THE DOCUMENT	NUMBER OF CORRECTLY SEGMENTED CHARACTER	ACCURACY (%)
1	128	100	78.12%
2	430	328	76.27%
3	189	135	71.42%
4	470`	403	85.74%
5	90`	76	84.44%
AVERAGE	1307	1042	79.72%

TABLE-IV: Odia Document Analysis on Character Segmentation in Different Font and Style

DOCUMENT	TOTAL NUMBER OF INPUT CHARACTER IN THE DOCUMENT	NUMBER OF CORRECTLY SEGMENTED CHARACTER	ACCURACY (%)
1	292	289	98.97%
2	154	140	90.90%
3	192	187	97.39%
4	149	142	95.30%
5	198	187	94.44%
AVERAGE	985	945	95.93%

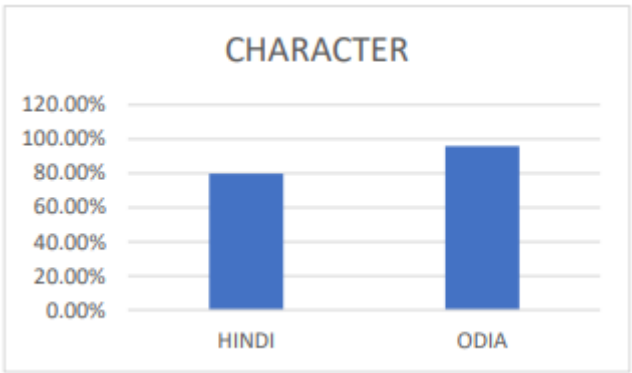


Fig 9. Analysis of Character Segmentation

VII. EXPERIMENTAL OBSERVATIONS

Character Segmentation in the Devanagari showed much lower response then Odia Script. The header line and the lower and upper modifier created complication in the segmentation of character in the script.

Shirorekha that is header line attribute holds an important role in Devanagari Script segmentation procedure which further leads to recognition procedure. This attribute is used for the identification of word limit. If the Shirorekha is absent from a word, it leads to recognition issues. Whereas, the presence of more than one header line (Shirorekha) adds disturbance for two text lines. Whereas in Odia script the lower and upper modifier created complication in the segmentation of the character which can future lead to recognition problem. Devanagari word can be recognized by the presence of number of Characters, vertical bars (Shirorekha) and Modifiers. Whereas the Odia word consists of information about number of characters and the number of modifiers. Devanagari and Odia word consists of complex mixture of all of the above elements, when we start segmenting the Characters the confusion created by all the Modifiers and the conjuncts makes the segmenting rate slow and low leading to difficult recognition rate. The constraints present in both the script leads to low recognition rate. The presence of Shirorekha in Deavaganari Script and the presence of Modifiers on both the script creates a huge lot of problems. The gap between the character and the modifier doesn't touch the core character at all, makes the situation more tedious.

VIII. FINDINGS AND CONCLUSION

The important findings based on data analysis and findings are as follows: (I) there is high degree of mean accuracy percent in the word segmentation between Hindi and Odia degraded scripts, the percentages are 90.09 and 92.3 respectively. In other words, the differences in percentage of accuracy is small 2.21 percent; (ii) on the other hand the mean accuracy percentage in character segmentation is low in Hindi language 79.72 percent while in case of Odia language it is 95.93, the different is 16.21.; and (iii) the result shows that the percentage of accuracy varies from word segmentation to that of character segmentation. From the experimental observation it can be concluded that segmentation rates depend upon the pre-processing and the degradation of the script and the styling of the script. Devanagari script (Hindi language) recognition is a tricky job due to presence of header line as compared to Odia script having less header lines. The accuracy level also can vary from one language to the other depending on the header line present on the language. In this paper, the problems are elaborated and the algorithm showing different accuracy rates in different script. However, it also led to scope of further and future research in the field of OCR by using different algorithms in order to improve the percentage of accuracy.

REFERENCES

1. C. V. Jawahar, M. N. S. S. K. Pavan Kumar and S. S. Ravi Kiran, "A bilingual OCR for Hindi-Telugu documents and its applications," Seventh International Conference on Document Analysis and Recognition, 2003. *Proceedings.*, Edinburgh, UK, 2003, pp. 408-412 vol.1, doi: 10.1109/ICDAR.2003.1227699.
2. Chaudhuri, Bidyut Baran and U. Pal. “An OCR system to read two Indian language scripts: Bangla and Devanagari (Hindi).” Proceedings of the Fourth International Conference on Document Analysis and Recognition 2., pp. 1011-1015 vol.2., 1997.
3. https://patents.justia.com/patent/9430703,Ming,Wei (Cupertino, CA, US)-United State Patent Method for Segmenting Text Words in Document Images Using Vertical Projections Of Centre Zone Of Characters, Application 20160180163, 2016, accessed on 29 April 2020 at 8:25 am.
4. V. Bansal and R. M. K. Sinha, “Integrating knowledge sources in Devanagari text recognition,” IEEE Trans. Syst. Man Cybern. A: Syst. Hum., vol. 30, no. 4, pp. 500–505, Jul. 2000.
5. Bansal, Veena & Sinha, R.M.K., “Segmentation of touching and fused Devanagari characters,” Pattern Recognition. 35., pp.875-893, 2002.
6. Garain, U., Chaudhuri, B.,” Segmentation of touching characters in printed Devanagari and Bangla scripts using fuzzy multifactorial analysis‖,” IEEE Trans. Syst. Man Cybern., Part C 32 (4), pp. 449–459, 2002.
7. N Tripathi and U Pal,” Handwriting segmentation of unconstrained Oriya text,” Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Sadhan, Vol. 31, Part 6, pp. 755–769, December 2006.
8. Subhadip Basu, Nibaran Das, Ram Sarkar, Mahantapas Kundu, Mita Nasipuri, Dipak Kumar Basu,” A hierarchical approach to recognition of handwritten Bangla characters,” Pattern Recognition 42, pp.1467 -1484, 2009.
9. K. Roy and U. Pal, “Word-wise Hand-written Script Separation for Indian Postal automation,” In 10th International Workshop on Frontiers in Handwriting Recognition, pp. 521-526, 2006.
10. Manoj Kumar et al. "Automatic Text Location from Complex Natural Scene Images", Second International Conference on Computer and Automation Engineering, Page(s): 594 – 597,2010.
11. https://scansnapcommunity.net/why-is-ocr-at-300-dpi-a-standard-2/, Shane Cooper, “Image Quality for Document Capture, is more DPI is always better?”, Parascript Blog ,2013, accessed on 21st May 2020 at 4:32 pm.
12. https://doi.org/10.1117/1.JEI.28.4.043030,Khader Mohammad, Aziz Qaroush, Muna Ayyesh, Mahdi Washha, Ahmad Alsadeh, Sos Agaian, "Contour-based character segmentation for printed Arabic text with diacritics," J. Electron. Imag. 28(4) 043030 ,30 August 2019
13. Ali, A.A.A., Suresha, M.,” Survey on Segmentation and Recognition of Handwritten Arabic Script.” *SN COMPUT. SCI.* 1, Article number192, pp. 3-31,2020.

AUTHORS PROFILE

Ipsita Pattnaik BTech from Guru Gobind Singh Indraprastha University and completed MTech from C-DAC Noida.

Tushar Patnaik currently The Joint-Director at C-DAC, Noida. His main research work focuses on Pattern Recognition, Image Processing, Feature Extraction, Computer Vision, Pattern Classification Digital Image Processing, Machine Learning, Feature Selection and Signal, Image and Video Processing.