Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.8 No.1, 2019, pp. 50-54

Character Segmentation and Skew Correction for Handwritten Devanagari Scripts: A Friends Technique

Ashok Kumar Bathla[1], Sunil Kumar Gupta[2] and Manish Kumar Jindal[3]
[1]Research Scholar, I. K. Gujral Punjab Technical University, Kapurthala, Punjab, India
[2]Associate Professor, Beant College of Engineering and Technology, Gurdaspur, Punjab, India
[3]Professor, P.U. Regional Centre, Muktsar, Punjab, India
E-Mail: Ashokashok81@gmail.com, skgbcetgsp@gmail.com, manishphd@rediffmail.com

Abstract **- Optical Character Recognition (OCR) technology allows a computer to "read" text (both typed and handwritten) the way a human brain does.Significant research efforts have been put in the area of Optical Character Segmentation (OCR) of typewritten text in various languages, however very few efforts have been put on the segmentation and skew correction of handwritten text written in Devanagari which is a scripting language of Hindi. This paper aims a novel technique for segmentation and skew correction of hand written Devanagari text. It shows the accuracy of 91% and takes less than one second to segment a particular handwritten word.**
Keywords:**Segmentation, Compound Character, Devanagari**

I. INTRODUCTION

The major challenge in segmentation involves removal of background noise, as when we are doing the digitization [1-3] or Binarization [1] of an image due to scanning problems, the noise gets embedded into the image. The other problems involve gray levels, noisy background, multiple skew levels, variable font size and different styles of writing.

Devanagari is main scripting language of Hindi, Dogri, Marathi and Sanskrit [2]. Unlike Urdu, it is written from left to right and there is no notion of upper and low case of the characters.It is a segmental writing system in which consonant vowels sequences are written as a unit. This scripting language contains fourteen vowels and thirty three simple consonants [1]. Apart from consonants and the vowels, it includes other consonants called 'Matras'. Matras includes the set of vowel modifiers that have been placed at the top, bottom, left and right of the character or conjunct. Pure consonants when combine with a half character lead to conjuncts (Character combined with the half Character).

"Shirorekha" is another term used when a horizontal [10] line above the characters that runs through when we write the words also called as Header Line. The word written in Hindi has three zones viz. lower, upper and middle. Upper zone includes some vowels and Matras above the horizontal line, lower zone includes the bottom modifiers below the base line and middle zone constitutes the basic characters. Above complete concept is shown in the Fig.1 to Fig.5.

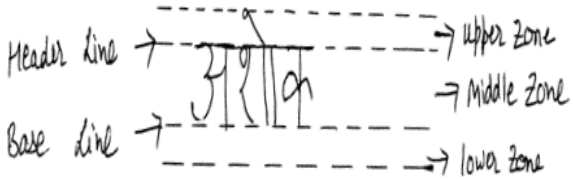


Fig. 1 Different zones for a word in Devanagari script

क	क़	ख	ख़	ग
ग़	घ	ङ	च	छ
ज	ज़	झ़	ञ	ट
ठ	ड	ड़	ढ	ढ़
ण	त	थ	द	ध
न	प	फ	फ़	ब
भ	म	य	र	ल
व	श	ष	स	ह

Fig. 2 Devanagari Script Consonants

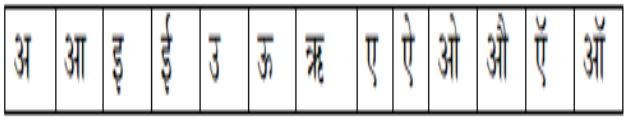


Fig. 3 Devanagari Script Vowels

Fig. 4Devanagari Script Matras or Vowel Modifiers

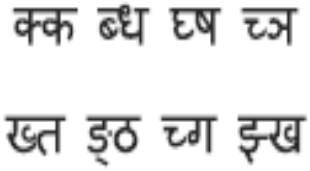


Fig. 5 Devanagari Script Conjuncts some examples

Major contributions of the research article are

1. To apply preprocessing technique with an aim to remove skewness in handwritten texts
2. To develop a novel algorithm for segmentation of preprocessed handwritten text written in Devanagari using vertical profile projection with an aim to get output as compound characters.

II. LITERATURE REVIEW

Since the last decade, significant research efforts have been put in the field of OCR in different languages. Bagg and Harit highlighted an extensive review of various techniques of character recognition for two popular Indian scripts viz. Bangla and Devanagari [13]. Authors observed that few more efforts are neededfor pre-processing phase of OCR. Bhattacharya *et al*. presented aproficient technique to extract regions having text from natural images of Bangla and Devanagariscripts [14]. In continuation, a fuzzy multi-factorial analysis based method of segmentation was highlighted for the segregation of characters that are touchingin printed Bangla and Devanagari text images [12]. Jayadevan *et al*. published a survey on offline recognition of the Devanagari script [15]. It has been observed that accuracy of the system decreases due to large number of false positives in segmentation process. Ma and Doermann developed system that reported accuracy of 88% and 95% on noisy and ideal images respectively using generalized comparison of Hausdroff image method (GHIC) [16].Further, Murthy *et al*. [17] performed little work on segmentation of text from images. Narang *et al*. presented novelclassification method text segmentation from text images of 40 classesbasically [18]. However, method proposed is highly intricate and costly in terms of time. Pal *et al*. has done recognition on character level on multi lingual written text images of Devanagari and Bangla scripts using the backdrop &forefront knowledge of the text image [12]. Thakral *et al*. shown a technique for segmentation and cluster detection which redress the issue like conjunct and touching characters and reported a accuracy of 95% for these and 88% for overlapping characters[6].

III. PROPOSED FRIENDS TECHNIQUE

In this section, a novel approach named as Friends Technique is proposed that overcomes some of the limitations of above discussed research studies. The step by step procedure is explained below.

Step 1

In this step, removal of skewness is performed on handwritten Devanagari text that is to be segmented. It is achieved using the same process used in Gauranga Mandal[2]. The authors removed skewness up to -45 to +45 degrees. However, we removed the skewness up to -5 to 5 degree. It is assumed that different handwritten styles never write in such manner which increases the skewness level more than 5 degree. This process reduces the time overhead of detecting and correcting skewness. Skewness algorithm works on the principle of tilting the handwritten text image to -5 to 5 degree with the increment of 1 degree at each step. After every step the algorithm calculates the height of the handwritten text image. The step with minimum height of the text image is chosen as a final text image free from any skewness. The whole process of skew correction is shown in the Fig.6.

Algorithm for skew correction is mentioned below

1. Consider co-ordinates (x,y) of all the pixels of handwritten word and find minimum value of y axis (min_y) and maximum value of y axis (max_y).
2. Calculate the height of the word, i.e. height = (max_y - min_y)
3. Rotate all the pixels of the word in anticlockwise direction by 1 degree i.e. rotation of any point (x,y) by certain angle θ with respect to the following pointx'= x_r+(x-x_r)cosθ-(y-y_r)sinθ and y'=y_r+(x-xr)sinθ +(y-y_r)cosθ
   (where x' and y' are new generated co-ordinate value, x and y are old co-ordinate value, x_r and y_r are co-ordinate of centroid of the word.)
4. Now calculate the height of the word again and again for each 1 degree of rotation (up to 5 degree anticlockwise and 5 degree clockwise).
5. Find the particular angle where the height of the handwritten word is minimum and stop the rotation.
6. If it is observed that height is minimum for more than one angle then, consider coordinates (x,y) of all the pixels of handwritten word and find minimum value of x axis (min_x) and maximum value of x axis (max_x)
7. Calculate the width of the word, i.e. width = (max_x - min_x)
8. Now check width of the word for the entire angle where same minimum height exists.
9. Consider the angle where the width value is maximum for the minimum height and stop the rotation.
10. Now the skew correction is almost done but not exactly, so calculate busy zone of the word by checking number of pixels available in each y co-ordinate to remove unwanted prolonged part.
11. Now consider only the busy zone of the handwritten word (don't consider those pixels which are outside of the busy zone) and repeat step 1 to step 9 once again.
12. The handwritten word is skew corrected.

Fig. 6 Process Depicting of removal of skewness.

Step 2

Second operation that is to be performed in Friends technique is to remove header line from the word or image by using horizontal profile projection [5,7,8]. Horizontal rows having maximum of black pixels will be treated as a header line and is removed then. For this, we have used Histogram technique as normalized method from 0 to 100 for better normalization [9] results.

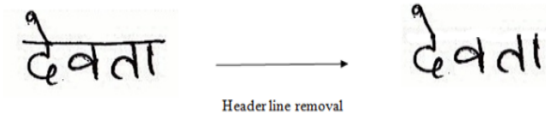


Fig. 7 Process showing removal of Header Line

Step 3

Then in the third phase of Friends technique a vertical profile projection of word is taken or we can say that the image of the text and the different characters are segmented where we found out the row having no black pixels. We will segment in this way the characters of word. By using this technique the Broken[11] and Discrete(correct) characters will get segmented but the touching characters will remain as such or our method will treat a two touching characters as one character.

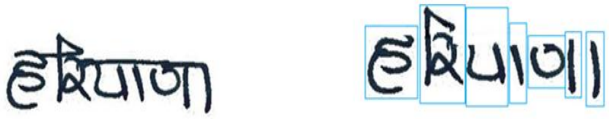


Fig. 8 Process showing Segmentation of Characters

Step 4

In this step noise is removed by discarding those chunks of pixels or cluster of pixels having size less than the 10% size of the segmented character or we can also fix a threshold [1] 5 to 10 pixels also, in this way unwanted noise can be removed from the image.

Fig. 9 Process or step showing removal of Noise

Step 5

In this step we calculate the average height of our segmented compound character. Here we are mentioning our segmented character as a compound character because in this technique upper zone matras and lower zone matras are not segmented differently. Here we have only separated the header line from the word rest all other matras. In other words we can say the symbols in upper, middle zone or lower zone are kept as such and also segmented all through under one process. So, the segmented characters with all zones symbols(upper, middle and lower) intact is called as a compound character.

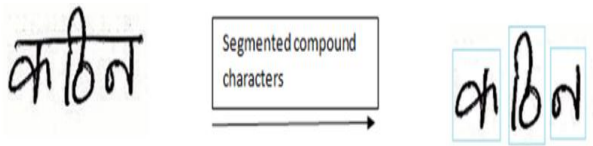


Fig. 10 Process of step showing Segmentation of Compound Character

Step 6

In this step, our main aim is to make the segmentation process more correct. We calculate the average height of all the segmented characters. However, if next consecutive character's height is less than 50% of average height then it is merged with previous character to form true compound character or correct compound character. It is worth to mention here that while in segmentation process of step 3 some of matras of lower or upper zone which originally belongs to previous character gets segmented as discrete character. Different characters are merged onto its true position to make segmentation process more clear, correct and robust.

Fig. 11 Correct Segmentation Process Depiction after merging character with previous character, having height less than 50% of average height.

Step 7

Co-ordinates are then arranged to form the word from the compound characters.

IV.COMPARATIVE ANALYSIS

Proposed friends techniqueis applied with a data set of words in which it is observed that it segmented correctly, partially segmented and wrongly segmented. List is shown in the Table I. A new method (the Friends Technique) is proposed for segmentation [5-9] of Devanagari text. The Friends technique efficiently segments the characters from the words. There are poorly and partially segmented characters from our proposed method also but that seems to be of very few in quantity. Furthermore, our method is more time proficient than its competitive ones in terms of execution time as our method takes times less than a second to successfully perform the entire operation. Character Segmentation results on 100 words dataset using proposed Friends technique is shown in Table II.

TABLE I RESULT ANALYSIS OF SEGMENTED IMAGES AFTER SKEW CORRECTION

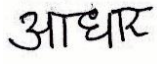
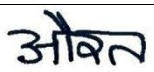
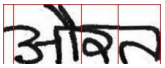
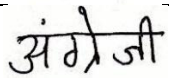
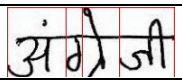
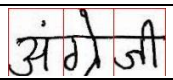
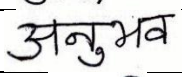
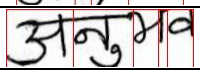
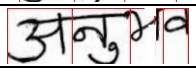
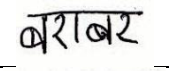
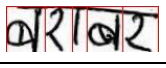
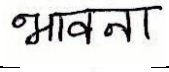
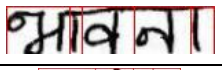
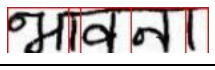
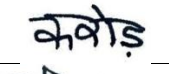
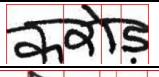
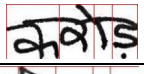
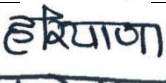
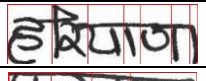
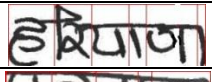
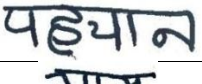
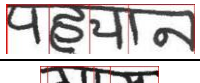
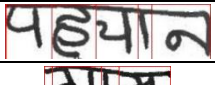
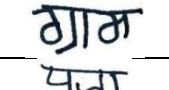
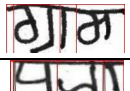
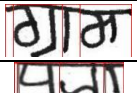
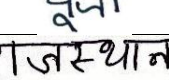
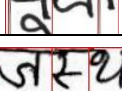
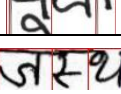
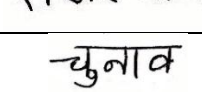
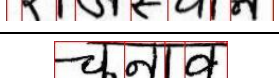
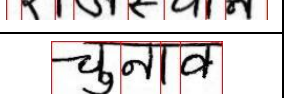
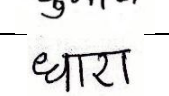
S. No.	Original Image	After Segmentation	Desired result
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			

TABLE II CHARACTER SEGMENTATION RESULTS ON 100 WORDS DATASET USING PROPOSED FRIENDS TECHNIQUE

Characters	Total
Correctly Segmented	337
Partially Segmented	14
Wrongly Segmented	22
Overall Accuracy	90.35%

V. CONCLUSIONS AND FUTURE SCOPE

Our proposed "Friends" technique achieved the accuracy of more than 90% in case of correctly segmented characters and nearly 4% for partially segmented characters and 5.9% for wrongly corrected characters. Moreover, it takes less than one second to execute the entire process which is faster than 70% with the related studies. In the close to prospect, the proposed Friends method can be practically applied on multi script documents on which characters are connected through a common aspect of having Header line (Shirorekha) on top of them.

REFERENCES

[1] K. Jindal, and R. Kumar, "A new method for segmentation of pre-detected Devanagari words from the scene images: Pihu method", *Computers and Electrical Engineering*, 2017.

[2] G. Mandal, "An Unprecedented Approach of Skew Detection and Correction for Online Bengali Handwritten Words", *IJARCET*, Vol. 7, No. 2, Feb 2018.

[3] D. V Sharma and G. S, Lehal, "An Iterative Algorithm for Segmentation of Isolated Handwritten Words in Gurmukhi Script", *IEEE International Conference ON Pattern Recognition*, Vol. 2, pp. 1022-1025, 2006.

[4] M. Kumar, 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*, Vol. 6, No. 2, pp. 58-63, 2014.

[5] P. Mangla and H. Kaur, "An End Detection Algorithm for segmentation of Broken and Touching characters in Handwritten Gurmukhi Word", *IEEE International Conference on Reliability, Infocom Technologies and Optimization*, pp. 1-4, 2014.

[6] B. Thakral and M. Kumar, "Devanagari Handwritten Text Segmentation for Overlapping and Conjunct Characters: A Proficient Technique", *IEEE International Conference on Reliability, Infocom Technologies and Optimization*, pp. 1-4, 2014.

[7] P. Mangla. And H. Kaur, "An End Detection Algorithm for segmentation of Broken and Touching characters in Handwritten Gurmukhi Word", *Institute of Electrical and Electronics Engineers (IEEE)*, pp. 1-4, 2014.

[8] V. Bansal and R. M. K. Sinha, "Segmentation of touching in Devanagari characters", *Proceeding, CVGIP, Delhi*, 1998.

[9] G. S. Lehal and C. Singh, "A technique for segmentation of Gurmukhi text", *Computer Analysis of image and patterns*, Vol. 21, No. 24, Springer, pp. 191-200, 2001

[10] A. K. Bathla., S.K.Gupta, and M.K. Jindal, "Identification and Recognition of Handwritten Devanagari text Using Machine Learning Classifiers ", *IJMDEBM*, Vol. 6, No. 2, pp. 21-25, April-June 2018.

[11] M. K. Sachan, G. S. Lehaland V. K. Jain, "A Novel Method to Segment Online Gurmukhi Script", *Information Systems for Indian Languages. ICISIL 2011. Communications in Computer and Information Science, Springer*, Vol. 13, No. 9, Berlin, Heidelberg, 2011

[12] U. Garain and B. B. Chaudhuri, "Segmentation of touching characters in printed devanagari and bangla scripts using fuzzy multifactorial analysis", *IEEE Systems, Man, and Cybernetics Society*, Vol. 32, No. 4, pp. 449-459, 2002.

[13] S. Bag, and G. Harit, "A survey on optical character recognition for Bangla and Devanagariscripts",*Sadhana*, Vol. 38, No.1, pp. 133-168, 2013

[14] U. Bhattacharya, S. K. Parui, and S. Mondal, "Devanagari and Bangla text extraction from natural scene images", *In: IEEE international conference on document analysis and recognition (ICDAR)*, pp. 171-175, 2009.

[15] R. Jayadevan, S. R. Kolhe, P. M. Patil, and U. Pal, "Offline recognition of Devanagari script- a survey", *IEEE Trans Sys Man Cybern Part C Appl Rev*, Vol. 41, No.6, pp. 782-796, 2011

[16] H.Ma,and D.Doermann, "Adaptive Hindi OCR using generalized Hausdorff image comparison", *ACM Trans Asian Lang Inf Process*, Vol. *2*, No. 3, pp. 193-218, 2003

[17] OVR. Murthy, S. Roy, V. Narang, M. Hanmandlu and S. Gupta , "An approach to divide pre-detected Devanagari words from the scene images into characters", *Signal Image Video Process* Vol. 7, No. 6, pp. 1071-1082, 2013

[18] V. Narang, S. Roy, OVR. Murthy, and M. Hanmandlu, "Devanagari character recognition in scene images", *In IEEE international conference on document analysis and recognition (ICDAR)*, pp. 902-906, 2013.