Pre-segmentation in Offline Handwritten words

Monika Kohli[1] and Satish Kumar[2]

[1*]Research Scholar Department of Computer Science and Applications, Panjab University, Chandigarh.
[2]Associate Professor Department of Computer Applications Panjab University,SSG Regional Centre Hoshiarpur, Punjab.

e-mail: monikakrajotia@gmail.com, satishnotra@yahoo.co.in
Corresponding Author: monikakrajotia@gmail.com

Abstract: The paper deals with extraction of individual components from handwritten word image consisting of shadowed characters and defining criteria to identify touching and non-touching component in a word image. Touching characters need to be segmented appropriately in order to recognize. In case, if word contains touching characters, pre-segmentation and segmentation are the two required phases. Pre-segmentation assures that each segment should contain either a single character or touching components. The resultant segment need to identify as touched or isolated. Isolated component extracted from the word image can be recognized after feature extraction. Identified touching components of the image need further segmentation so that it can be segregated into individual components. Objects area criteria is used to embark upon the problem of shadowed characters. This paper also proposed analytic approach based technique- PCSW (Pixel Continuity Slope and Width technique) which help in differentiating between touching and non-touching(isolated) characters in a word image. The database for experimentation consists of legal amount words containing touching characters consisting of 1530 words by 15 different writers(db1) and 250 words dataset taken from database provided by ICDAR(db2). Implementation of PCSW achieved the accuracy of 98.16% and 96.80% on db1 and db2 respectively.

Keywords-PCSW, Segmentation, Shadowed characters, Touching characters.

(Received October 1st, 2019 / Accepted November 25th, 2019)

1. INTRODUCTION

Complex dataset to Devanagari script(Hindi) and variations in handwritten data makes the task of optical recognition complex. Pre-processing, segmentation, feature extraction and recognition are the various phases of optical character recognition. Segmentation phase of Optical Character Recognition (OCR) requires a technique to segment word string into individual components. If word contains touching characters, pre-segmentation and segmentation are the two required phases. Pre-segmentation assures that each segment should contain either a single character or touching components. The resultant segment need to identify as touched or isolated. Isolated component extracted from the word image can be recognized after feature extraction. Identified touching components of the image need further segmentation so that it can be segregated into individual components. In case of touching characters, components need to be segregated without artifacts of the neighbouring components into individual components. In literature, VPP is often used for character segmentation. D.V.Sharma and G.S.Lehal in[1] used Vertical Projection Profile (VPP) to identify middle zone characters. S. Ramachandrula in[2] also used VPP for segmentation after taking skeleton of the image. Authors also looks into the over-segmentation and under-segmentation problem. V.P.Dhaka and M.K.Sharma[3] used pixel based segmentation technique- Pixel Plot and Trace and Replot and Re- trace (PPTRPRT) for word and character segmentation in handwritten text. Ladwani et. al.

in [4] used morphological operations for pre-processing and segmentation. Segmentation of touching characters on the basis of joint points is given by S. Kapoor and V. Verma in[5].Touching characters needs to be segmented properly for increasing the recognition accuracy but prerequisite is to find the suitable candidate for the same. Various techniques like Average width, aspect ratio, water reservoir principle are used in the literature to find the presence of touching characters. To identify touching and non-touching components, Water Reservoir technique was introduced by U.pal et.al. [6] for the segmentation of unconstrained handwritten connected numerals. It uses Statistical analysis based projection profiles technique for the identification of a touching character and water reservoir based technique for the segmentation of existing touching characters. This technique was later used for segmentation of handwritten text[7]. Little progress has been seen in segmentation of touching characters in Devanagari script[8][9][10] as compared to other languages like Bangla[6][7][11], Kannada[12][13][14], Punjabi[1][15] . Complex character set makes the task of segmentation and thus recognition challenging. Unavailability of benchmark database for touching characters adds more to the challenge. There is a room for research in this area. This paper is organized as follows. Section II of the paper describes the preprocessing of the word image under consideration and the segmentation issue in case of shadowed characters. Algorithm-A1 is proposed which facilitates the extraction of individual components in case of shadowed characters. This paper also deals with the identification of isolated and touched characters which is an analytic approach based technique- Pixel Continuity Slope Width technique (PCSW). Database used and experimental results are reported in Section III. Conclusion and future scope is given in Section IV.

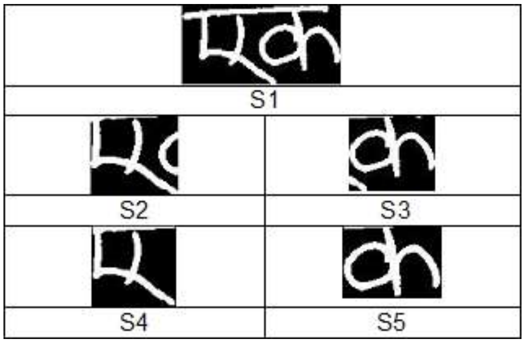


Figure 1. Example of Shadowed Characters

2. PRE-PROCESSING

The first phase is the Pre-processing of handwritten word image to eliminate those elements from the image which are not required in further processing phases. Preprocessing techniques like Binarization, smoothing, noise removal are applied for pre-processing. Techniques of pre-processing are applied on the image captured with scan resolution of 300 dpi. Binarization is performed using im2bw function using threshold value 0.9, noise removal using median filter approach.

2.1 PRE-SEGMENTATION

After pre-processing, segmentation of individual components is required. Headline is one of the characteristic of Devanagari Script(Hindi). Headline connects the middle and upper zone characters except in cases like Bindu(.) and Chandrabindu(◌ँ) . Handwritten data may or may not be joined properly using headline as headlined can be skewed or absent. Variation in number of pixels in each column belonging to headline adds more difficulty in detecting the headline. The isolated components extracted in case of shadowed characters include the neighboring components which is undesirable. The issue of shadowed characters is resolved in this paper using algorithm-A1 as given in section 2.1.1.

2.1.1 Shadowed characters

The sample given in image S1 in Figure 1 is the handwritten sample from the database which consists of shadowed characters. Word string is segmented using connected components approach. Segmentation involves dividing the word image into smaller parts in order to obtain probable character boundaries. An ideal segmentation gives the exact character boundaries, where a character lies in exactly one segment. However, this ideal segmentation is not easily achievable. In Figure 1, S2 and S3 shows the result of segmentation in the word image given in S1. Both segments i.e. S2 and S3 include neighbouring component unit or sub-unit. Segment S1 adds sub-unit of the right neighbour and segment S3 adds sub-units of the left neighbour to their respective bounded region. The algorithm-A1 is used to remove these artifacts and the results thus obtained are shown in S4 and S5.

Algorithm A1:

Step 1: Calculate number of components (nc) in a bounded region(br) of image(img).

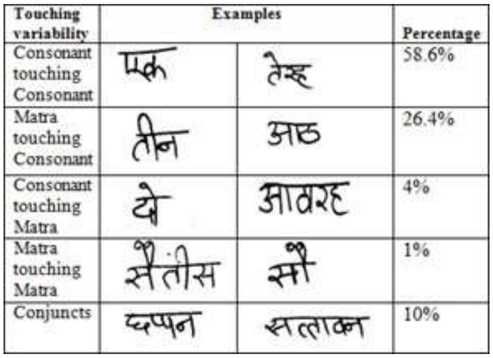
Step 2: If nc>1, then find area(nci), where i=1 to max(nc).

Step 3: Find nci with x(area(nci))

Step 4: Save nci.

2.1.2 Identifying character as isolated or touching character using PCSW (Pixel Continuity Slope Width technique)

After segmentation of word string into sub-images or units, the need is to identify components as non -touched (isolated) or touched characters. Isolated characters are one which does not require any further segmentation and are fit for the next phase i.e. features extraction and recognition. But in handwritten data, strokes often touch each other while writing. It results in touching components which is not a part of the character set of a script. Touching components requires a segmentation technique to separate into individual components. Two hypotheses are considered to differentiate touched and non-touched characters. One is the width of the character if greater than average width. Other is by analyzing the vertical pixel continuity to identify vertical stroke in the starting or ending of component. In order to identify the isolated and touching components, touching variability of the components are analyzed on the legal amount word database in this paper. Percentage of touching components variability is calculated for 1780 words and results are given in Figure 2.



Touching variability	Examples		Percentage
Consonant touching Consonant	एक	तेरह	58.6%
Matra touching Consonant	तीन	आठ	26.4%
Consonant touching Matra	दो	आठरह	4%
Matra touching Matra	सैंतीस	सौ	1%
Conjuncts	छप्पन	सत्तावन	10%

Figure 2. Percentage of Touching components in Handwritten dataset of legal amount words

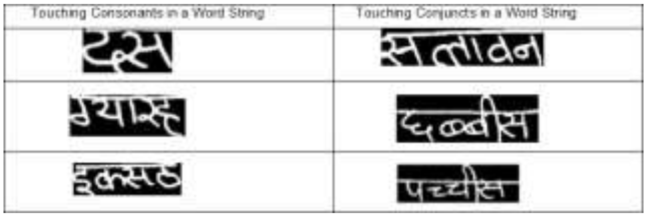

Touching Consonants in a Word String
Touching Conjuncts in a Word String
दस
सत्तावन
ग्यारह
छब्बीस
इकसठ
पच्चीस

Figure 3. Touching consonants and conjuncts

It has been noticed that there is high percentage of variability i.e. 58.6% in which two consonants touch each other. 26.4% of variability in which preceding matra (◌ा) or as a part of other matra like (◌ि, ◌ी, ◌ो, ◌ौ) touches a consonant and 10% in case of conjuncts.4% variability in case preceding consonant touches matra (◌ा) or as a part of other matra like (◌ि, ◌ी, ◌ो, ◌ौ) which is considered as case3.

PCSW (Pixel Continuity Slope Width technique) is applied on 1780 legal amount words which include touching components in the middle zone of word image. 1530 legal amount words are collected from 15 writers of different age group and gender and 250 words are collected from database provided by ICDAR for experimentation.

Results reported in the literature examines conjuncts and touching consonants using the average width in Devanagari script (Hindi) but the approach fails considering other touching variability(Figure 3). Following are the 3 cases on the basis of which we will examine the touched or isolated component. Case1 deals with touching consonants and conjuncts. Case2 and Case3 are considered as segmented components starting and ending with vertical line are neglected by average width criteria. A segmented component in which vertical line is present in the beginning is discussed in Case2. Case3 is discussed to find the vertical line in the end of the segmented component in which detected vertical line is distinguished as a matra (◌ा) or part of matra (◌ि, ◌ी, ◌ो, ◌ौ) in the middle zone touching a character or a part of character itself.

Case 1: If two consonants in middle zone are touching each other or if conjuncts are considered, then average width criteria is used. Case 2: If a matra(◌ा) or middle zone part of matras like (◌ि, ◌ी, ◌ो, ◌ौ) preceding a consonant, touch each other(Figure.4, Figure. 5). Character set of Devanagari script (Hindi), consists of approximately 66%characters that ends with vertical line (◌ि, ◌ी, ◌ो, ◌ौ) These characters are rarely found to be touched with matra(◌ा) or part of matra (◌ि, ◌ी, ◌ो, ◌ौ). Analyzing legal amount words database figure out that remaining 34% characters are more prone to get touched with other component and results in touching components.

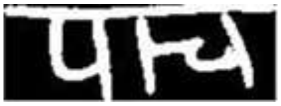


Figure. 4

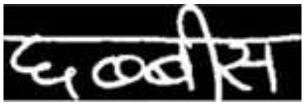


Figure. 5

Case 3: If consonant preceding a matra(◌ा) or middle zone part of matras like (◌ि, ◌ी, ◌ो, ◌ौ) touch each other(Figure. 6, Figure. 7). After finding contiguous vertical pixels in the end of the segmented component, a column wise segmentation based on 3-neighbour pixels of the skeletonized segmented component is done to obtain two segments. Left segment thus obtain is without the vertical stroke which is fed into the recognizer. Recognition results are obtained using Convolution Neural Network(CNN) using python interface in MATLAB R2015b. The segmented images are tested using CNN model. The model is trained using 103700 training images and 18300 test images of size 32x32. The result thus obtained, if does not belong to class containing half consonants, then it is treated as a touched segmented component else it is treated as a character in which vertical line is a part of character itself.

Technique proposed in this paper, used PCSWD (Pixel Continuity Slope Width and Density Technique) which considers conjuncts, touching consonants and matra touching the consonant in a word image. Algorithm-A2 results are shown in Figure 8.

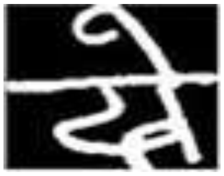


Figure. 6

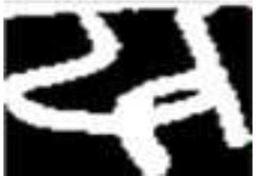


Figure. 7

Algorithm-A2 - PCSW(Pixel Continuity Slope Width Technique)

```
r-Row
c -Column
Cn -No. of characters
Aw -Average Width of character
Cj-Conjuncts (half character touches full character)
Cs -Touching consonants
Cl1-Matra touching character or matra in the beginning of the segment.
Cl2-Character touching matra or matra in the ending of the segment.
Vd -Vertical line .
Wc-Width of a character.
imc-part of the image under consideration.
Vh-Continuous pixels height.
imh-Height of the image.
Vs-Continuous pixels slope.
ims-Slope of the image.
imskel-Skeletonized image.
Pnv-the nearest pixel with 3- neighbouring pixels to the left of Vh.
Vnc -closest pixel to vertical line.
vr-row Vnc.
vc-column Vnc.

Step 1: If Aw<Wc (im), then the cropped image is Cs or
Cj.

%Average Width(Aw)= (Width/(no. of character)-1) is
calculated.

Step 2: else if Wc (im)>threshold[threshold is calculated
on the basis of stroke width]
imc=(1:r,1:0.25*c)
Vh=max(hough(imc))
if Vh> 0.80*imh && Vs> threshold, then the cropped
image is Cl1.

% Vertical line detection in the starting of the image
considering height and slope.

Step 3: else if Wc (im)>threshold[threshold is calculated on
the basis of stroke width]
imc=(1:r, 0.75*c: max(c))
Vh=max(hough(imc))
if height(Vh)> 0.80*imh and Vs> threshold, then the cropped
image is Cl2.

%Vertical line detection in the ending of the image
considering height and slope.

 imskel = bwmorph(imc ,’skel’) ;
%Skeletonization is used to obtain pixels with 3 neighbouring
pixels.

find neigh_pix=neighbor(3);
[rp cp]=neigh_pix
Vnc=find (neig_pix==max(cp))
[vr vc]=Vnc
Find Vnc (vc)<min(Vcline)

% Vnc, closest pixel to the left of  Vline to divide image into
left and right segment.

limg =crop imc( start column, start row, vc, max(vr))
class=CNN(limg); %if class belongs to half consonants, then
non-touched, else touched.
```

3. EXPERIMENTAL RESULTS

The handwritten character database is the character images scanned with HP scannerwith300dpi resolution.

The experiment is performed in MATLABR2015b under Microsoft Windows environment with X86 based PC, 2.40GHz CPU and 4GB RAM.

3.1. Database used for Experiment

Two databases are used for experimentation. One database consists of legal amount 1530 words by 15 different writers and other dataset consisting of 250 words from legal amount words database provided by ICDAR. These 250 words are selected manually on the basis of presence of touching components. PCSW (Pixel Continuity Slope Width technique) is verified manually.

3.2. Results

The experimental evaluation of PCSW is verified manually. Table 2 shows the accuracies obtained with both datasets(db1 and db2). It is observed that 98.16% accuracy is achieved using PCSW (Pixel Continuity Slope Width technique) with db1 and 96.80% with db2.

S. No.	Database	Words	Result
1	db1	1530	98.16%
2	db2	250	96.80%

Table 2. Accuracy of PCSW technique with db1 and db2.

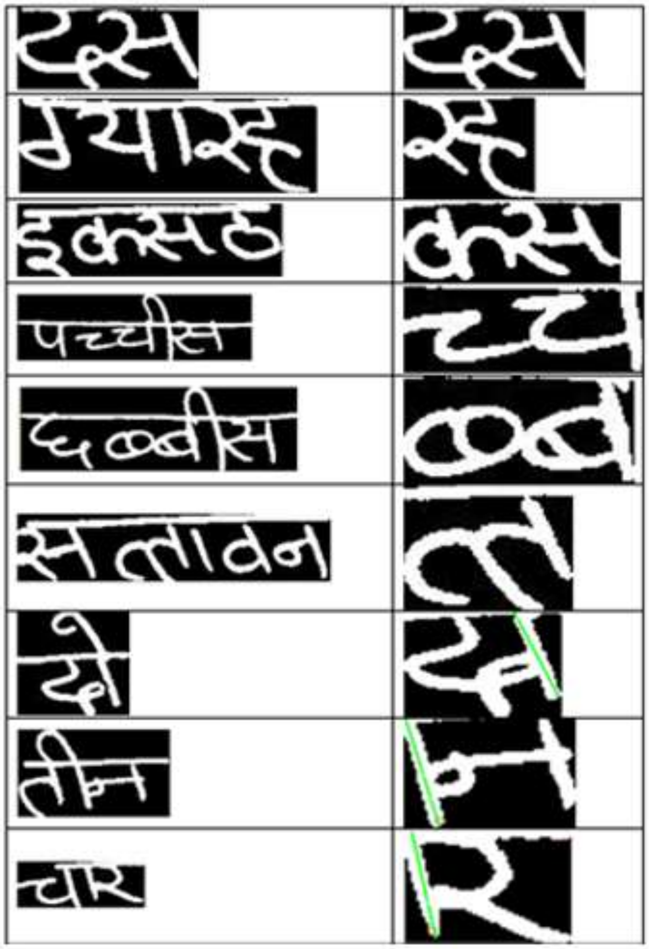


Figure 8. Algorithm-2 Output

4. Conclusion

This paper focused on essential part of Offline handwritten optical character recognition in Devanagari script i.e. Pre-segmentation phase. This phase assures that each segment should contain either a single character or touching components. The resultant segmented sub-unit thus obtained, need to identify as isolated or touching component. Future work will focus on the segmentation of touching characters. Properly segmented character is a challenging task in Offline Handwritten Devanagari Script (Hindi) and very few papers in this regard are available. The proposed method can be extended to include the character segmentation in words containing overlapping, touching and broken characters in the printed and handwritten document as well.

References

[1] D. V. Sharma and G. S. Lehal, An iterative algorithm for segmentation of isolated handwritten words in Gurumukhi script in Proceedings - International Conference on Pattern Recognition,2006, vol. 2, pp. 1022–1025.

[2] R. H. Ramachandrula, Sitaram. Jain Shrang,Offline Handwritten Word

Recognition in Hind in Proceeding of the workshop on Document Analysis and Recognition, 2012, no. ii, pp. 49–54.
[3]V.P.DhakaandM.K.Sharma,An efficient segmentation technique for Devanagari offline handwritten scripts using the Feedforward Neural Network Neural Comput. Appl., vol. 26, no. 8, pp. 1881–1893, 2015.
[4] V.M.Ladwani and L.Malik, Novel approach to segmentation of handwritten Devnagari word Proc.- 3rd Int.Conf. Emerg.Trends Eng.Technol.ICETET2010,pp.219–224, 2010.
[5] S. Kapoor and V. Verma, Fragmentation of handwritten touching characters in devanagari script Int. J. Inf. Technol. Model. Comput., vol. 2, no. 1, 2014.
[6] U. Pal, A. Behaid, C. Choisy, Touching numeral segmentation using water reservoir concept Pattern Recognit. Lett., vol. 24, no. jan, pp. 261–272, 2003.
[7] U. Pal and S. Datta, Segmentation of Bangla unconstrained handwritten text in Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 2003, vol. 2003–Janua, no. ICDAR, pp. 1128–1132.
[8] B. Shaw, S. K. Parui, and M. Shridhar, Offline handwritten Devanagari word recognition: A holistic approach based on directional chain code feature and HMM Proc. - 11th Int. Conf. Inf. Technol. ICIT 2008, pp. 203–208, 2008.
[9] B. Shaw, A Segmentation Based Approach to Offline Handwritten Devanagari Word Recognition Pattern Recognit., pp. 256–257, 2008.
[10] A. S. Ramteke and M. E. Rane, Offline Handwritten Devanagari Script Segmentation Int. J. Sci. Res., vol. 1, no. 4, pp. 142–145, 2012.
[11] U. Garain and B. B. Chaudhuri, Segmentation of touching characters in printed devnagari and bangla scripts using fuzzy multifactorial analysis in Proceedings of the InternationalConferenceonDocumentAnalysisandRecognition,ICDAR,2001,vol.2001–Janua,no.4,pp.805–809
[12] M. Venkatesh, V. Majjagi, and D. Vijayasenan, Implicit segmentation of Kannada characters in offline handwriting recognition using hidden Markov models, Implicit arXiv Prepr. arXiv1410.4341 (2014)., pp. 1–6, 2014.
[13] C.Naveena and V.N.Manjunath Aradhya, Handwritten character segmentation for Kannada scripts Proc. 2012 World Congr. Inf. Commun. Technol. WICT 2012, pp. 144–149, 2012.
[14] H. R. Mamatha and K. Srikantamurthy, Morphological Operations and Projection Profiles based Segmentation of Handwritten Kannada Document Int. J. Appl. Inf. Syst., vol. 4, no. 5, pp. 13–19, 2012.
[15] M. Kumar, M. K. Jindal, and R. K. Sharma, Segmentation of Isolated and Touching Characters in Offline Handwritten Gurmukhi Script Recognition Int. J. Inf. Technol. Comput. Sci., vol. 6, no. 2, pp. 58–63, 2014.