

Segmentation of Touching Characters in Devanagari

Veena Bansal and R. M. K. Sinha

veena@iitk.ernet.in rmk@iitk.ernet.in

Department of Computer Science and Engineering
Indian Institute of Technology, Kanpur 208016 India

Abstract- This paper presents an algorithm for segmentation of touching Devanagari characters (also referred to as conjuncts) into its constituent symbols and characters. Proposed algorithm extensively uses structural properties of the script. Statistical information about the height and width of character boxes, which are vertically separate from their neighbours, is used to hypothesize character boxes to be touching character boxes. The recognition rate of 85% has been achieved on the segmented touching characters.

1 Introduction

Many algorithms have been proposed for segmentation of touching characters [4, 5, 6, 7, 8] for Roman script.

In this paper, we have considered the problem of conjunct segmentation in the context of Devanagari script¹. Consider the word **राष्ट्रीय** which is pronounced as ‘rashtriya’. The preliminary segmentation phase based on projections (vertical and horizontal) yields the following units:

र | ष्ट | य and the top modifier ण .

This process does not separate lower modifiers and leaves the conjuncts unsegmented. Some additional touching characters used in Devanagari are shown in figure 1. Statistical analysis of height and width information of the character boxes indicate that ष्ट is a

possible composite character and has a possible lower modifier. The output of the classification process is also taken into consideration before attempting further segmentation of the possible composite character. In this case, the image box ष्ट is rejected by the classification process [9]. Therefore, image box ष्ट is further segmented to obtain the following units:

दु ट and the lower symbol , .

In section 2, the algorithm for segmentation of touching characters is presented. The algorithm for lower modifier separation is discussed elsewhere [1]. In section 3, we present results of our investigation followed by conclusion in section 4.

Touching characters	Constituent characters
ष	दु
ह	ण
ख	स
म	स
न	ण
म	दु

Figure 1: Some of the Touching characters used in Devanagari script.

2. Segmentation of Touching Characters

We refer to the conjunct image by image(conj_left, conj_right, conj_top, conj_bottom) where conj_left, conj_right, conj_top and conj_bottom are the left, right, top and bottom coordinates of the minimum sized upright bounding box of the image. Before discussing the algorithms, we need to define the follow-

¹Devanagari is the script for Hindi which is official language of India. It is also the script for Sanskrit, Marathi and Nepali languages. It is used by more than 300 million people on the globe.

junct is reached or the pixels strength in the next column is more than the pixel strength in the column under examination.

In case of class H_2 characters, we look at the left one third of the image and locate the right most column which contains 50% or more black pixels of conjunct height. Starting from this column to the middle column of the conjunct image, we examine each column. We stop when the pixel strength in the next column is more than the pixel strength in the column under examination or the middle column is reached. The present column is the right boundary column of the first constituent character. Please refer to figure 2(c).

For extracting the second constituent character of the conjunct, the continuity of the *collapsed horizontal projection* is checked. The *collapsed horizontal projection* corresponding to a Devanagari character image has continuity. In case of *end bar* and *middle bar* characters, the *collapsed horizontal projection* corresponding the character image to the left of the bar also has continuity. However, if we vertically chop δ columns from the image from the left, where δ is a positive integer greater than or equal to the *penwidth*, continuity of the corresponding *collapsed horizontal projection* is lost.

The search for the left boundary of the second constituent character starts from *conj_right* if no bar is present and immediately to the left of the bar if it is present. This point is referred to as *temp_conj_right*. We make an initial guess for the left column of the second constituent character. We refer to this column to as *temp_left2* and set it to the $temp_conj_right - 2 * penwidth$. We now make a collapsed horizontal projection of the image enclosed by *temp_conj_right*, *temp_left2*, *conj_top*, *conj_bottom*. This is referred to as *HPr*. If *HPr* has no discontinuity and the height

of *HPr* is more than 1/3rd of *conj_height*, *temp_left2* becomes the left boundary of the second constituent character (referred to as *left2*). Otherwise, we move *temp_left2* to further left in steps of one column and modify *HPr*. We stop moving *temp_left2* further if *HPr* has no discontinuity and required number of rows are present. The present value of *temp_left2* becomes the left boundary of the second constituent character. However, if *temp_left2* has reached *left_onethird* and still a break column has not been located, the search is abandoned and no segmentation point is suggested. The magnified conjunct images and their segmentation have been depicted in figure 2. Please refer to figure 2(e) and (f).

If *left2* is less than *right1*, both the segmentation points are ignored and no segmentation point is suggested. If *right1* and *left2* are same or the difference ($right1 - left2$) is same or less than double of pen width, the segmentation is accepted. However, the *right1* is moved to the left by the pen width. Please refer to figure 2(g).

3 Results The structural segmentation algorithm has been tested on 18 document pages from two different magazines. The number of conjuncts in the test documents are about 5%. Out of all the conjunct and composite characters, about 90% conjuncts has been marked for further segmentation based on the output of the classification process (see [9, 10, 3] and width and height information of characters. During testing, it was observed that sometimes a composite character was substituted by another Devanagari character. The recognition of the constituent characters obtained after segmentation of touching character is about 88% which matches with the overall performance of the system [3]. The results of touching character segmentation are summarized in tables 1. However, the algorithm is capable of segmenting all

	total number of chars.	overall recognition	number of touching chars.	touching char. recognition	touching chars. substitution
Font I	20856	18162 87.08%	1172 5.61%	981 83.70%	191 16.29%
Font II	16356	14359 87.79%	802 4.90%	678 84.53%	124 15.46%

Table 1: Performance of the system for Font I and II.

the conjuncts when suggested to do so.

A sample text page and the text after recognition is shown in figure 3. The segmented image after preliminary segmentation and lower modifier segmentation is also shown in the same figure. The image after conjunct segmentation and the OCR output after post-processing [9] are also presented in the same figure.

4 Conclusion We have introduced a new method for the segmentation of the conjuncts for Devanagari script. This strategy is based on the structural properties of the script. The right and left part of the images are extracted independently.

Acknowledgments A part of this work was supported by TDIL sponsored project.

References

- [1] Veena Bansal and R.M.K. Sinha, *Segmentation of touching and fused Devanagari characters*, Technical Report, TRCS-97-247, I.I.T. Kanpur, India, 1997.
- [2] Veena Bansal and R.M.K. Sinha, Partitioning and Searching Dictionary for Correction of Optically-Read Devanagari Character Strings, *Technical Report*, TRCS-97-246, I.I.T. Kanpur, India, (1997).
- [3] Veena Bansal and R.M.K. Sinha, *Integrating Knowledge Sources in Devanagari Text Recognition System*, Technical Report, TRCS-97-248, I.I.T. Kanpur, India, 1997.
- [4] R.G. Casey and G. Nagy, Recursive segmentation and classification of composite character patterns, *Proc. 6th International Conference Pattern Recognition, Munich, Germany* 1023-1026 (1982).
- [5] S. Kahan, T. Pavlidis and H. S. Baird, On the recognition of printed characters of any font and size, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9, 274-287 (1987).
- [6] Su Liang, Shridhar, M. Ahmad, Segmentation of Touching Characters in Printed Document Recognition, *Pattern Recognition*, 27, 825-840 (1994).
- [7] Shuichi Tsujimoto and Haruo Asada, Resolving Ambiguities in Segmenting Touching Characters, *Structured Document Image Analysis*, H. S. Baird et. al., eds., Springer-Verlag, U.S.A. (1992).
- [8] Jairo Rocha and Theo Pavlidis, A Shape Analysis Model with Applications to a Character Recognition System, *IEEE Transactions on PAMI*, 16(4) 393-403, (1994).
- [9] R. M. K. Sinha and Veena Bansal, On Devanagari Document Processing, *IEEE International Conference on Systems, Man and Cybernetics, Vancouver, Canada* (1995).
- [10] R.M.K. Sinha and Veena Bansal, On automating trainer for construction of prototypes for Devanagari text recognition, *Technical Report*, TRCS-95-232, I.I.T. Kanpur, India, (1995).

(a) The Input Text

भारत में गरीबी ग्रामीण क्षेत्र में ही विद्यमान है ऐसी धारणा चिन्तन के स्तर पर मानी जाती रही है। इसीलिए गरीबी उन्मूलन के अधिकांश कार्यक्रम ग्रामोन्मुख रहे हैं। गरीबी की विकरालता तथा विशालता निस्सन्देह ग्रामीण क्षेत्रों में विराट रूप से परिलक्षित होती है, किन्तु शहरी क्षेत्र इससे अछूता हो, ऐसा नहीं है। शहरी क्षेत्रों में

(b) The image after preliminary segmentation and lower modifier separation

भ | र | त | म | ग | र | ब | ी | ग | र | अ | म | प |
क्ष | त्र | म | ह | । | व | द्य | म | न | ह | ए | स | ।
ध | र | ण | ।
। च | न्त | न | क | स्त | र | प | र | म | न | । | ज | त | ।
र | ह | । | ह | । | इ | स | । | ल | ए | । | ग | र | । | ब | ।
उ | न्म | ल | न | क |

(c) The image after conjunct segmentation

भ | र | त | म | ग | र | ब | । | अ | म | प | क्ष | त्र | म
ह | । | व | द्य | म | न | ह | ए | स | । | ध | र | ण | ।
। च | न्त | न | क | स्त | र | प | र | म | न | । | ज | त | ।
र | ह | । | ह | । | इ | स | । | ल | ए | । | ग | र | । | ब | । | उ | न्म | ल | न
क

(d) The output of the classification process; the word has been underlined if the true word is the second or third choice

भारत में गरीबी ग्रामीण क्षेत्र में ही विद्यमान है ऐसी धारणा चिन्तन के स्तर पर मानी जाती रही है। इसीलिए गरीबी उन्मूलन के अधिकांश कार्यक्रम ग्रामोन्मुख रहे हैं। गरीबी की विकरालता तथा विशालता निस्सन्देह ग्रामीण क्षेत्रों में विराट रूप से परिलक्षित होती है किन्तु शहरी क्षेत्र इससे अछूता हो ऐसा नहीं है। शहरी क्षेत्रों में

Figure 3: