Character Segmentation of Handwritten Bangla Text by Vertex Characterization of Isothetic Covers

Soumen Bag, Partha Bhowmick
Computer Sc. & Engg. Department
IIT Kharagpur, India
{*soumen, pb*}*@cse.iitkgp.ernet.in*

Gaurav Harit
Computer Sc. & Engg. Department
IIT Rajasthan, India
gharit@iitj.ac.in

Arindam Biswas
Computer Sc. & Engg. Department
BESU, Shibpur, India
barindam@gmail.com

***Abstract*—Segmentation of cursive handwriting is one of the most challenging problems in the area of handwritten character recognition. In this paper, we propose a novel approach towards character segmentation in a handwritten document. It is based on the vertex characterization of outer isothetic polygonal covers so that each cover corresponds to a particular word or part of a word. The proposed method has the potential to segment skewed text without deskewing them. Experiment is done on several Bangla handwritings of different individuals. The average success rate is 96.04%. This method can be considered as a significant preprocessing step towards the development of a handwritten Bangla OCR system.**

***Keywords*-Bangla handwriting; Character segmentation; Isothetic cover; OCR.**

I. Introduction

The segmentation of characters in a word for optical character recognition (OCR) should be such that each segment resembles a character. Features are extracted from these components and classified as particular characters belonging to the existing character set. Many works have been done on text line, word, and character segmentation for printed documents of Indian languages [1]. But segmentation of cursive handwriting still remains one of the most challenging problems in the area of handwritten character recognition. The inter-line distance variability, baseline skew variability, and large-scale similarity/dissimilarity in the shape of different characters written by different individuals render the problem even more difficult. Hence, over the last decade, different techniques such as, recursive contour following [2], water reservoir [3], graph-based [4], and fuzzy features with MLP classifier [5, 6] have been reported for our specific domain, i.e., handwritten Bangla (also called 'Bengali') character segmentation.

We see that the present methods perform skew detection and correction before character segmentation. They also use training-based model and different heuristics based on the prior knowledge of the various handwritings for character segmentation. Apart from that, very few handwritten OCR methods have used character segmentation as an intermediate step. But character segmentation is one of the primary preprocessing steps for character recognition. In view of this,

Figure 1. Top: Three handwritten words; Middle & Bottom: Segmented characters with their corresponding outer isothetic covers (IOC) in red and blue respectively.

we propose a novel character segmentation method from Bangla handwritten text documents (Fig.1). The method is based on vertex characterization of the outer isothetic covers corresponding to words in a text. Using the vertex characterization, each cover is partitioned into several sub-polygons that represent the characters constituting a word. The novelty of the proposed method is that it can perform character segmentation of skewed documents without deskewing them and is applicable for handwritten Bangla text without incorporating any heuristics or a prior knowledge.

II. Proposed Methodology for Character Segmentation

The proposed handwritten character segmentation method has three stages: construction of outer isothetic cover (OIC), extraction of headline and baseline, and vertex characterization of isothetic covers for character segmentation.

A. Construction of Outer Isothetic Cover

Given a scanned document page, we binarize it using Otsu's algorithm [7]. Currently we are working with only text documents. To construct the outer isothetic cover, we apply the algorithm TIPS [8]. We impose an isothetic set of grid lines having the grid size $g = 1$ on the binarized image. Now, the grid points are traversed in the row-major order (object always lies left during traversal) until a 90^0 vertex ('start vertex') is found. If only one quadrant incident at a grid point has an object containment, then the particular grid point is marked as 90^0 vertex of the isothetic polygon (Fig. 2). There are another 15 different

978-0-7695-4599-8/11 $26.00 © 2011 IEEE
DOI 10.1109/NCVPRIPG.2011.12

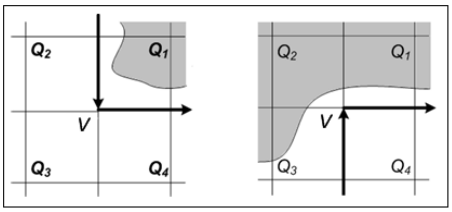

Figure 2. Left: 90^0 vertex; Right: 270^0 vertex [9].

arrangements considering object containments of these four quadrants. Detailed discussion is reported in [8]. All these vertex arrangements are skew angle free. If a document is written with a reasonable skew angle, then also the definition of these vertices will not be changed. This property helps us to handle skewed documents without deskewing them. After getting the 'start vertex', subsequent grid points are classified, marked as 'visited', and the direction is determined from each such grid point until the start vertex is reached. This completes the outer isothetic cover corresponding to a word or part of a word. The procedure iterated over the entire set of unvisited grid points subsequently derives the vertex sequences of all the isothetic covers corresponding to the words in the input document. Fig. 6(b) shows the example of outer isothetic covers (OIC) of a Bangla handwritten document in Fig. 6(a).

B. Extraction of Headline and Baseline

Headlines and baselines of the words are extracted from the OICs by analyzing their horizontal chords. The detailed methodology is reported in [9]. But this method performs not so good for few particular cases with grid size $g = 1$. Fig. 3 shows the failure of baseline extraction for a handwritten word. To overcome this problem, we have modified the algorithm. For detecting the headline of each, we search the maximum value of the sum of lengths of all equi-ordinate horizontal edges in the upper half of the OIC. For baseline detection, we search the maximum of the same in the lower half of the OIC. The improved result is shown in Fig. 3. The overall headline and baseline detection of the document in Fig. 6(a) is shown in Fig. 6(c).

C. Vertex Characterization of Isothetic Covers

Each OIC is represented by an ordered set of vertices $\mathcal{S} = \langle v_1, v_2, \ldots, v_n \rangle$ in an anticlockwise direction. At the time of constructing an OIC, we traverse the grid points in the row-major order to detect the first 90^0 vertex ('start vertex'). According to the construction rule of OIC (i.e., object always lies left during traversal), the traversal happens in an anticlockwise direction. Now, we analyze the vertex characteristics of the OIC to detect the segmentation points of each word image. The steps are given below.

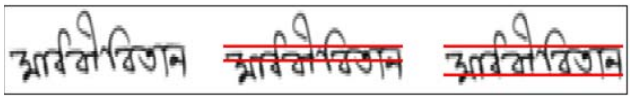

Figure 3. Left: Handwritten word; Middle & Right: Extraction of headline and baseline (in red) using method [9] and our proposed method respectively.

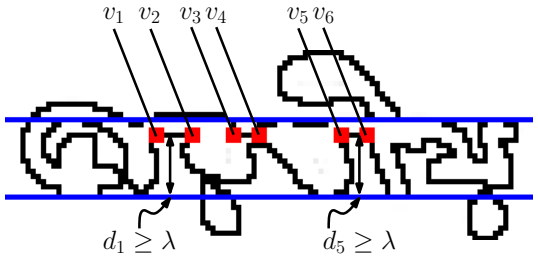

Figure 4. Detection of segmentation points on OIC of a word image.

1) Traverse all the vertices in the set $\mathcal{S}$ to detect 270^0 vertices of the OIC. From them, consider the ordered set of vertices lying in between the headline and the baseline of the OIC to prepare the set $\mathcal{S}' \subset \mathcal{S}$. This approach discards the modifiers or parts of the modifiers for further consideration.
2) After detecting the 270^0 vertices, we partition the set $\mathcal{S}'$ into two subsets, namely $\mathcal{S}_h$ and $\mathcal{S}_b$, depending on the isothetic path lengths of the vertices in $\mathcal{S}'$ from the headline and baseline respectively. The *isothetic path length* of a vertex $v \in \mathcal{S}'$ from the headline (baseline) is given by the length of the traversed path (in anticlockwise direction) from v through the successor vertices of the OIC in $\mathcal{S}$ up to the grid point where it meets the headline (baseline) for the first time. After the partition, for each vertex in $\mathcal{S}_h$ ($\mathcal{S}_b$), its isothetic path length from the headline (baseline) is less than that from the baseline (headline). The procedure is as follows.
After detecting a 270^0 vertex v, we start traversing from v according to the sequence of vertices in set $\mathcal{S}$. If we reach the headline earlier than the baseline, then we insert the vertex v in the set $\mathcal{S}_h$; otherwise, in the set $\mathcal{S}_b$. In principle, we are trying to detect the valleys in a word image looking from the baseline. For this purpose, we consider the set $\mathcal{S}_b$ for the next step. In Fig. 4, vertices v_1, v_2, v_3, and v_4 (marked in red) are four 270^0 vertices in which $\langle v_1, v_2 \rangle$ and $\langle v_3, v_4 \rangle$ belong to the set $\mathcal{S}_b$ and $\mathcal{S}_h$ respectively. So, from the next step we consider only $\langle v_1, v_2 \rangle$ and discard $\langle v_3, v_4 \rangle$ for further computation.
3) From the set $\mathcal{S}_b$, we find two consecutive 270^0 vertices having the same height from the baseline. There are 4 different arrangements (considering object containments of four quadrants) of this type of vertex. In Fig. 5, these arrangements are marked as A_1, A_2, A_3, and A_4. From these 4 arrangements, we can get 6 different combinations by taking any two of them as a

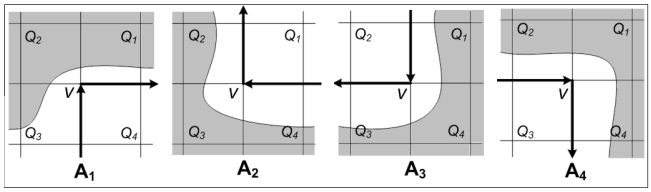

Figure 5. 4-different arrangements of 270^0 vertex.

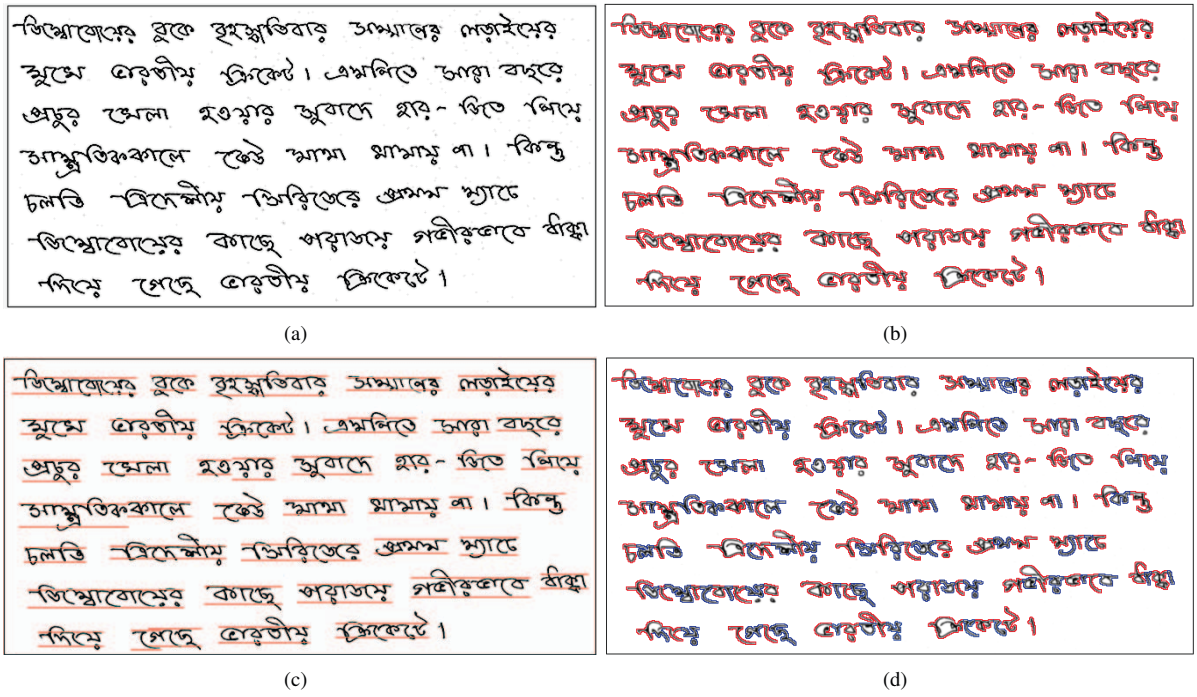

Figure 6. Handwritten Bangla character segmentation. (a) Bangla handwritten text; (b) Outer isothetic covers (OIC, in red) of text document with grid size $g = 1$; (c) Headline and baseline (in red) of segmented words; (d) Segmented characters are represented with their corresponding OICs marked by alternative colors.

pair. Interestingly, out of these six combinations, only one combination (A_1, A_4) yields a *valley*, which is defined as two consecutive vertices from $\mathcal{S}_b$ having same y-coordinate. Whatever be the writing style in Bangla, for the OIC of each word there are valleys parallel to the headline of the word (Fig. 4). So, we detect these valleys by verifying the pairs $\langle A_1, A_4 \rangle$ from $\mathcal{S}_b$. For example, in Fig. 4, the pairs $\langle v_1, v_2 \rangle$ and $\langle v_5, v_6 \rangle$ satisfy (A_1, A_4) combination, thus resulting two valleys.

4) After detecting the above pairs, we measure the distance, d of a vertex of each pair from the baseline along y-axis. If $d \geq \lambda$, where λ is half the distance between the headline and the baseline, then we partition the OIC by a diagonal joining the midpoint of the valley and a joint on the OIC lying immediately above it.
5) Repeat the above steps for all remaining OICs to get the set of segmented characters. As we have constructed the OIC with grid value 1, initially it helps us to get isolated covers for few characters those have large intermediate gaps between them. So, if we find any unvisited vertex, then this vertex belongs to the isolated cover. So, we take that vertex and traverse the whole outer cover to get the segmented character.

Fig. 6(d) shows the character segmentation output of a handwritten Bangla text document in Fig. 6(a). The segmented characters are represented with their corresponding OICs marked by alternative colors.

III. Experimental Results

We have implemented the algorithm in C using OpenCV 2.0 in Linux platform. The proposed method is tested on a variety of Bangla handwritten documents. The sample images are not slant- and slope-corrected, and written in black ink on white paper. Each such image is digitized using a HP scanjet 5590 scanner at 300 dpi resolution. Result of character segmentation on a Bangla handwriting has been already shown in Fig. 6 with necessary explanations. For experimental purpose, we have taken 6 different Bangla texts written by 10 different individuals. The total number of samples of handwritten words is 14,650. Fig. 7 shows a result of character segmentation for a small part of a Bangla text written by different persons with 96.42% average success rate. A part of the detailed experimental results are tabulated in Table I. Here we have reported the results of 6 different handwritten texts written by first 5 persons. The performances of the remaining 5 persons are also similar compared with the reported results. The average success rate for our whole dataset is 96.04%.

The proposed method does not perform well for few pathological cases. Sometimes it suffers from under-segmentation and sometimes from over-segmentation. When a lower portion of a basic character touches with the lower portion of some other character or modifier, the outer polygonal cover is constructed by considering these two characters as a single character. In Fig. 8(a), the lower part of the first character touches the lower portion of a vowel modifier. It performs over-segmentation for the characters 'aa' (Fig. 8(b)), 'bor-gio-jo' etc. It breaks these characters

Table I
DETAILS OF THE PERFORMANCE OF SEGMENTATION TECHNIQUE ON THE DIFFERENT HANDWRITTEN TEXTS.

Text id	Total no. of words	Total no. of characters	Person id	No. of faulty segmented characters	Success rate (%)
T_1	95	289	P_1	10	96.54
			P_2	12	95.85
			P_3	11	96.19
			P_4	11	96.19
			P_5	12	95.85
T_2	165	488	P_1	18	96.31
			P_2	21	95.70
			P_3	20	95.90
			P_4	19	96.11
			P_5	18	96.31
T_3	247	732	P_1	28	96.17
			P_2	31	95.77
			P_3	29	96.04
			P_4	30	95.90
			P_5	29	96.03
T_4	264	798	P_1	28	96.49
			P_2	35	95.61
			P_3	31	96.12
			P_4	33	95.86
			P_5	30	96.24
T_5	310	935	P_1	35	96.26
			P_2	40	95.72
			P_3	37	96.04
			P_4	38	95.94
			P_5	36	96.15
T_6	384	1165	P_1	44	96.22
			P_2	50	95.71
			P_3	47	95.97
			P_4	48	95.88
			P_5	45	96.14

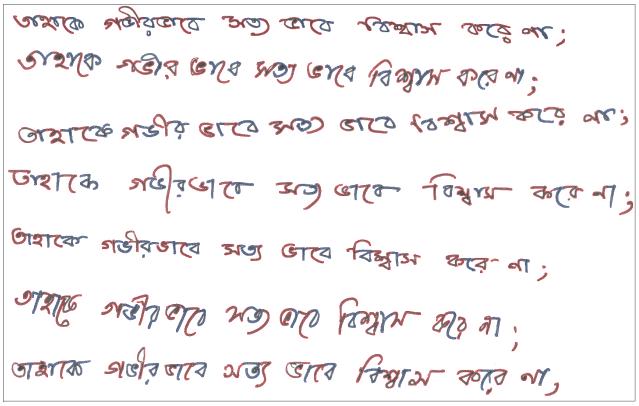

Figure 7. Handwritten character segmentation of a text written by different persons.

into small parts but those are not separate characters. Apart from these constraints, the proposed method works well for handwritten documents. To perform an evaluation with the state-of-the-art results in the literature, we compare the results of the present work with the results of some available algorithms reported in the literature. Our own dataset is used here for the comparison of results. A detailed comparison of different methods is shown in Table II. Regarding the accuracy rate, our method does perform better than other methods. As seen from experimental results, our method is applicable for reasonably skewed documents without deskewing them and does not incorporate any type of heuristics or classifiers for segmentation.

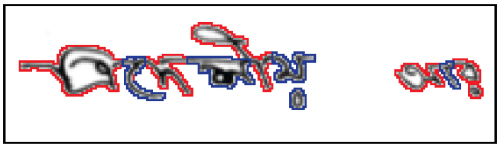

Figure 8. Failure in handwritten character segmentation. Left: Under-segmentation; Right: Over-segmentation.

Table II
COMPARISON OF OUR RESULTS WITH DIFFERENT ALGORITHMS.

Methods	Input pattern	Accuracy rate (%)
Recursive contour following [2]	Isolated word	91.10
Water reservoir model [3]	Text document	95.15
Graph-based model [4]	Isolated word	92.45
Fuzzy technique [5]	Isolated word	94.12
Fuzzy features and MLP classifier [6]	Isolated word	94.37
Proposed method	**Text document**	**96.04**

IV. CONCLUSION

We have proposed a novel handwritten character segmentation method by characterizing the vertices of outer isothetic polygonal covers corresponding to words in text documents. The proposed method is tested on different handwritten Bangla documents written by different individuals. We have obtained very promising results with an average accuracy rate of 96.04%. But this method is inefficient for a few particular cases stated earlier. In future, we shall extend our work as a significant preprocessing step towards the development of an integrated handwritten OCR system.

REFERENCES

[1] U. Pal and B. B. Chaudhuri, "Indian script character recognition: A survey," *Patt. Rec.*, vol. 37, pp. 1887–1899, 2004.

[2] A. Bishnu and B. B. Chaudhuri, "Segmentation of Bangla handwritten text into characters by recursive contour following," in *Proc. ICDAR*, 1999, pp. 402–405.

[3] U. Pal and S. Datta, "Segmentation of Bangla unconstrained handwritten text," in *Proc. ICDAR*, 2003, pp. 1128–1132.

[4] A. Roy, T. K. Bhowmik, S. K. Parui, and U. Roy, "A novel approach to skew detection and character segmentation for handwritten Bangla words," in *Proc. DICTA*, 2005, pp. 203–210.

[5] S. Basu, R. Sarkar, N. Das, M. Kundu, M. Nasipuri, and D. K. Basu, "A fuzzy technique for segmentation of handwritten Bangla word images," in *Proc. Intl. Conf. on Computing: Theory and Appl.*, 2007, pp. 427–433.

[6] R. Sarkar, N. Das, S. Basu, M. Kundu, M. Nasipuri, and D. K. Basu, "A two-stage approach for segmentation of handwritten Bangla word images," in *Proc. ICFHR*, 2008, pp. 403–408.

[7] N. Otsu, "A threshold selection method from gray-level histogram," *IEEE Trans. SMC*, vol. 9, no. 1, pp. 62–66, 1979.

[8] A. Biswas, P. Bhowmick, and B. B. Bhattacharya, "Construction of isothetic covers of a digital object: A combinatorial approach," *Journal of Visual Communication and Image Representation*, vol. 21, pp. 295–310, 2010.

[9] A. Sarkar, A. Biswas, P. Bhowmick, and B. B. Bhattacharya, "Word segmentation and baseline detection in handwritten documents using isothetic covers," in *Proc. ICFHR*, 2010, pp. 445–450.