A Modified Thinning Strategy to Handle Junction Point Distortion for Bangla Characters

Soumyadeep Ghosh and Soumen Bag
Department of Computer Science and Engineering
International Institute of Information Technology, Bhubaneswar, India
{somghosh1988, bagsoumen}@gmail.com

***Abstract*—Thinning which is an important preprocessing step for character recognition is often subject to several kinds of distortion. Junction point distortion is a major imperfection in thinned images especially for handwritten Indian scripts due to the presence of large number of complicated junctions in them. Such distortion does allow the optical character recognition (OCR) systems to exploit the properties of these junctions for character recognition. We present a novel methodology to reduce distortion at junction regions adjoining the *mātrā* for Bangla script. Our method uses geometric properties of the junctions to solve the problem. We have tested our approach on our own data set consisting of a variety of isolated handwritten character images by different writers and have got promising results.**

***Keywords*—Bangla; Distortion; Extrapolation; Geometric properties; Junction point; Thinning.**

I. Introduction

Thinning of character images is one of the most fundamental preprocessing operations for several document image processing applications. The thinned skeleton is required to preserve the structural property of the characters. Producing good quality thinned skeletons of handwritten character images still continue to be a challenge in this field [1] due to the variability in writing style, spilling of ink on poor quality paper and usage of pens which produce thick strokes. Due to these reasons thinning is subject to several distortions like spurious strokes, spurious loops, recession of stroke length, and shape distortion at junction points. The quality of the thinned skeleton affects the feature extraction and finally the recognition accuracy of the OCR system [2].

In the last three decades a large number of methods for thinning of character images have been reported in literature. However very few of them have attempted to reduce the distortions in thinned skeletons. Chouinard and Plamondon [3] uses line following on the contour to reduce distortions at curvature regions. They use angular deviations on the contour to locate junction regions. However their strategy do not use domain knowledge and would not give good results for Indian scripts. Sung-Bae and Dong-Hyeop [4] have reduced the occurrences of spurious loops and shape distortions of touching Hangul (Korean script) characters, although they have not taken junction point distortion into consideration. The quality of the character image sometimes induces distortions like spurious strokes and spurious loops. On the other hand shape distortion at junction points are a common phenomenon when written with ink pens on poor quality paper which produces thick strokes. They are also caused due to different writing styles of the same character. Junction point distortion is extremely difficult to do away with and no dedicated method for handling it (for any script) has been reported in the literature. Indian scripts have lot of complicated junctions and the thinned skeletons sometimes completely lose the structural properties of the junctions. Thus the opportunity to use structural characteristics of the junctions as a feature in character recognition has been a challenge in modern OCR systems.

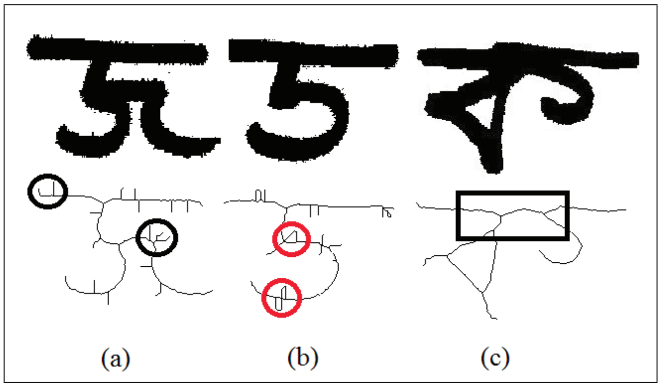


Fig. 1. Types of distortion in thinned skeletons. (a) Black circles indicate spurious strokes; (b) Red circles indicate spurious loops; (c) Region under rectangle suffered junction point distortion.

Figure 1 shows different kinds of distortion in handwritten Bangla characters. It can be seen that spurious strokes (Fig. 1(a)) and loops (Fig. 1(b)) occur in character images with noisy contour due to spilling of ink along the contour of the character on low quality paper. Figure 1(c)) shows a completely deformed junction in a thinned skeleton of the Bangla character (ক), due to irregular handwriting. Figure 2 compares thinning results of printed and handwritten character images. It can be seen that junction point distortion is evident in thinned skeletons of both printed and handwritten characters, although the distortion is more pronounced for handwritten samples. We observed that some specific characters of the Bangla character set suffered more junction point distortion than others, depending on the type of junctions present in them. Some of the are identified in Fig. 2.

The rest of this paper is organized as follows. Section II describes the nature of junction point distortion for Bangla script. Section III describes our proposed methodology for the remedy of junction point distortion. Experimental results and related discussion are reported in Section IV. The concluding notes are given in Section V.

978-1-4799-2608-4/14/$31.00 ©2014 IEEE

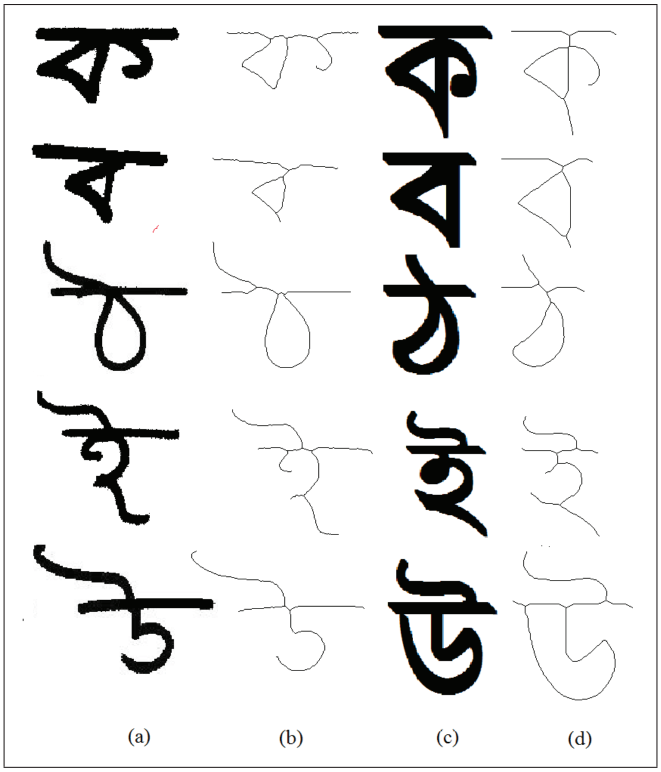


Fig. 2. Comparison of junction point distortion for some printed and handwritten characters of Bangla script. (a) Input character (handwritten); (b) Thinned result; (c) Input character (printed); (d) Thinned result.

II. Junction Point Distortion

Most characters in every Indian script contain one or more junctions. Junctions are intersection regions of more than two strokes. These junctions contain essential information, which can be exploited for OCR. But due to high distortion at junction points, OCR systems for Indian scripts have faced challenges to consider junction points as a feature for recognition. In this paper we have attempted to reduce junction point distortion for Bangla basic characters. Such thinned skeletons are good candidates for development of a more efficient feature set for Bangla OCR.

A. Properties of Bangla script

Bangla is the second most popular language in India and the fifth most popular language in the world [2]. The alphabet of the modern Bangla script consists of 11 vowels and 39 consonants. These characters are called basic characters. Most of the Bangla characters have a headerline called *mātrā*, which touch the upper part of the character. Depending on the character it covers at most the entire width of the character. Due to the presence of this headerline the junctions become more complicated. This headerline is easily identified due to the presence of some geometrical properties which has been utilized in our proposed approach.

B. Junction point distortion in Bangla characters

Figure 3 shows how different writing style of the same character result in different degrees of junction distortion in their corresponding thinned skeletons. The figure shows three samples of the Bangla character (ক). The 3^{rd} column of the figure shows the distorted junctions in magnified form. In the same figure, (1^{st} row 3^{rd} column), the stroke segments $\mathcal{S}_{pq}$, $\mathcal{S}_{qu}$, $\mathcal{S}_{uv}$ is the headerline (*mātrā*) of the input character. $\mathcal{S}_{xy}$ represent a stroke segment from point '*x*' to '*y*', which are points on the thinned skeleton. Ideally five strokes should have originated from a single point for this character. Also there should have been only one junction point in this region (enclosed in red rectangle in 1^{st} row 2^{nd} column) of the character. Instead there are three junction points in the region, namely $\mathcal{J}_q$, $\mathcal{J}_r$ and $\mathcal{J}_u$ where $\mathcal{J}_x$ denotes a junction point '*x*' on the thinned skeleton. But the junction lost its structural property on all the three samples of the thinned skeletons of the character (ক) as shown in Fig. 3.

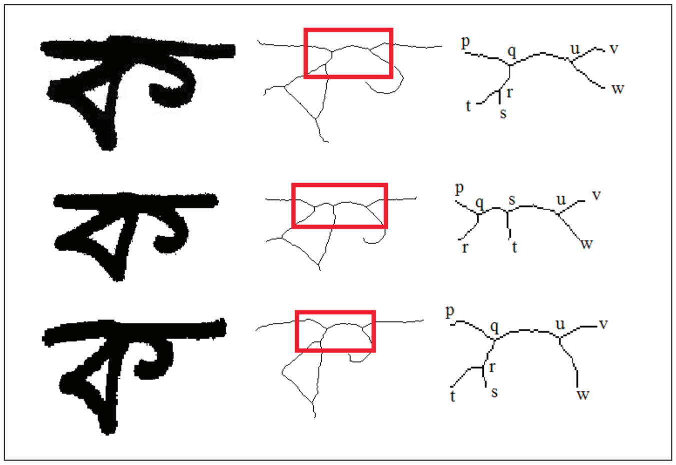


Fig. 3. Nature of distorted junction in different handwritten samples of the same character. 1^{st} column: Input characters; 2^{nd} column: Thinned results with distorted junctions indicated with red rectangles ; 3^{rd} column: Distorted junctions magnified.

III. Proposed Method

We propose a step by step modified thinning strategy which will reduce junction point distortion to a considerable extent. The method is summarized as follows. First we find out the average stroke width of the character image. We note that junction regions are much wider than the average stroke width. Now, we can use any iterative or parallel thinning algorithm to thin the input image partially, so the input character image, except the junction regions are thinned. Now the junction regions are extracted from the partially thinned image. We analyze the junctions geometrically and find out the junction from which maximum strokes emerge. We locate the *mātrā* in this junction region and reconstruct it. Next we locate other strokes and connect them to the *mātrā* using extrapolation. We explain the proposed method in detail in the following subsections.

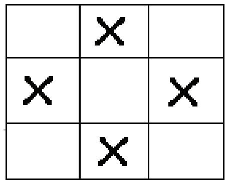
A. Average stroke width

We find out the average stroke width of the character in both the horizontal and vertical directions given by $\mathcal{W}_h$ and $\mathcal{W}_v$ respectively and take their mean as the average stroke width denoted by $\mathcal{D}_{avg}$. It is an essential property that the junction regions are much wider than the average stroke width of the character. This property helps us to partially thin the image so that the character is thinned except the junction regions.

B. Partial thinning

We use any iterative or parallel thinning method to thin the image. Iterative or parallel thinning methods are known as raster scan based thinning methods and share a common property that they remove pixels from the contour of the character, until a single pixel width skeleton is left. They apply the same set of masks repeatedly in a loop, each time doing a raster scan of the image and removing pixels. The loop terminates when no pixel is removed by an entire raster scan of the image. The input image is transformed into a single pixel width skeleton.

We run the loop of the iterative or parallel thinning method exactly $\mathcal{D}_{avg}$ number of times. We have used the thinning algorithm of Huang *et al.* [5] for our experimentation. After this process we are left with a partially thinned image $\mathcal{P}_t$. The junction regions are not entirely thinned since they were much wider than $\mathcal{D}_{avg}$. But the rest of the strokes are thinned completely. There may be more than one junction in the input character. Figure 5(b) shows partially thinned result of character image given in Fig. 5(a).



	×	
×		×
	×	

Fig. 4. Structuring element used for erosion operation.

C. Extracting the target junction

We apply the morphological operation of *opening* on the partially thinned result for extracting the target junction. *Opening* is defined as *dilation* of the *erosion* of an image. We apply *erosion* to eliminate single pixel regions using the structuring element given in Fig. 4. The junction regions remain and we apply *dilation* operation on it. *Dilation* is applied because *erosion* removes pixels from the contour of the junction regions as well. Figure. 5(b) is transformed to Fig. 5(c) after the *opening* operation. Thus we extract the junction regions from the partially thinned image without affecting the geometric property of these regions.

It is evident that there might be more than one junction region in the input character. So the partially thinned image may have more than one unthinned region as can be seen in Fig. 5(b). We extract that region which is expected suffer maximum deformation. We use component labeling [6] to extract each junction region. Figure 5(c) shows each junction region of the input character that was not completely thinned in Fig. 5(b). We perform geometrical analysis of the regions (explained in the next subsection) to identify that junction from which maximum strokes emerge. That junction is our target junction. The proposed approach will work to reduce distortion in this junction region. Our approach do not take into consideration multiple junction point distortion in the character.

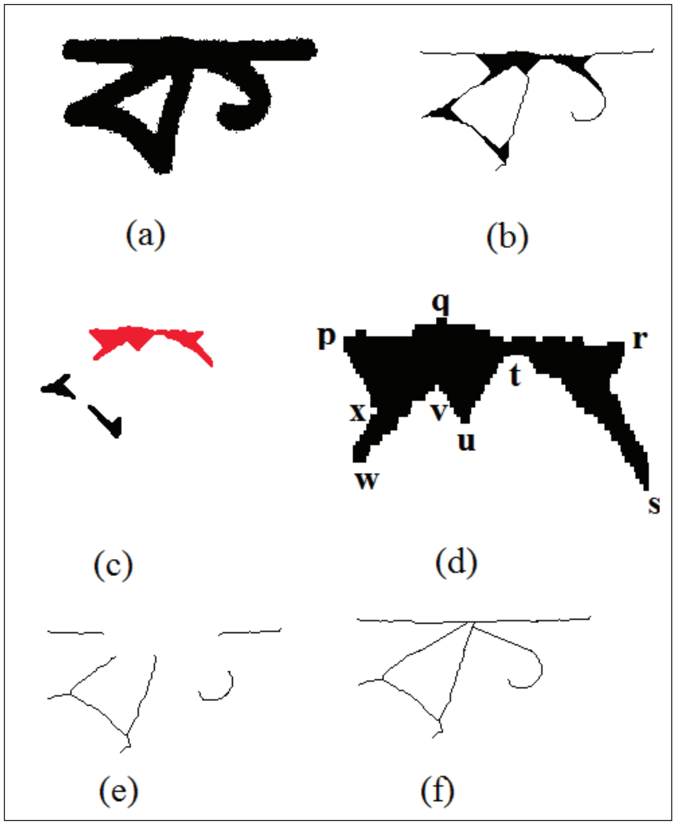


Fig. 5. Stepwise illustration of the proposed thinning method. (a) Input character; (b) Result of partial thinning; (c) Unthinned regions extracted using *opening* (target region shown in red); (d) Target region magnified; (e) Target region removed from partially thinned image and thinning reapplied; (f) Final thinning result after extrapolation.

D. Analysis of the junctions

Figure 5(d) shows a detailed representation of one of the junction regions given in Fig. 5(c; highlighted in red). We analyze each junction (Fig. 5(c)) one by one. First we find out the contour of the unthinned junction region. Then we traverse along the contour in clockwise direction and look for possible sharp convex turns. Each sharp convex turn is an evidence of a stroke, which emerge out of this junction. In Fig. 5(d) we can see that there are five sharp convex turns namely, $\mathcal{K}_p$, $\mathcal{K}_r$, $\mathcal{K}_s$, $\mathcal{K}_u$, and $\mathcal{K}_w$. $\mathcal{K}_x$ represents a turn on the contour where 'x' is a point on the contour of the region. There are also some concave turns on the same region namely, $\mathcal{K}_t$, $\mathcal{K}_v$ and $\mathcal{K}_x$. However, strokes do not emerge from concave turns. The method of analysis for each region is formally explained as follows.

1 Let there be n points on the contour of the junction region $\mathcal{R}_i$ and V=$\{p_1, p_2,...., p_3\}$ be an ordered set of the points on the contour, assuming the traversal of the points in clockwise direction. let $\mathcal{L}_{x,y}$ be the straight line from the point x to the point y and let the slope of this line be denoted by θ_x.

2 We start from a point p_i on the contour and find the slope θ_i of the straight line $\mathcal{L}_{i,i+2}$ which is the straight line from p_i to p_{i+2}. We ignore p_{i+1} because this approximation gives us better results for detecting the sharp change in slopes of the consecutive straight

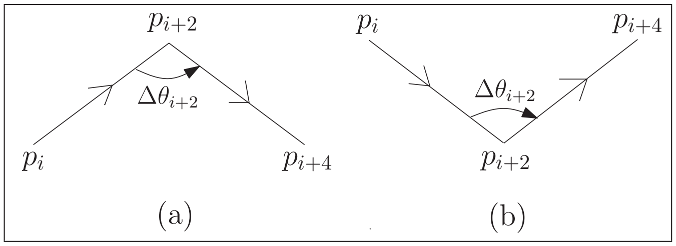


Fig. 6. Convex and concave turns at the point p_{i+2}, arrows showing the direction of traversal. (a) Convex turn at p_{i+2}; (b) Concave turn at p_{i+2}.

lines. Next we take the slope of the straight line $\mathcal{L}_{i+2,i+4}$ given by θ_{i+2}.

3 Then we find the change in slope of these two consecutive straight lines $\mathcal{L}_{i,i+2}$ and $\mathcal{L}_{i+2,i+4}$ by taking the absolute difference of θ_i and θ_{i+2}, and we denote it by $\Delta\theta_{i+2}$ as shown in Fig. 6. If this change in slope $\Delta\theta_{i+2}$ is greater than a threshold we can detect it as a sharp turn at the point p_{i+2}.
We repeat Step 2 and 3 for all points on the contour of the region.

4 A stroke will emerge from only those sharp turns on the contour of the junction which are convex in nature. So we find out the convexity of the point p_{i+2} by the method given below [7].

$$\Delta(p_{i-1}, p_i, p_{i+1}) = \begin{vmatrix} 1 & 1 & 1 \\ x_{i-1} & x_i & x_{i+1} \\ y_{i-1} & y_i & y_{i+1} \end{vmatrix} \tag{1}$$

The convexity of a point p_i can be determined by calculating the value of twice the signed area of a triangle formed by the point $p_i(x_i, y_i)$ and its two adjacent points $p_{i-1}(x_{i-1}, y_{i-1})$ and $p_{i+1}(x_{i+1}, y_{i+1})$. Equation 1 gives the formula for calculating it. If the value of $\Delta(p_{i+1}, p_{i+2}, p_{i+3})$ is positive then the point p_{i+2} has convex property. If the value is equal to 0 then the point p_{i+2} has the same property as the previous point p_{i+1}. Thus, if p_{i+2} is convex and if $\Delta\theta_{i+2}$ is greater than 65° (which is taken as the threshold) then the point p_{i+2} is a position from which a stroke originates in this junction area. The process is illustrated Fig. 6. We repeat this process for each junction area. We calculate the number of strokes in each junction area and select that junction area from which maximum number of strokes originate. This junction area is removed from the partially thinned image and we reapply thinning on this image to get Fig. 5(e). We then perform extrapolation (explained in the next section) to reconstruct this junction area.

E. Extrapolation

Figure 7(a) shows the partially thinned region after the target junction has been removed for the Bangla character (ব). Now thinning is reapplied for the remaining junction regions which were not completely thinned. In Fig. 5(e), we see that the target junction area needs to be reconstructed by extrapolation of the disconnected strokes to produce the final thinned skeleton shown in Fig. 5(f). We define the following terminology for explaining our extrapolation method.

1) Local orientation: We find out the end points $\mathcal{E}_{ai}$ and $\mathcal{E}_{bi}$ of each stroke ϕ_i as shown in Fig. 7(b). Now we find the average slope $\theta_{avg(a)}$ and $\theta_{avg(b)}$of the preceding *k* points of $\mathcal{E}_{ai}$ and $\mathcal{E}_{bi}$ respectively. The value *k* is calculated as 15% of the total number of points (say *n*) in ϕ_i. Thus $\theta_{avg(a)}$ and $\theta_{avg(b)}$ represent the local orientations of the stroke ϕ_i.

To locate the *mātrā* which are almost horizontal strokes (we have not considered skewed character images) we can check for strokes having local orientations almost close to 0° (may vary by ±10°) at both the endpoints. However some Bangla characters do not contain a *mātrā*. We do not apply our method on those characters. If the matra cannot be located in the character(using our method) we skip this extrapolation step and apply normal thinning on those characters. The extrapolation procedure is illustrated as follows.

1 After identifying the *mātrā*, we need to find out the endpoints $\mathcal{E}_{a1}$ and $\mathcal{E}_{a2}$ which need to be extrapolated, as can be seen in Fig. 7(b). To locate these points we measure the distance between each pair of endpoints of two corresponding *mātrā* strokes and if the distance is less than the threshold $\mathcal{D}_f$ we consider them as points which need to extrapolated. The threshold $\mathcal{D}_f$ is taken as the distance between the farthest points on the target junction region identified in the previous section.

2 After $\mathcal{E}_{a1}$, $\mathcal{E}_{a2}$ (of the *mātrā* stroke) is identified we apply extrapolation, so that they meet. We use the local orientations for extrapolation of the *mātrā*.

3 After the *mātrā* is reconstructed the rest of the endpoints $\mathcal{E}_{a3}$ and $\mathcal{E}_{b3}$ need to joined with the *mātrā* (Fig. 7(b)). But before we do this, we check the distance from $\mathcal{E}_{a3}$ and $\mathcal{E}_{b3}$ to the *mātrā*. If it is greater than $\mathcal{D}_f$, we skip our method and apply normal thinning on those characters. This is necessary since our approach do not take into consideration those junctions which are not adjacent to the *mātrā*. If it is not so (i.e., the target junction is adjacent to the *mātrā*), we take the average of local orientation at $\mathcal{E}_{a3}$ and the orientation of the shortest straight line between $\mathcal{E}_{a3}$ and the *mātrā* (which has already been reconstructed). Similar process is carried out to extrapolate $\mathcal{E}_{b3}$. This averaging helps to preserve the structural property of the region.

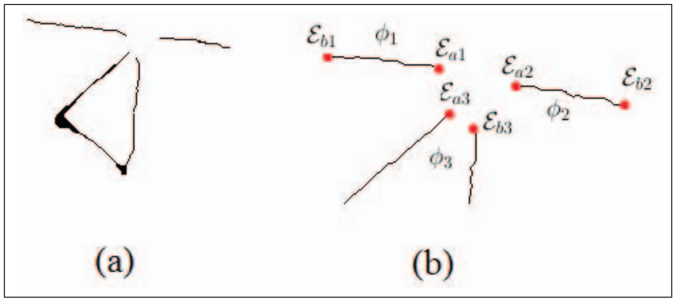


Fig. 7. Extrapolation process. (a) Partially thinned result after removing target region; (b) Target region magnified, ϕ_1, ϕ_2 and ϕ_3 are the strokes and their end points are highlighted in red.

IV. Experimental Results and Discussion

This section presents the experimental results and related discussion of our proposed method.

A. Dataset

We have taken about 509 handwritten Bangla characters written by 3 different kind of pens by 4 different writers. We have scanned them at 200 dpi with HP Officejet 5610. We have considered only isolated and non-skewed characters for our experimental purpose. All the programs were written in MATLAB (R2010a) on the Windows-7 platform.

B. Test Results

We compared our results with the thinning algorithms of Huang *et al.* [5] and Datta-Parui [8] which are well known raster scan based thinning methods. Some of the results that we have obtained have been reported in Fig. 8. Our results exhibited very less junction point distortion and produced structurally similar junctions for the same characters written by different individuals with different pens. Our method gave good results even on some extremely irregular handwritten character images. However it should be noted that the proposed method works only on junctions adjoining the *mātrā*. On some of the irregular handwritten characters, as can be seen in Fig. 8, thinned results produced by our method also exhibited some amount of junction point distortion, although it was reduced to a considerable extent. On some occasions however we also succeeded in getting all the strokes in the target junction region to meet at one point. Creating the ideal junction point for all handwritten character images will be difficult for any thinning algorithm, due to high variations in handwriting as shown in Fig. 3. However our proposed method produced thinned skeletons which is a considerable improvement over all previous thinning methods. Our method produced skeletons which are very good candidates where the nature of junctions can be utilized as a feature for optical character recognition. In future we will utilize this work to propose a feature extraction method which incorporates the nature of the junction points as a major characteristic for Bangla character recognition.

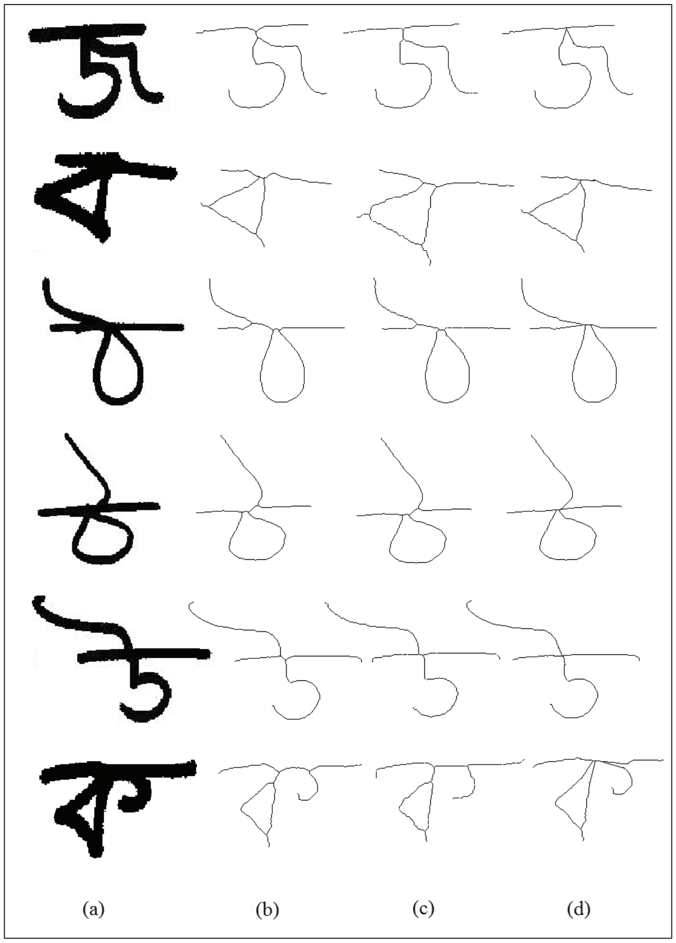


Fig. 8. Experimental Results. (a) Input Character Images; (b) Thinning results using Huang *et al.* [5]; (c) Thinning results using Datta-Parui [8]; (d) Thinning results of our proposed method.

V. Conclusion

One of the major distortions that has not been handled by thinning algorithms is the junction point distortion. Removing junction point distortion is extremely difficult especially for handwritten characters. Indian scripts exhibit very high junction point distortion in their corresponding thinned skeletons. Junction regions of the same character in different handwritten samples show different structural properties, due to the inherent variability of writing style of different individuals. In order to utilize the characteristic of junctions as a feature for optical character recognition it is desirable that the same character written by different individuals with different pens on different quality paper show similar junction point characteristics. Our method presents an elegant and effective approach for tackling junction point distortion for basic Bangla characters. It will pave the way for exploiting the structure of the junction points to use it as an effective feature for Bangla character recognition. In future we will utilize this work to propose a feature extraction method which incorporates the nature of the junction points as a major characteristic for Bangla character recognition.

References

[1] S. Bag and G. Harit, "Skeletonizing character images using a modified medial axis-based strategy," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, no. 7, pp. 1035-1054, 2011.

[2] S. Bag and G. Harit, "A survey on optical character recognition for Bangla and Devanagari scripts," *Sadhana*, vol. 38, pp. 133–168, 2013.

[3] C. Chouinard and R. Plamondon, "Thinning and segmenting handwritten characters by line following," *Machine Vision and Applications*, vol. 5, pp. 185–197, 1992.

[4] C. Sung-Bae and H. Dong-Hyeop, "Extracting intuitive strokes in complex structured patterns with domain knowledge," *Engineering Applications of Artificial Intelligence*, vol. 16, pp. 65–73, 2003.

[5] L. Huang, G. Wan, and C. Liu, "An improved parallel thinning algorithm," in *Proceedings of the International Conference on Document Analysis and Recognition,*, pp. 780–783, 2003.

[6] A. Rosenfeld and A. C Kak, ' *Digital picture processing*, 2^{nd} Edition, vol 1 and 2, Academic press, New York, 1982.

[7] S. Bag, P. Bhowmick, and G. Harit, "Detection of structural concavities in character images—A writer-independent approach," in *Proceedings of the Indo Japan Conference on Perception and Machine Intelligence*, pp. 260–268, 2012.

[8] A. Datta and S. K. Parui, "A robust parallel thinning algorithm for binary images," *Pattern Recognition*, vol. 27, pp. 1181–1192, 1994.