Topological Features for Recognizing Printed and Handwritten Bangla Characters

Soumen Bag, Partha Bhowmick

Department of CSE
IIT Kharagpur
India

Gaurav Harit

Department of CSE
IIT Rajasthan
India

Contents

- ✓ Contribution
- ✓ Properties of *Bangla* script
- ✓ Proposed Character Recognition Method
- ✓ Experimental Results
- ✓ Conclusion

Contribution cont.

- Recognition of Bangla characters by developing topological features which have the capability to capture the distinguishing aspects of Bangla characters - both basic and compound.

- Topological features are described by different skeletal convexities of strokes. Such skeletal convexities act as invariant features for character recognition.

Contribution

- Experiment is done on a benchmark datasets of printed and handwritten Bangla basic and compound character images.

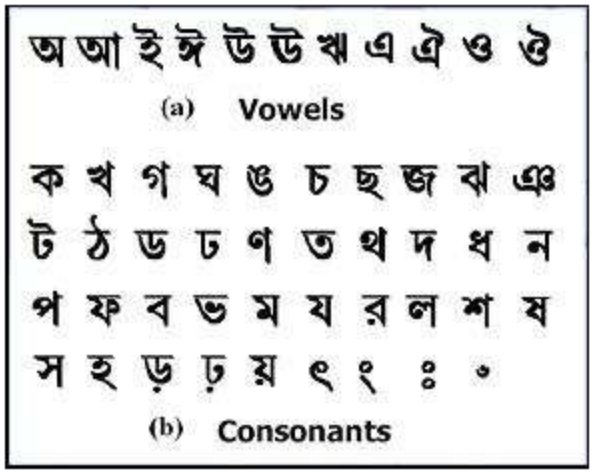
- The experimental results demonstrate the efficacy of our proposed method comparing with other methods.

Properties of Bangla script cont.

- Bangla (Bengali) is the second most popular language in India and fifth most popular language in world.
- The script name of this language is also called Bangla.
- This script has 11 vowels and 39 consonants. These characters are named as Basic characters.
- This script has near about 250 compound/conjunct characters. Conjunct characters are formed by combining 2 or 3 basic characters together.

Properties of Bangla script

- Most of the characters have a header line named Matra.

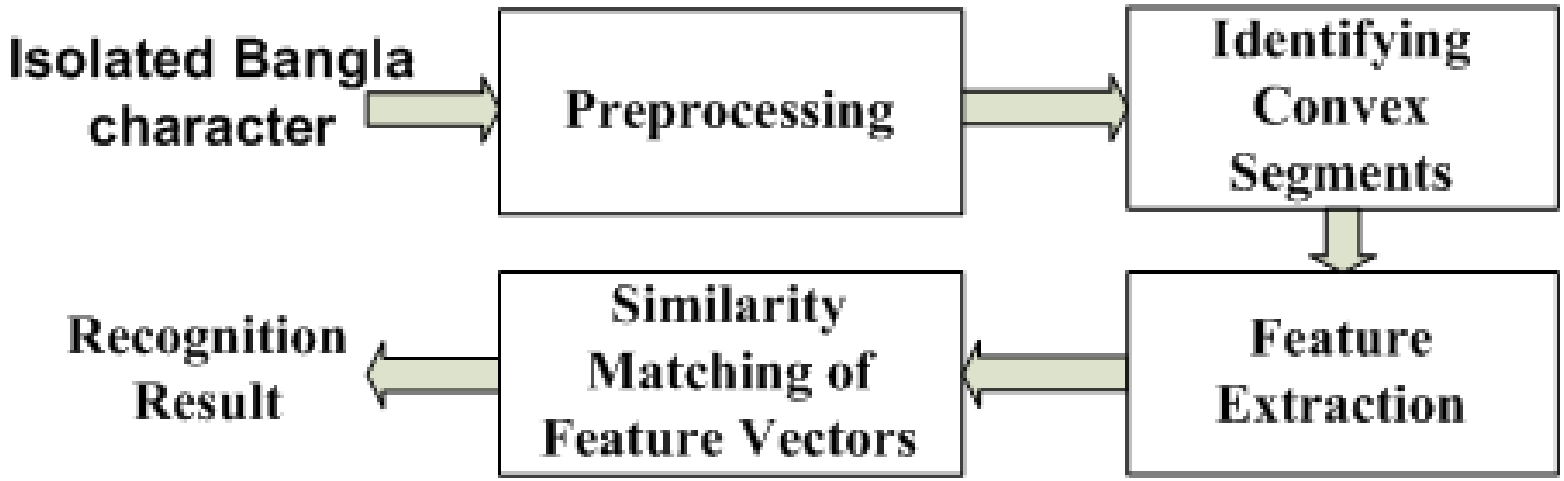


Basic characters

Conjunct characters

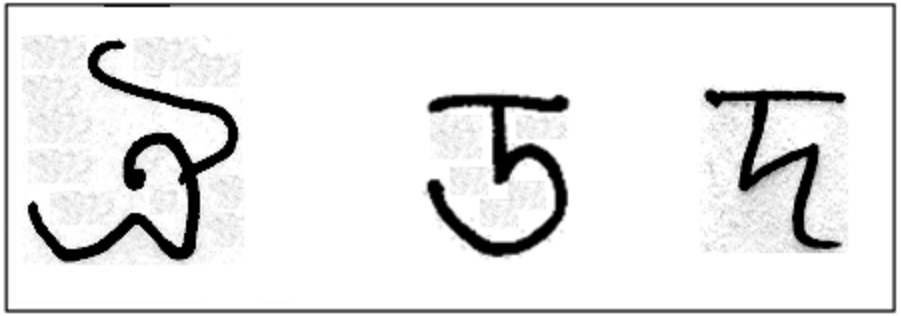
Proposed Method cont.

The algorithm is divided into Four phases:

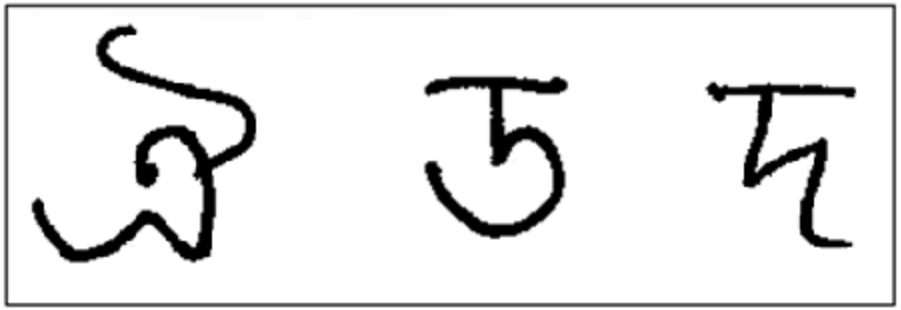


Preprocessing cont.

1. Binarize the given scanned character image.



Input images



Binarized images

Preprocessing cont.

2. Character images are converted to single pixel thick images by a medial-axis based thinning strategy[1].

Binarized images

Skeleton images

[1] S. Bag and G. Harit, ``A medial axis based thinning strategy and structural feature extraction of character images,” in *Proc. ICIP*, 2010, pp. 2173–2176.

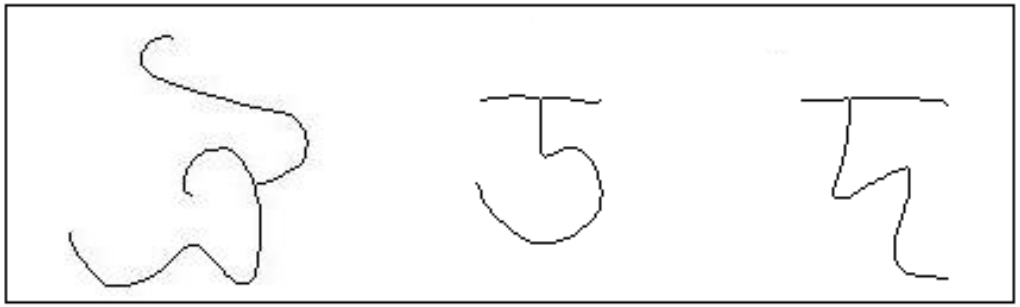
Preprocessing cont.

3. For noisy images, the proposed thinning results in undesired small concave and convex regions.

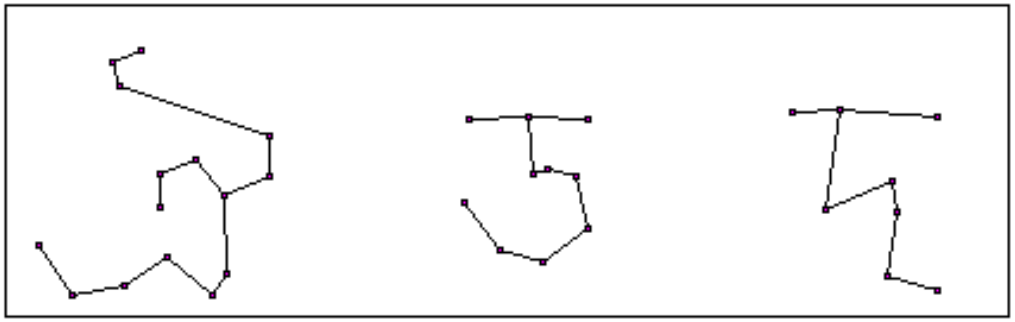
✓ To solve this problem, we apply a straight line approximation method[1] on thinned images.

[1] P. Bhowmick and B. B. Bhattacharya, ``Fast polygonal approximation of digital curves using relaxed straightness properties,'' *IEEE Trans. PAMI*, vol. 29, no. 9, pp. 1590–1602, 2007.

Preprocessing



Skeleton images



Straight line approximation results

✓ The approximation results often contain deviation of thinned images at the junction points. To solve this problem, we perform junction point refinement.

Identifying Convex Segments cont.

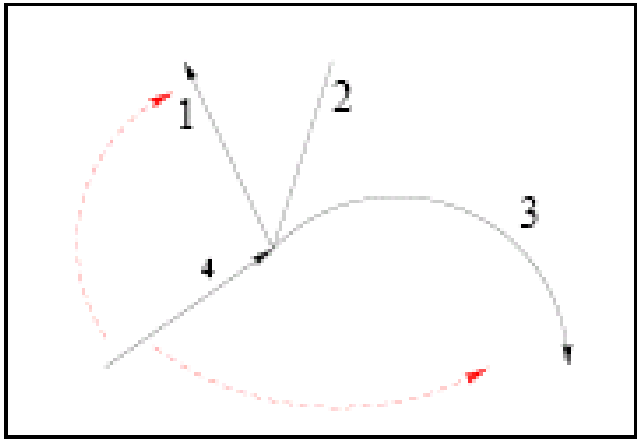
- This phase has Three parts:

  - Path traversal
  - Detection of concavity and convexity
  - Segmenting character strokes into convex regions

Path Traversal cont.

- Traversal start from any end point and instantiate a new path with an unique ID. Each node is associated with the IDs of the paths passing through that node.

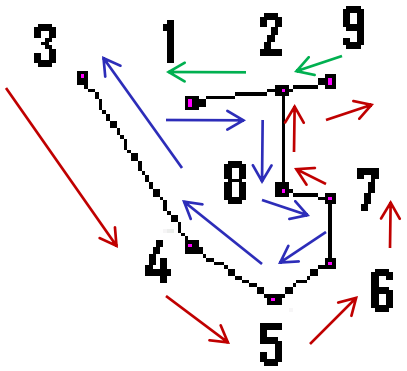
- When a junction is encountered, we choose the first branch towards the counter clock-wise side.



Path Traversal cont.

- We proceed past the junction point and continue traversal on the identified branch. Other junction points encountered on the path are traversed using the same policy.
- The path terminates when it reaches another end point of the skeleton or if it reaches back to the starting point (in case of circular traversal).
- A new path would now be traversed from some other end point of the skeleton.

Path Traversal



Path ID	Visited points
P_1	1-2-8-7-6-5-4-3
P_2	3-4-5-6-7-8-2-9
P_3	9-2-1

Detection of concavity/convexity cont.

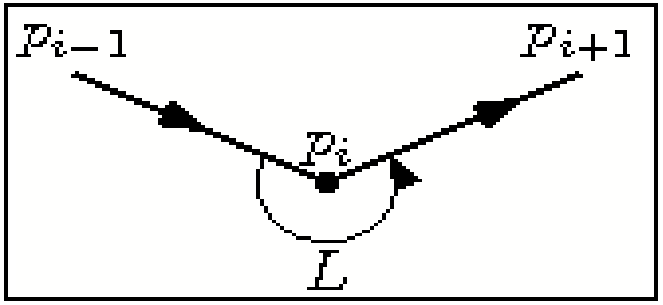
- To detect the concavity/convexity of a point p_i , we need to consider its two adjacent points p_{i-1} and p_{i+1}.

- Consider $p_{i-1}\,(x_{i-1},\ y_{i-1})$, $p_i\,(x_i,\ y_i)$, and $p_{i+1}\,(x_{i+1},\ y_{i+1})$ as the three vertices of a triangle. Then twice the signed area of this triangle is given by

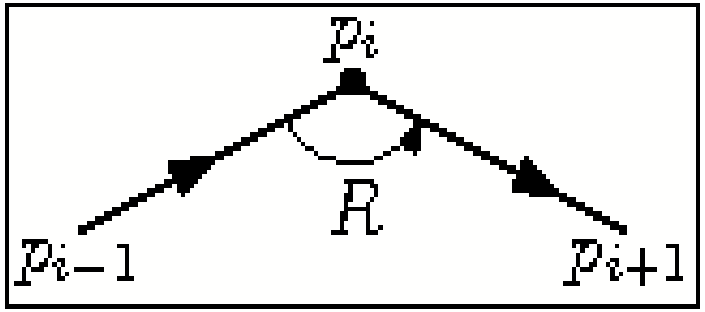
$$\Delta(p_{i-1}, p_i, p_{i+1}) = \begin{vmatrix} 1 & 1 & 1 \\ x_{i-1} & x_i & x_{i+1} \\ y_{i-1} & y_i & y_{i+1} \end{vmatrix}$$

Detection of concavity/convexity cont.

- If $\Delta(\cdot) < 0$, then the point p_i has a concave property and it marks as L.
- If $\Delta(\cdot) > 0$, then p_i has a convex property and it marks as R.



Concave



Convex

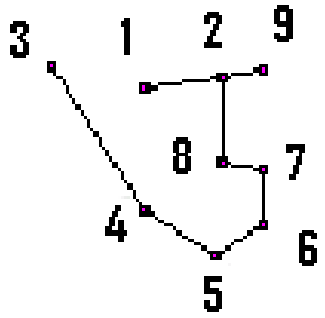
Detection of concavity/convexity cont.

- If $\Delta(p_{i-1}, p_i, p_{i+1}) = 0$, then the point p_i has the same property of its previous point p_{i-1}.

- An end point is assigned the same label as that of the adjacent point.

Segmenting Character Strokes cont.

- After detecting the concavity/convexity of all the points, we get a list *L = {R, R, L, L, R, L, …}*, where *L / R* indicates the concavity/convexity of a point.



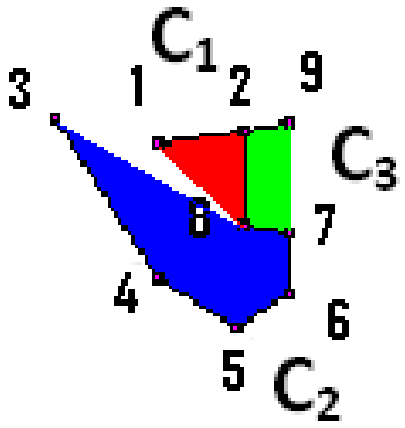
P_1:

1	2	8	7	6	5	4	3
	R	L	R	R	R	R	R

P_2:

3	4	5	6	7	8	2	9
	L	L	L	L	R	R	R

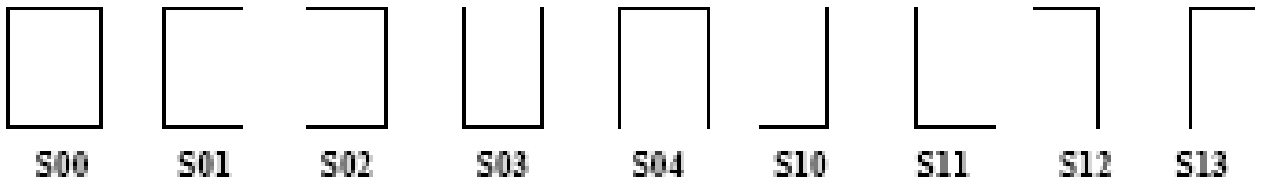
Segmenting Character Strokes



Convex Segment	Approximation points
C_1	1-2-8
C_2	8-7-6-5-4-3
C_3	7-8-2-9

Feature Extraction cont.

- Each concave segment is approximated by a shape prototype selected from a fixed set of shape primitives.



- S00 : This corresponds to a closed region. This is detected during graph traversal.
- S01 : $x_d > y_d$. The *x* coordinate of end point is greater than *x* coordinate of other points.

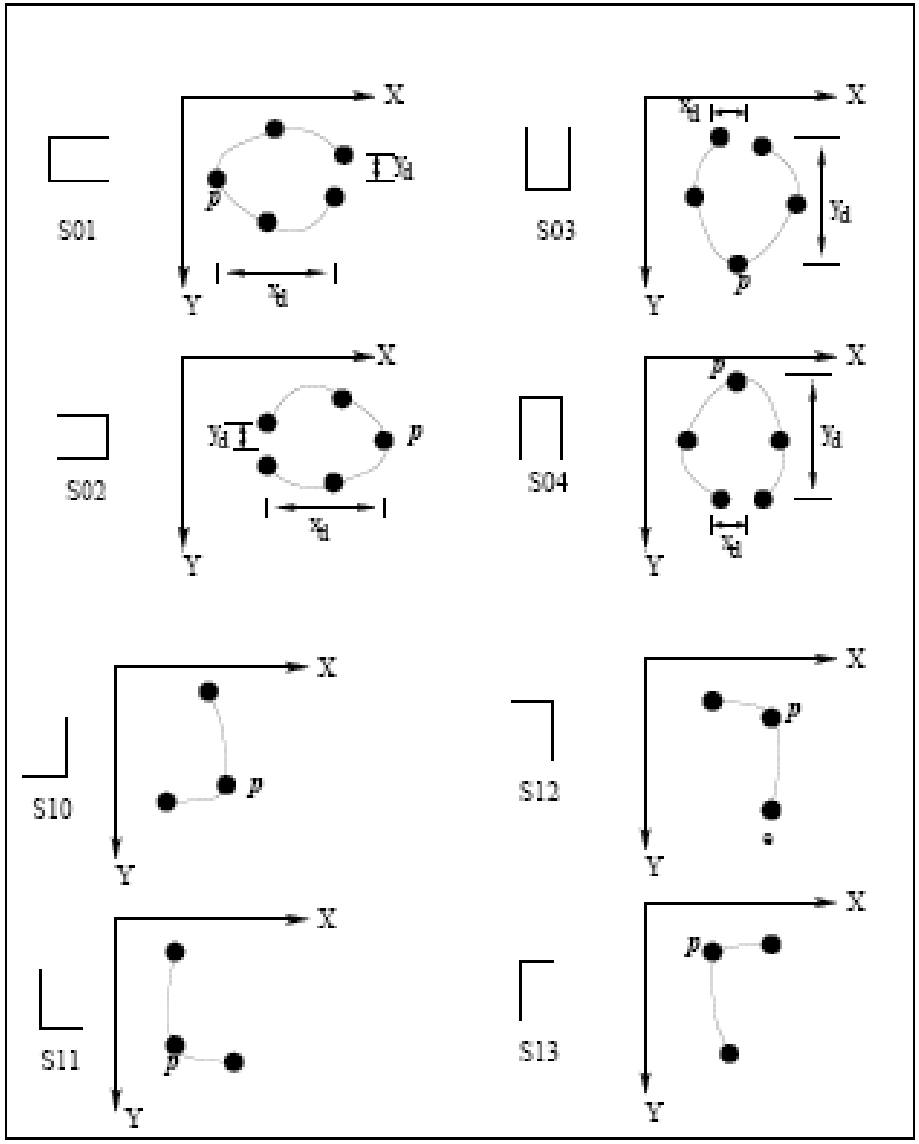
Feature Extraction cont.

- S03 : $y_d > x_d$. The y coordinate of end point is less than y coordinate of other points.
- S10 : $x_d = 0$ and $y_d = 0$. The orientation of shapes is worked out by examining the relative orientation of points relative to the line joining the end points.
- The shape descriptor for a shape segment comprises:

  (1) The ID of the shape primitive

  (2) The pair (N_i, D_i) for each of its adjacent shape primitives.

Feature Extraction



$X_d = 0$ if $x_{e1} \le x \le x_{e2}$ or $x_{e2} \le x \le x_{e1}$

$= |x - x_e|$ otherwise

Similarity of Feature Vectors cont.

- To identify a given character we compute its feature similarity score with each of the templates of Bangla characters.
- The given character is labeled depending on which template receives the highest match score.

$$\text{match score} = \sum_{\forall i \in G} w_i \; m_i$$

G: Set of shape primitives; w_i : Assigned weight of a shape primitive *i*

m_i : the degree of match for the primitive shape *i*

Similarity of Feature Vectors

$$m_i = \frac{1}{|A_i|} \sum_{j \in A_i} \texttt{match}(\mathcal{N}_j, \mathcal{D}_j)$$

$|A_i|$: Total number of adjacent shape primitives to the *i* th primitive

$\texttt{match}(\mathcal{N}_j, \mathcal{D}_j)$: Returns *1* if the adjacent shape primitives match in terms of their shape IDs and relative direction, else returns *0*.

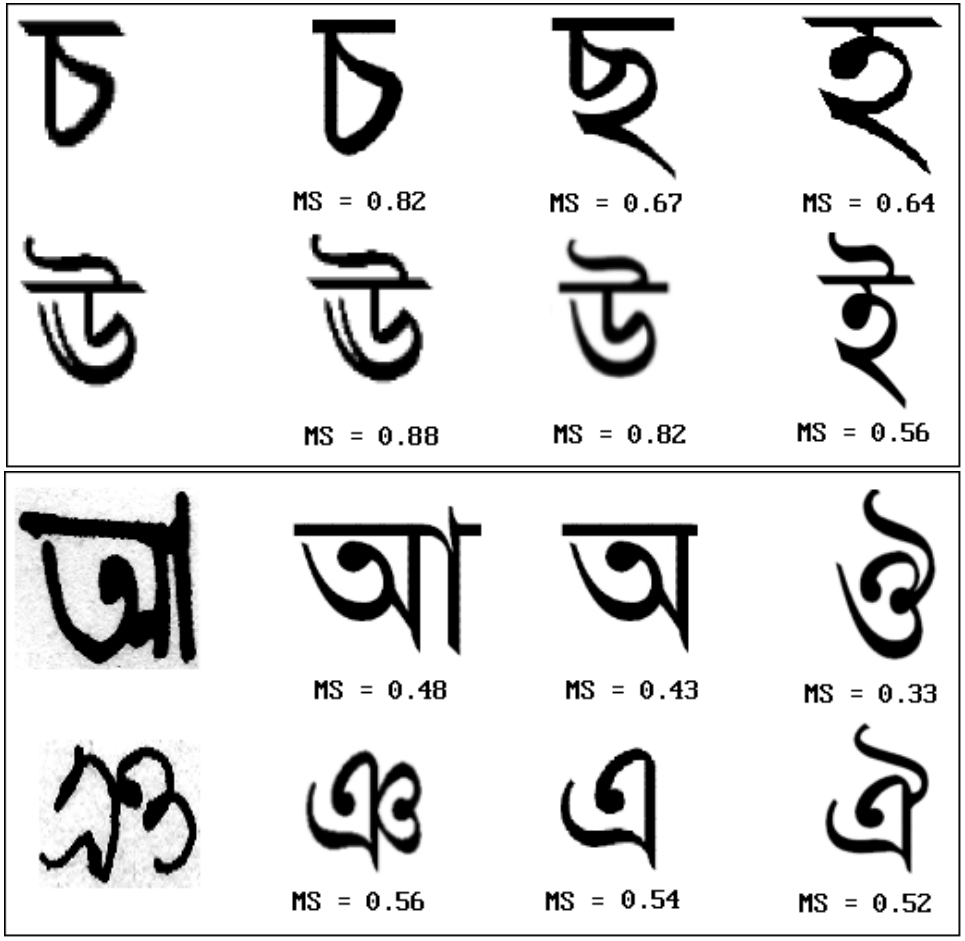
Experimental Results cont.

Information of different test datasets used for experiment

Dataset type	Dataset collected at	# distinct characters	Sample size
Printed basic	IIT Kharagpur	50	20
Handwritten basic	ISI Kolkata[1]	50	20
Printed compound	IIT Kharagpur	165	20
Handwritten compound	IIT Kharagpur	165	20

[1] www.isical.ac.in/~ujjwal/download/database.html

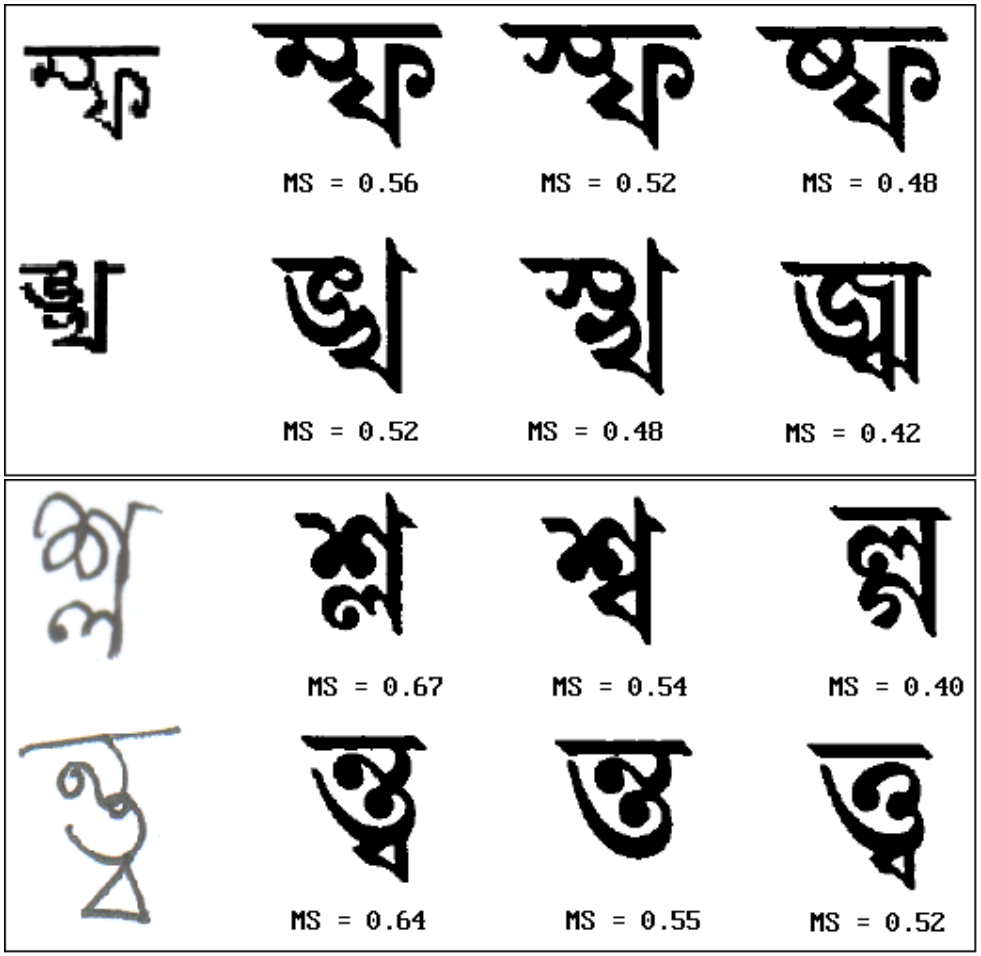
Top Three Matches as per their Matching Score (MS) cont.



Printed basic

Handwritten basic

Top Three Matches as per their Matching Score (MS)



Printed compound

Handwritten compound

Experimental Results cont.

Bangla basic character recognition rates based on different choices

Character type	# top matches considered	Recognition rate (%)	
		Printed	Handwritten
Basic	1	98.6	96.2
	2	99.1	97.1
	3	99.4	98.3
	4	99.7	98.9
	5	99.8	99.1

Experimental Results

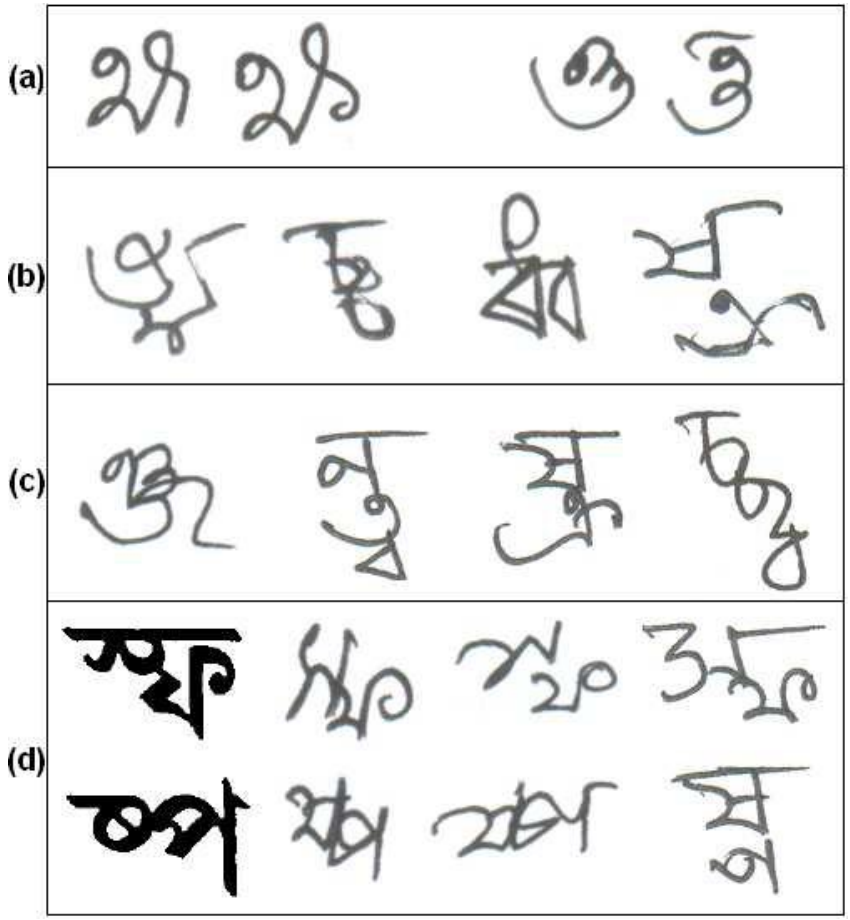
Bangla compound character recognition rates based on different choices

Character type	# top matches considered	Recognition rate (%)	
		Printed	Handwritten
Compound	1	88.4	86.1
	2	89.1	87.2
	3	89.7	87.8
	4	90.2	88.2
	5	90.3	88.3

Comparison among different Bangla OCR Methods

Methods	Input pattern	Feature set	Recognition rate (%)
Chaudhury's *Pattern Recognition, 31(5), 531-549, 1998*	Printed basic	Structural and template	96.4
Bhattacharya's *Proc. ICVGIP, 817-828, 2006*	Handwritten basic	Local chain code histogram	91.8
Sural's *Pattern Recognition Letters, 20, 771-782, 1999*	Printed compound	Fuzzy-based	83.5
Pal's *Proc. Int. Conf. Info. Tech., 208-213, 2007*	Handwritten compound	Gradient	85.2
Proposed method	**Printed and handwritten basic and compound**	**Topological**	**98.6 (printed basic) 96.2 (handwritten basic) 88.4 (printed compound) 86.1(handwritten compound)**

Failure Cases



Similar-shaped characters

Very poor handwriting

Complex structure of characters

Deviation of shape of handwritten characters from the model

Conclusion cont.

- In this paper, we have proposed a novel topological feature extraction method for Bangla OCR system.
- We have detected convex-shaped segments formed by the character strokes. The topological feature set captures the spatial layout of convex segments.
- The proposed method has been tested on printed and handwritten Bangla characters. We have obtained promising results comparing with other methods.

Conclusion

- From experimental results, it is shown that structural features, when formulated properly, are potentially enough to handle small variations in characters.

- In future, we shall extend our work to improve the recognition rate of and to make it an integral component of a Bangla OCR system.

Thank you!