ISSN:2249-5789

A Cluster Based Hierarchical approach for the recognition of Bengali Handwritten Character

Satarupa Bagchi Biswas[1] , Smritikona Barai[2], Sandipan Dutta[3]

1 .Asst. Professor, Dept. of IT, Heritage Institute of Technology, Kolkata.

2 .Asst. Professor, Dept. of IT, Heritage Institute of Technology, Kolkata.

3. .Asst. Professor, Dept. of IT, Heritage Institute of Technology, Kolkata..

Abstract

This work proposes a cluster based hierarchical technique for recognition of handwritten Bengali character. Due to the consecutive appearance of Bengali text segmentation of Bengali character is not very easy. For successful recognition of handwritten Bengali character proper segmentation is an important criterion. In this respect the present approach trying to provide a complete solution for Bengali handwritten character recognition. Here mainly we will concentrate on single Bengali character recognition.

Keywords: Bengali Handwritten character recognition, Clustering, OCR, Segmentation.

1. Introduction

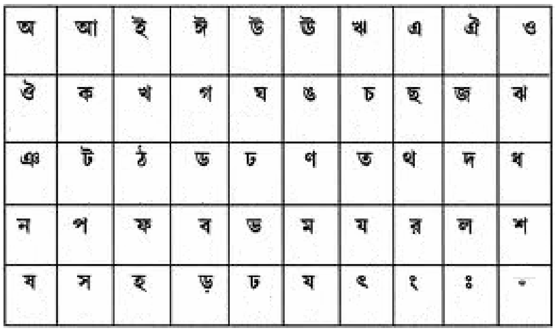
Bengali is the second most used language in India. Moreover certain languages like Manipuri, Ahamia have the similar script like Bengali and also Bengali is the official language of Bangladesh. Successful recognition helps to office automation and saves huge amount of time and effort. Though there are lots of commercially available system is there but yet they can be further extended to handwritten text. Thus handwritten character recognition research for Bengali script has a lot of significance.

Research works on optical character recognition (OCR) for printed Indian scripts including Bengali [1] are found in the literature. A survey of Indian script character recognition research is also available in [2]. However there are not very significant research works done so far. Unfortunately the technology used for printed character cannot be extended for handwritten character recognition
Due to the numerous style of writing and complex nature of Bengali character recognition of Bengali character is really a challenging one. No generalized rules can be formed to recognize character. The number of characters in basic *Bangla* alphabet is 50 which is much larger than that of Roman alphabet. Many algorithms/schemes for handwritten character recognition [3,4] exist and each of these has its own merits and demerits.The most important aspect of a handwriting recognition scheme is the selection of an appropriate feature set which is independent with respect to shape variations caused by various writing styles. A large number of feature extraction methods are available in the literature [5].

Several approach like stroke based, chain code based approach has already developed for successful recognition of handwritten Bengali character. But as handwriting varies from person to person so in our work we proposes a scheme which is fully dependent on rules of Bengali character writing.

2. Basics of Bengali Character Set

The Bengali script evolved from the Siddham, which belongs to the Brahmic family of scripts, along with the Devanagari and other written systems of the Indian subcontinent. Among 50 characters 11 are considered as vowel (Sarabarna) and rest are considered as consonant (Byanjanbarna).The script starts from left to right and there is no concept of capital or small letters like English character set. The Bengali character set is depicted below:



অ	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও
ঔ	ক	খ	গ	ঘ	ঙ	চ	ছ	জ	ঝ
ঞ	ট	ঠ	ড	ঢ	ণ	ত	থ	দ	ধ
ন	প	ফ	ব	ভ	ম	য	র	ল	শ
ষ	স	হ	ড়	ঢ়	য়	ৎ	ং	ঃ	ঁ

Figure 1: Bengali Character Set

2.1 Data Collection:

Writing style varies from person to person. In most of the previous work the data was collected in laboratory. Here we circulated a specially designed form (figure 2) among 25 persons to collect the samples of different writing styles. In the above specified form, 50 separate boxes are provided to enter one character at a time. Each box is divided into three rows; the topmost is dedicated for the curved line above headline (called ‘Matra’), the middle row is for writing the body of the character, and the bottom row is reserved for the dot (called ‘Bindu’) that some characters contain. Some examples are given below:

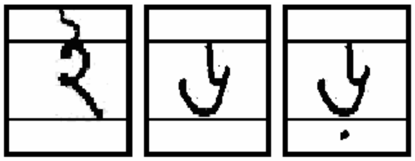


Figure 2: Data Collection Form

These sample forms are scanned and thus the sample data set is prepared. Now, from the whole character set, five characters (one vowel ঊand four consonants ঙ, ঞ, ণ, ঁ) are not considered for recognition, as it has been observed that, either they are not used independently, or their frequency of use is negligible in Bengali language.

3. Recognition Methodology:

To recognize handwritten Bengali character, the ‘sample’ (here sample refers to the scanned handwritten character to be recognized) needs to be pre-processed before applying segmentation.

3.1 Pre-processing:

The pre-processing includes binarization of the sample, thinning the sample, white space removal and extended headline removal.

3.1.1 Binarization:

The first step of pre-processing is to convert the image into a bi-tonal image. A bi-tonal image only contains two tones or colors, white and black. If a pixel is black, then it is considered to be a part of the sample, and if it is white, then it is a part of the background.

The following images show the output of this step:

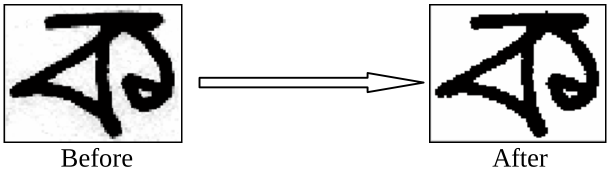


Figure 3: The effect of Binarization.

3.1.2 Thinning:

The next step of pre-processing is to apply ‘thinning’ to the bi-tonal sample which gives the ‘skeletal representation’. The algorithm, we implemented for this purpose, iteratively deletes pixels inside the shape to shrink it without shortening it or breaking it into parts considering 8 neighbouring pixels in the 3 by 3 neighbourhood. It has the following steps:

For each black pixel

Step1: Find out the number of black neighbour pixels.

Step2: Find out number of transitions from black to white (or white to black) in the neighbourhood.

Step3: The subjected pixel is marked if any of the following cases is true

Case1: all neighbours are black.

Case2: all neighbours are black except one.

Case3: all neighbours are white.

Case4: all neighbours are white except one.

Case5: number of transitions from black to white (or white to black) is less than or equal to one.

Step4: Convert all marked pixels to white.

The following images show the output of this algorithm:

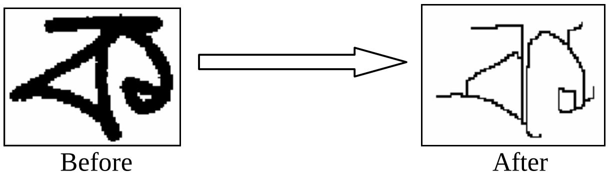


Figure 4: The effect of Thinning.

3.1.3 White Space Removal & Extended Headline Reduction:

The sample may contain unnecessary white space around itself. So, the next step is to remove these extra white pixels to improvise the relative study of samples for better recognition.

In addition to this white space removal, this process also reduces the extended headline of the sample (if required). The following images are showing how this step worked on the sample.

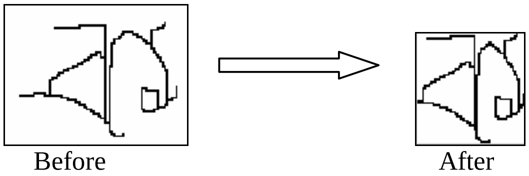


Figure 5: The effect of White Space Removal.

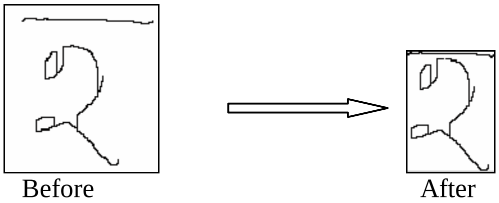


Figure 6: The effect of Extended Headline Reduction.

3.2 Preparation of Hierarchical Clusters based on Segmentation:

In this step, standard recognized Bengali character samples are segmented to extract unique characteristics (features). Based on these features, test samples can be categorized into their respective clusters in a hierarchical manner, such that, every leaf node of the hierarchy will contain one unique character cluster. The algorithm defined to categories the test samples into their respective clusters using the extracted features is as follows:

Step1: *The sample contains vertical line covering at least 75% of its height (Say it is called the **spine** of the sample).* If yes, go to Step 1.1, else go to Step 1.2.

Step1.1: *The sample contains two vertical parallel lines, each covering at least 75% of its height.* If yes, go to Step 1.1.1, else go to Step 1.1.2.

Step1.2: *The sample contains a curved line above the headline.* If yes, go to Step 1.2.1, else go to Step 1.2.2.

Step1.1.1: *The sample contains a curve ∪ at the lower half.* If yes, it is identified as the Bengali vowel আ (aa), else go to Step 1.1.1.1

Step1.1.2: *The sample contains a curved line above the headline.* If yes, go to Step 1.1.2.1, else go to Step 1.1.2.2.

Step1.2.1: *The sample contains a closed loop.* If yes, go to Step 1.2.1.1, else it is identified as the Bengali vowel উ (u).

Step1.2.2: *The sample contains a curve ∪ at the lower half.* If yes, go to Step 1.2.2.1, else go to Step 1.2.2.2.

Step1.1.1.1: *The sample contains a curved line above the headline.* If yes, it is identified as the Bengali vowel ঋ (rhi), else it is identified as the Bengali consonant ঝ (jha).

Step1.1.2.1: *The sample contains no black pixel to the right hand side of its spine.* If yes, it is identified as the Bengali consonant ঢ (dha), else go to Step 1.1.2.1.1.

Step1.1.2.2: *The sample contains maximum number of black pixels to the left hand side of its spine.* If yes, go to Step 1.1.2.2.1, else go to Step 1.1.2.2.2.

Step1.2.1.1: *The sample contains a curve ∪ at the lower half.* If yes, go to Step 1.2.1.1.1, else go to Step 1.2.1.1.2.

Step1.2.2.1: *The sample contains a loop.* If yes, go to Step 1.2.2.1.1, else go to Step 1.2.2.1.2.

Step1.2.2.2: *The sample contains a loop.* If yes, go to Step 1.2.2.2.1, else it is identified as the Bengali consonant দ (da).

Step1.1.2.1.1: *The sample contains a loop on the right hand side of the spine.* If yes, it is identified as the Bengali consonant ট(tta), else it is identified as the Bengali vowel ঐ (oi).

Step1.1.2.2.1: *The sample contains more black pixels at the lower half.* If yes, it is go to Step 1.1.2.2.1.1, else go to Step 1.1.2.2.1.2.

Step1.1.2.2.2: *The sample contains black pixels on the left hand side of its spine.* If yes, it is go to Step 1.1.2.2.2.1, else go to Step 1.1.2.2.2.2.

Step1.2.1.1.1: *The sample contains no headline.* If yes, it is identified as the Bengali vowel ঔ(ou), else it is identified as the Bengali consonant ঠ(ttha).

Step1.2.1.1.2: *The sample contains (curve at right hand side.* If yes, it is identified as the Bengali vowel ঈ (ii), else it is identified as the Bengali vowel ই (i).

Step1.2.2.1.1: *The sample contains loop at its bottom side.* If yes, it is identified as the Bengali consonant ড় (ddra), else go to Step 1.2.2.1.1.1.

Step1.2.2.1.2: *The sample contains more black pixels at right hand side.* If yes, it is identified as the Bengali consonant জ (ja), else it is identified as the Bengali consonant ড (dda).

Step1.2.2.2.1: *The sample contains a vertical straight line covering less than 50% of its height.* If yes, it is identified as the Bengali consonant ছ (chha), else go to Step 1.2.2.2.1.1.

Step1.1.2.2.1.1: *The sample contains a) curve in the left half.* If yes, it is identified as the Bengali consonant য় (yya), else go to Step 1.1.2.2.1.1.1.

Step1.1.2.2.1.2: *The sample contains less black pixels at the lower half.* If yes, it is go to Step 1.1.2.2.1.2.1, else go to Step 1.1.2.2.1.2.2.

Step1.1.2.2.2.1: *The sample contains (curve at left hand side.* If yes, it is identified as the Bengali consonant ক (ka), else it is identified as the Bengali consonant ফ (pha).

Step1.1.2.2.2.2: *Horizontal scan lines cross the sample at two or less points.* If yes, it is identified as the Bengali consonant চ (cha), else go to Step 1.1.2.2.2.2.1.

Step1.2.2.1.1.1: *The sample contains (curve.* If yes, it is identified as the Bengali consonant ভ (bha), else go to Step 1.2.2.1.1.1.1.

Step1.2.2.2.1.1: *The sample contains negligible number of black pixels at the bottom left corner.* If yes, it is identified as the Bengali consonant ং (anushwar), else go to Step 1.2.2.2.1.1.1.

Step1.1.2.2.1.1.1: *The sample contains a loop in the upper half.* If yes, it is identified as the Bengali vowel এ (e), else go to Step 1.1.2.2.1.1.1.1.

Step1.1.2.2.1.2.1: *The sample contains no loop.* If yes, it is identified as the Bengali consonant গ (ga), else go to Step 1.1.2.2.1.2.1.1.

Step1.1.2.2.1.2.2: *The sample contains a curve ◡at the lower half.* If yes, it is identified as the Bengali vowel অ (a), else go to Step 1.1.2.2.1.2.2.1.

Step1.1.2.2.2.2.1: *The sample contains only one loop.* If yes, it is identified as the Bengali consonant ঢ (ddha), else it is identified as the Bengali consonant ঢ় (dhra).

Step1.2.2.1.1.1.1: *The sample contains two) curves horizontally.* If yes, it is identified as the Bengali vowel ও (o), else it is identified as the Bengali consonant ত (ta).

Step1.2.2.2.1.1.1: *Horizontal scan lines cross the sample at two or less points.* If yes, it is identified as the Bengali consonant ঃ (visargha), else go to Step 1.2.2.2.1.1.1.1.

Step1.1.2.2.1.1.1.1: *Horizontal scan lines cross the sample at more than three points.* If yes, it is identified as the Bengali consonant ল (la), else go to Step 1.1.2.2.1.1.1.1.1.

Step1.1.2.2.1.2.1.1: *The sample contains one loop.* If yes, it is identified as the Bengali consonant প (pa), else it is identified as the Bengali consonant শ (sha).

Step1.1.2.2.1.2.2.1: *The sample contains loop at left half.* If yes, it is go to Step 1.1.2.2.1.2.2.1.1, else go to Step 1.1.2.2.1.2.2.1.2.

Step1.2.2.2.1.1.1.1: *The sample contains one loop only.* If yes, it is identified as the Bengali consonant ৎ (khand-ta), else it is identified as the Bengali consonant হ (ha).

Step1.1.2.2.1.1.1.1.1: *The sample contains one loop.* If yes, it is identified as the Bengali consonant ন (na), else it is identified as the Bengali consonant র (ra).

Step1.1.2.2.1.2.2.1.1: *The sample contains loop at top half.* If yes, it is go to Step 1.1.2.2.1.2.2.1.1.1, else it is identified as the Bengali consonant ম (ma).

Step1.1.2.2.1.2.2.1.2: *The sample contains (curve at top left quadrant.* If yes, it is go to Step 1.1.2.2.1.2.2.1.2.1, else go to Step 1.1.2.2.1.2.2.1.2.2.

Step1.1.2.2.1.2.2.1.1.1: *The sample contains (curve at left hand side.* If yes, it is identified as the Bengali consonant খ (kha), else it is identified as the Bengali consonant থ (tha).

Step1.1.2.2.1.2.2.1.2.1: *The sample contains (curve at bottom left quadrant.* If yes, it is identified as the Bengali consonant ঘ (gha), else it is identified as the Bengali consonant স (sa).

Step1.1.2.2.1.2.2.1.2.2: *The sample contains) curve at left hand side.* If yes, go to Step 1.1.2.2.1.2.2.1.2.2.1, else it is identified as the Bengali consonant ব (ba).

Step1.1.2.2.1.2.2.1.2.2.1: *Vertical scan lines at the mid crosses the sample at three points.* If yes, it is identified as the Bengali consonant ষ (shha), else it is identified as the Bengali consonant য (yya).

4. Results & Discussion:

To test the effectiveness of the above Bengali Handwritten character recognition procedure, handwritten samples from each character class are fed to it. For example-suppose ষ and উ are the test samples to be detected. After preprocessing, the detection algorithm is to be applied on the processed samples. In the first case, the detection algorithm will give the following resulting steps:

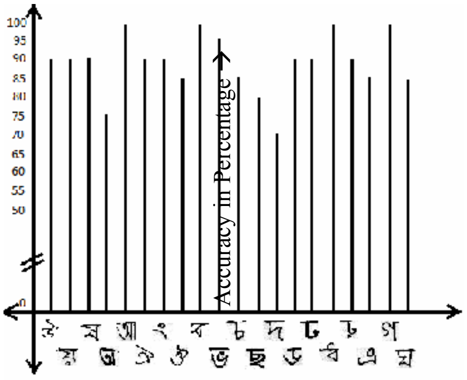
Feature description	**Present in sample?(Y/N)**	**Next step to follow**
Spine	Y	1.1
Vertical line parallel to spine	N	1.1.2
Curved line above the headline	N	1.1.2.2
Maximum number of black pixels to the left hand side of the spine	Y	1.1.2.2.1
More black pixels at the lower half	N	1.1.2.2.1.2
Less black pixels at the lower half	N	1.1.2.2.1.2.2
Less number of black pixels in the top left corner	N	1.1.2.2.1.2.2.1
Loop at left half	N	1.1.2.2.1.2.2.1.2
(curve at top left quadrant	N	1.1.2.2.1.2.2.1.2.2
) curve at left hand side	Y	1.1.2.2.1.2.2.1.2.2.1
Vertical scan line at the mid has three crossections	Y	The sample is identified as the Bengali consonant ষ (shha).

Whereas, in the second case, the detection algorithm will give the following resulting steps:

Feature description	**Present in sample?(Y/N)**	**Next step to follow**
Spine	N	1.2
Curved line above the headline	Y	1.2.1
Closed loop	N	The sample is identified as the Bengali vowel উ (u).

In the above described algorithm, several comparisons of the number of black pixels present at different sides and quadrants of a sample have been made. For this purpose, certain threshold values for comparison have been identified on a trial and error basis to achieve better results.

The accuracy of the algorithm has been measured for each of the Bengali character classes and the mean of these values can be treated as the overall accuracy of the algorithm. The accuracy measured for some of the Bengali characters are graphically depicted below:



Bengali Handwritten Characters →

Figure 7: Accuracy chart

and, the overall accuracy of the algorithm is approximately 87.34%.

5. Conclusion:

In this paper we have presented a procedure to recognize Bengali handwritten characters using hierarchical clustering. Successful implementation of this process will make it possible to translate Bengali manuscripts into other languages, to convert manuscripts to printable formats and many such applications. This procedure can be further improvised to include the recognition of whole Bengali handwritten words, sentences and texts containing compound characters (called ‘Yuktakshar’), punctuation marks etc.

6. References:

[1] Chaudhuri, B. B., Pal, U.: A Complete Printed Bangla OCR System. Pattern Recognition, Vol. 31. (1998) 531-549

[2] Pal, U., Chaudhuri, B. B.: Indian Script Character Recognition: A Survey: Pattern Recognition, Vol. 37 (2004) 1887-1899.

[3] Plamondon, R., Srihari, S. N.: On-Line and Off-Line Handwriting Recognition: AComprehensive Survey. IEEE Trans. Patt. Anal. and Mach. Intell., Vol. 22 (2000)
63-84

[4] Arica, N., Yarman-Vural, F.: An Overview of Character Recognition Focused on Off-line Handwriting. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Vol. 31 (2001) 216 - 233

[5] Trier, O. D., Jain, A. K. and Taxt, T.: Feature Extraction Methods for Character Recognition - A Survey. Pattern Recognition, Vol. 29 (1996) 641 – 662