A Novel Topographic Feature Extraction Method for Indian Character Images

Soumen Bag and Gaurav Harit

Indian Institute of Technology Kharagpur, Kharagpur-721302, India
{soumen,gharit}@cse.iitkgp.ernet.in

Abstract. In this paper, we present novel features based on the topography of a character as visible from different viewing directions on a 2D plane. By topography of a character we mean the structural features of the strokes and their spatial relations. In this work we develop topographic features of strokes visible with respect to views from different directions (e.g. North, South, East, and West). We consider three types of topographic features: closed region, convexity of strokes, and straight line strokes. We have tested the proposed method on printed and handwritten Bengali and Hindi isolated character images. Initial results demonstrate the efficacy of our approach.

Keywords: Convexity, Indian script, OCR, Thinning, Topographic feature.

1 Introduction

Feature selection and extraction has wide range of application for pattern recognition. It plays an important role in different classification based problems such as face recognition, signature verification, optical character recognition (OCR) etc.. From past several decades, many feature selection and extraction methods are reported [19] for Indian character recognition. The accuracy rate of OCR depends on feature sets and classifiers [16]. Next we begin a brief description about few important feature sets used in optical character recognition for Bengali and Hindi documents.

Chaudhuri and Pal [5] proposed the first complete printed Bengali OCR in 1998. In this method, nine different strokes are used as primary feature set for recognizing basic characters and template-based features are used for recognizing compound characters. To recognize handwritten basic and compound characters, Das *et al.* [6] used shadow, longest run, and quad-tree based features. Dutta and Chaudhuri [7] proposed topological features, such as junction points, hole, stroke segments, curvature maxima, curvature minima, and inflexion points of character images for performing printed and handwritten Bengali alpha-numeric character recognition. To detect convexity of Bengali numerals, Pal and Chaudhuri [13] used water-flow model. They also used topological and statistical features for preparing feature set. Bhowmick *et al.* [4] proposed a stroke-based feature set for recognizing Bengali

N. Meghanathan et al. (Eds.): CCSIT 2011, Part I, CCIS 131, pp. 358–367, 2011.

handwritten character images. In this method, ten stroke based features which indicate the shape, size and position information of a digital curve with respect to the character image, are extracted from character images to form the feature vector. Majumdar [11] have introduced a new feature extraction method based on the curvelet transform of morphologically altered versions of an original character image. Table 1 gives a summary of different feature sets used in Bengali OCR systems.

Table 1. Different feature sets used in Bengali OCR systems

Method	Feature set
Chaudhuri and Pal [5]	Structural and template features
Das *et al.* [6]	Shadow, longest run and quad-tree based features
Dutta and Chaudhury [7]	Structural and topological features
Pal and Chaudhuri [16]	Watershed, topological, and statistical features
Bhowmick *et al.* [4]	Stroke based feature
Majumdar [11]	Curvlet coefficient features

The first complete OCR on Devanagari was introduced by Pal and Chaudhuri [12]. In this method, they have used structural and template features for recognizing basic, modified, and compound characters. To recognize real-life printed documents of varying size and font, Bansal and Sinha [3] proposed statistical features. Later Pal *et al.* [15] used the same gradient features for recognizing handwritten Devanagari characters. Bajaj *et al.* [2] used density, moment of curve and descriptive component for recognizing Devanagari handwritten numerals. Sethi and Chatterjee [17] proposed a set of primitives, such as, global and local horizontal and vertical line segments, right and left slant, and loop for recognizing handwritten Devanagari characters. Sharma *et al.* [18] used directional chain code information of the contour points of the characters for recognizing handwritten Devanagari characters. Table 2 gives a summary of different feature sets used in Devanagari OCR systems.

Table 2. Different feature sets used in Devanagari OCR systems

Method	Feature set
Pal and Chaudhuri [12]	Structural and template features
Bansal and Sinha [3]	Statistical features
Bajaj *et al.* [2]	Density, moment of curve and descriptive component features
Sethi and Chatterjee [17]	Line segments, slant, and loop
Pal *et al.* [15]	Gradient features
Sharma [18]	Directional chain code information of contour points

All of the above mentioned methods do not consider shape variation for extracting features. But in Indian languages, a large number of similar shape type characters (basic and conjunct) are present. From that point of view, we have proposed novel features based on the topography of a character to improve the performance of existing OCR in Indian script, mainly for Bengali and Hindi documents. The major features of the proposed method are listed as follows:

1. The main challenge to design an OCR for Indian script is to handle large scale shape variation among different characters. Strokes in characters can be decomposed into segments which are straight lines, convexities or closed boundaries (hole). In our work we consider the topography of character strokes from 4 viewing directions. In addition to the different convex shapes formed by the character strokes, we also note the presence of closed region boundaries.
2. The extracted features are represented by a shape-based graph where each node contains the topographic feature, and they all are placed with respect to their centroids and relative positions in the original character image.
3. This approach is applicable for printed as well as handwritten text documents of different languages.

This paper is organized as follows. Section 2 describes the proposed topographic features extraction method. Section 3 contains the experimental results. This paper concludes with some remarks on the proposed method and future work in Section 4.

2 Proposed Topographic Feature Extraction Method

In this section we describe our proposed feature extraction method based on measurement of convexity of character strokes from different directions.

2.1 Preprocessing

Given a scanned document page we binarize it using the Otsu's algorithm [8]. Currently we are working with documents with all text content. For Bengali or Hindi documents the entire word gets identified as a single connected component because of the *mātrā/shiro-rekhā* (head line) which connects the individual characters. For this case we separate out the individual *aksharā* within a word by using the character segmentation methods reported in [14] (for Bengali) and [10] (for Hindi).

2.2 Thinning to Get Skeletonized Image

Before extracting the features, character images are converted to single pixel thick images. But to retain the proper shape of thinned character images is a big challenge. Lot of works have been done on thinning. But most of the works are reported for English, Chinese, and Arabic languages [9]. From this point of view, Bag and Harit have proposed an improved medial-axis based thinning strategy for Indian character images. Details of the methodology is reported in [1]. Few results are shown in Fig. 1.

2.3 Topographic Features Extraction

Topographic features are classified into three categories: closed region, convexity of strokes, and straight line strokes.The convexity of curve is detected from different directions. Here we consider four directions (North, South, East, and West) for convexity measurement.

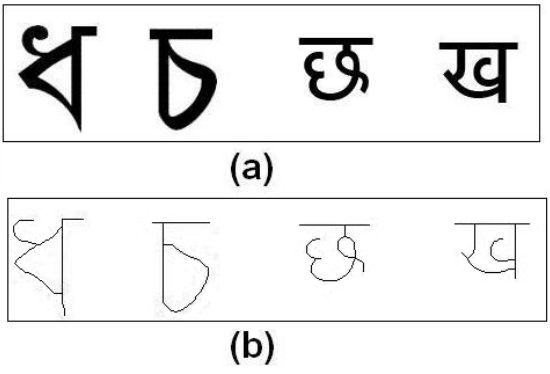

Fig. 1. Skeletonization results: (a) Input image; (b) Skeleton image

Definition of convexity: The convexity of a shape is detected by checking its convergence towards a single point or a cluster of points, connected by 4-connectivity. For any convexity shape, if we move down word then we shall reach to a single point or a flat region which is a set of 4-connected pixels. For detecting the stroke convexity, we have used this concept. Fig. 2 shows the different convexities of character images detected from different view directions. In the figure, different colors are used to mark convexities detected from different view directions (Red, Blue, Green, and Magenta for North, South, East, and West respectively).

The following steps discuss the methodology to detect convexity from North direction.

1. Prepare a database containing different convex shape. Each convex shape is defined according to their structural property. Fig. 3 shows different shapes stored in the database.

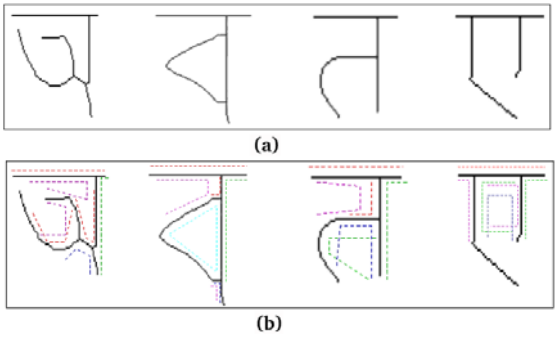

Fig. 2. (a) Skeleton image; (b) Different convex shapes marked by different colors (Red, Blue, Green and Magenta for North, South, East, and West respectively)

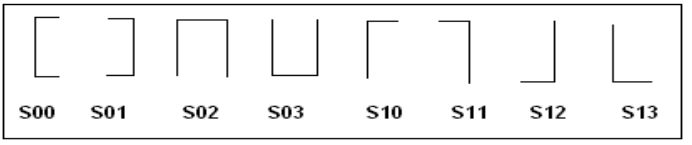

Fig. 3. Different convex shapes

2. Detect horizontal-start (H_s), horizontal-end (H_e), vertical-start (V_s), and vertical-end (V_e) points of the thinned image. Here $[H_s, H_e]$ and $[V_s, V_e]$ are the horizontal and vertical limits of the input image respectively (see Fig. 4).
3. Given a thinned binary image, it is scanned from top to bottom and left to right, and transition from white (background) to black (foreground) are detected. The whole scanning is done from H_s to H_e and V_s to V_e in horizontal and vertical directions respectively.
4. Suppose, a single scan from left to right generates a sequence of pixels $\langle x_1, x_2, x_3, ..., x_i, x_{i+1}, ..., x_n \rangle$ where each pixel is a cut point between horizontal scan line and character image. Now, if the two consecutive pixels $(x_i \ \& \ x_{i+1})$ are not 4-connected neighbors, then put them in an array of points $P[1 : N, 1 : 2]$ where each cell $P[i]$ contains one pixel with its x, y coordinates in $P[i][1]$ and $P[i][2]$ respectively.
5. Continue the above steps for all remaining horizontal scan until we get a single point or a set of horizontally 4-connected neighbor pixels with value $\geq \zeta$ (set to 5).
6. The detected set of pixels are matched with the defined convex shapes stored in the database to get a specific convex shape.
7. The above steps are repeated for detecting the convexity from remaining three directions, i.e. South, East, and West. Only difference is that, for South direction, the scan is done from bottom to top and left to right, and for East/West directions, scan is done from top to bottom and right to left (for East)/left to right (for West) directions. Finally, a set of different convex shapes are generated to prepare topographic feature set.

The above steps are used for detecting convexity of character strokes from different directions. Now for detecting closed region, we use the concept of connected

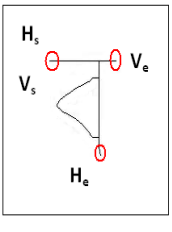

Fig. 4. Horizontal and vertical limits of an image

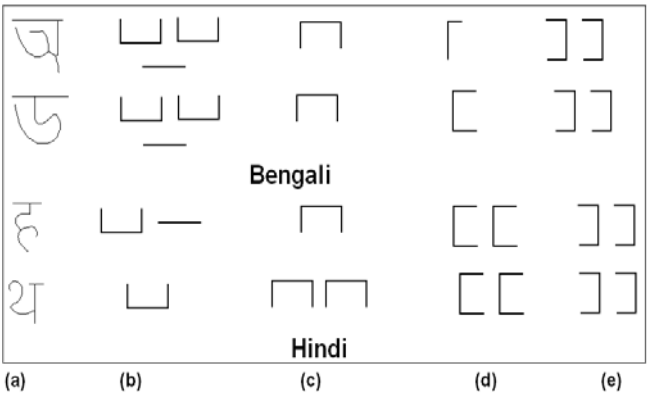

Fig. 5. Topographic features of thinned character images: (a) Skeleton image; (b)-(e) Features from North, South, East, and West direction

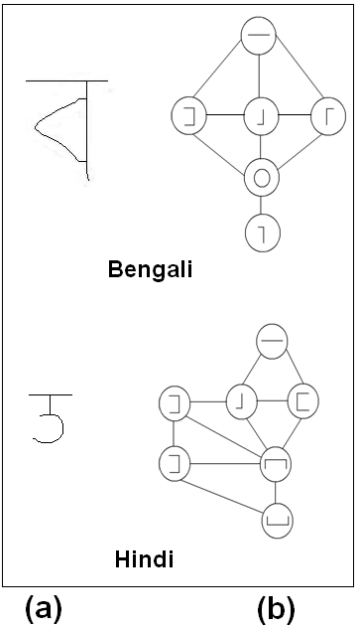

Fig. 6. (a) Skeleton image; (b) Shape-based graph

component analysis. If the pixels of the stroke segment are connected (i.e. each pixel has two 8-connected neighbors), then the stroke is identified as **closed region**.

If the number of horizontally 8-connected neighbor pixels is $\geq \Omega$ (set to 20), then we can say that **straight line** is present. For Bengali and Hindi language, most of the characters have headline. By observing the structural shape of these two type of characters, we can say that the straight line will detect only from

North direction. For this reason, we do not use straight detection method for other three directions.

Fig. 5 shows the different topographic features extracted from Bengali and Hindi thinned character images.

2.4 Graphical Representation of Topographic Features

After detecting the convexity from four directions, we design a shape-based graph to represent the feature set of a particular character. The steps are given below.

1. Suppose, a thinned character image has k number of topographic components $< T_1, T_2, ..., T_k >$.
2. For each component, calculate the centroid(X_i, Y_i) using equation 1.

$$X_i = \frac{\sum_{j=1}^{N_i} p_{ij}}{N_i} \; ; \qquad Y_i = \frac{\sum_{j=1}^{N_i} p'_{ij}}{N_i} \tag{1}$$

where N_i is the total number of black pixels of the i^{th} topographic component, p_{ij} and p'_{ij} are the x-coordinate and y-coordinate of the i^{th} pixel of the j^{th} topographic component respectively.

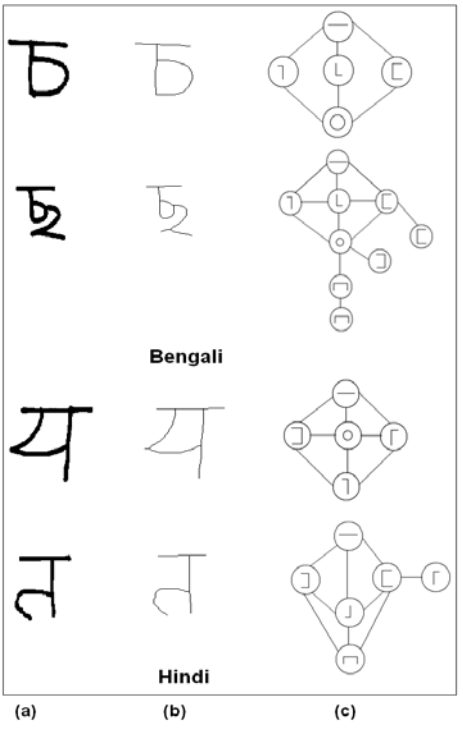

Fig. 7. Topographic feature extraction of handwritten character images: (a) Input image; (b) Skeleton image; (b) Shape-based graph

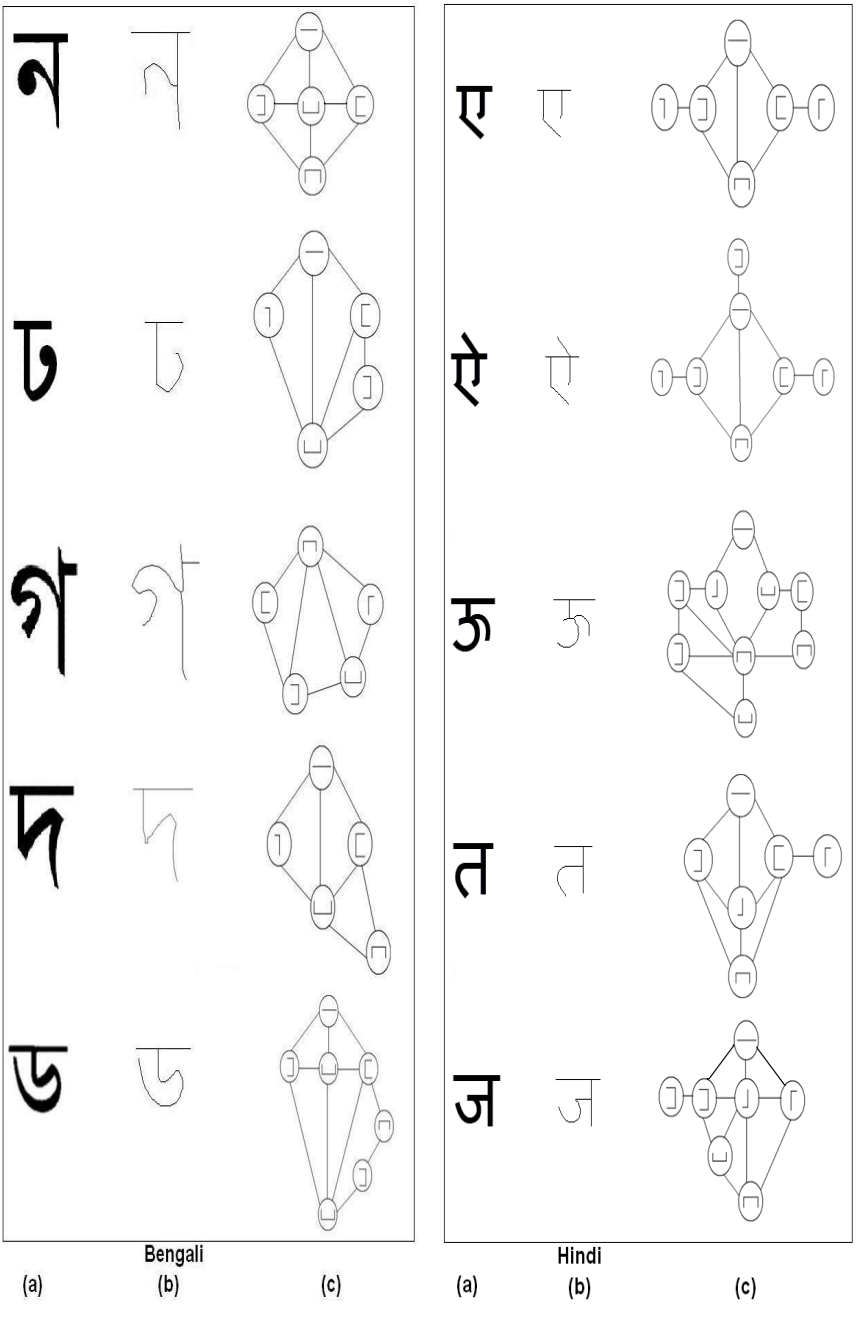

Fig. 8. Topographic features of printed character images: (a) Input image; (b) Skeleton image; (c) Shape-based graph

3. Design an undirected graph $G = (V, E)$, where V is the set of vertices containing different topographic features and E is the set of edges. All vertices are placed based on the coordinates of there centroid. Now add edges among these vertices with respect to their relative positions in the thinned character image.

Fig. 6 shows the shape-based graph of thinned Bengali and Hindi character images. This graphical model gives a clear pictorial difference among different feature set of different character images.

3 Experimental Results

We collected characters from several heterogeneous printed and handwritten documents of Bengali and Hindi. The number of characters in the testing set is 1250 (600 for Bengali and 650 for Hindi). All the characters are collected in a systematic manner from printed and handwritten pages scanned on a HP scanjet 5590 scanner at 300 dpi. At first, images were thinned by Bag and Harit thinning method [1]. This thinning method has the capability to preserve the shape of character images at junction points and end points. Then we tested our proposed feature extraction method on these thinned images to get topographic features. Fig. 8 shows the shape-graph representation of topographic features of printed Bengali and Hindi character image. The algorithms is implemented in C++ programming language using OpenCV 2.0 on Unix/Linux platform.

We applied this proposed method on handwritten Bengali and Hindi character images. Fig. 7 shows few results. It is observed that the performance is satisfactory for handwritten characters also.

4 Conclusion

In this paper, we have proposed a novel topographic feature extraction method for Indian OCR systems. This feature set captures close region, convexity of stroke from different direction, and flat region of thinned character images. The proposed method is tested on printed and handwritten Bengali and Hindi documents and we have obtained promising results. The proposed feature set helps to discriminate two similar type characters properly. In future, we shall extend our work to extract topographic features for other popular Indian languages and make it as a script independent feature set for designing multi-lingual OCR in Indian scripts.

References

1. Bag, S., Harit, G.: A medial axis based thinning strategy for character images. In: Proceedings of the 2^{nd} National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, Jaipur, India, pp. 67–72 (2010)
2. Bajaj, R., Dey, L., Chaudhury, S.: Devnagari numeral recognition by combining decision of multiple connectionist classifiers. Sadhana 27, 59–72 (2002)

3. Bansal, V., Sinha, R.M.K.: Integrating knowledge sources in Devanagari text recognition system. IEEE Transactions on Systems, Man, and Cybernetics–Part A: Systems and Humans 30(4), 500–505 (2000)
4. Bhowmick, T.K., Bhattacharya, U., Parui, S.K.: Recognition of Bangla handwritten characters using an MLP classifier based on stroke features. In: Proceedings of the 11^{th}International Conference on Neural Information Processing, Kolkata, India, pp. 814–819 (2004)
5. Chaudhuri, B.B., Pal, U.: A complete printed Bangla OCR system. Pattern Recognition 31(5), 531–549 (1998)
6. Das, N., Das, B., Sarkar, R., Basu, S., Kundu, M., Nasipuri, M.: Handwritten Bangla basic and compound character recognition using MLP and SVM classifier. Journal of Computing 2(2), 109–115 (2010)
7. Dutta, A., Chaudhury, S.: Bengali alpha-numeric character recognition using curvature features. Pattern Recognition 26(12), 1757–1770 (1993)
8. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Prentice Hall, USA (2008)
9. Lam, L., Lee, S.W., Suen, C.Y.: Thinning methodologies—A comprehensive survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 14(9), 869–885 (1992)
10. Ma, H., Doermann, D.: Adaptive Hindi OCR using generalized hausdorff image comparison. ACM Transactions on Asian Language Information Processing 2, 193–218 (2003)
11. Majumdar, A.: Bangla basic character recognition using digital curvelet transform. Journal of Pattern Recognition Research 2(1), 17–26 (2007)
12. Pal, U., Chaudhuri, B.B.: Printed Devnagari script OCR system. Vivek 10, 12–24 (1997)
13. Pal, U., Chaudhuri, B.B.: Automatic recognition of unconstrained off-line Bangla handwritten numerals. In: Proceedings of the 3^{rd} International Conference on Advances in Multimodal Interfaces, Beijing, China, pp. 371–378 (2000)
14. Pal, U., Datta, S.: Segmentation of Bangla unconstrained handwritten text. In: Proceedings of the 7^{th} International Conference on document Analysis and Recognition, Edinburgh, Scotland, pp. 1128–1132 (2003)
15. Pal, U., Sharma, N., Wakabayashi, T., Kimura, F.: Offline handwritten character recognition of Devnagari script. In: Proceedings of the 9^{th} International Conference on Document Analysis and Recognition, Curitiba, Brazil, pp. 496–500 (2007)
16. Pal, U., Wakabayashi, T., Kimura, F.: Comparative study of Devnagari handwritten character recognition using different feature and classifiers. In: Proceedings of the 10^{th} International Conference on Document Analysis and Recognition, Barcelona, Spain, pp. 1111–1115 (2009)
17. Sethi, K., Chatterjee, B.: Machine recognition of constrained hand-printed Devnagari. Pattern Recognition 9, 69–77 (1977)
18. Sharma, N., Pal, U., Kimura, F., Pal, S.: Recognition of offline handwritten Devnagari characters using quadratic classifier. In: Proceedings of the 5^{th} Indian Conference on Computer Vision, Graphics and Image Processing, Madurai, India, pp. 805–816 (2006)
19. Trier, O.D., Jain, A.K., Taxt, T.: Feature extraction methods for character recognition–A survey. Pattern Recognition 29(4), 641–662 (1996)