Handwritten Bangla Compound Character Recognition using Gradient Feature

U. Pal[1], T. Wakabayashi[2] and F. Kimura[2]

[1]Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata-108, India
Email: umapada@isical.ac.in
[2]Graduate School of Engineering, Mie University, 1577 Kurimamachiya-cho, Tsu, Mie, Japan

Abstract

Recognition of handwritten characters of Indian script is difficult because of the presence of many complex shaped compound characters (cluster characters) as well as variability involved in the writing style of different individuals. This paper deals with recognition of off-line Bangla handwritten compound characters using Modified Quadratic Discriminant Function (MQDF). The features used for recognition purpose are mainly based on directional information obtained from the arc tangent of the gradient. To get the feature, at first, a 2 X 2 mean filtering is applied 4 times on the gray level image and a non-linear size normalization is done on the image. A Roberts filter is then applied on the normalized image to obtain gradient image. Next, the arc tangent of the gradient (direction of gradient) is initially quantized into 32 directions and the strength of the gradient is accumulated with each of the quantized direction. Finally, the frequencies of these directions are down sampled using Gaussian filter to get 392 dimensional feature vectors. Using 5-fold cross validation technique we obtained 85.90% accuracy from a dataset of Bangla compound characters containing 20,543 samples.

1. Introduction

Recognition of handwritten characters has been a popular research area for many years because of its various application potentials. Some of its potential application areas are postal automation, bank cheque processing, automatic data entry, etc. There are many pieces of work towards handwritten recognition of Roman, Japanese, Chinese and Arabic scripts. Various approaches have been proposed by the researchers towards handwritten character recognition [1]. One of the widely used approaches is based on Neural Network. Here the network architecture is, at first, trained by a set of training data and then the trained networks classify the input. Some researcher used structural approach, where each pattern class is defined by structural description and the recognition is performed according to structural similarities. Statistical approach is also applied to character recognition [2]. It is insensitive to pattern noise and distortion but accurate modeling of statistical information is a tedious task. Combination of structural and statistical methods is also used by the researchers for character recognition [3]. Among others, support vector machines [4], Fourier and wavelet description [5], fuzzy rules [6], tolerant rough set [7], are reported in the literatures. Although India is a multi-lingual and multi-script country much research has not been done towards the recognition of handwritten Indian characters. In this paper, we propose a system towards the recognition of off-line handwritten Bangla *compound* characters (cluster characters).

Many pieces of work have been done towards the recognition of Indian printed scripts and at present OCR systems are commercially available for some of the printed Indian scripts [8]. Several pieces of research work exist on Indian printed characters but only a few attempts have been made towards the recognition of off-line handwritten Indian characters [8]. Among off-line handwritten work on Indian scripts, maximum research has been done for Bangla. Systems are available for off-line Bangla handwritten numerals and characters [9-13]. Also some systems have been developed for unconstrained Bangla handwritten word recognition for Indian postal automation [14].

Although some pieces of work have been done towards the recognition of handwritten Bangla numerals, only a few research papers are available for the recognition of Bangla handwritten characters. Recently Roy et al. [9] proposed a quadratic classifier based approach for Bangla basic character recognition. Basu et al. [10]

proposed an MLP based scheme for the recognition of Bangla characters and the feature set used for recognition includes 24 shadow features, 16 centroid features and 36 longest-run features. Bhattacharya et al. [11] proposed a hybrid scheme for recognition of handwritten Bangla basic characters. Rahman et al. [12] proposed a multistage approach for handwritten Bangla character recognition and the major features used for the multistage approach include *matra/shirorekha*, upper part of the character, disjoint section of the character, vertical line, double vertical line etc.

To the best of our knowledge no work is published on Bangla handwritten compound character recognition. In this paper, we propose a scheme for unconstrained off-line Bangla handwritten compound character recognition. Compound character recognition is more difficult than basic character recognition because of the complex shaped nature.

The features used in this paper for recognition purpose are mainly based on directional information obtained from the gradient [15,16]. To get the feature, at first, a non-linear size normalization technique is applied on the gray image obtained after the mean filtering of the image. The normalized image is then segmented to 49 x 49 blocks. A Roberts filter is then applied to obtain gradient image. Next, the arc tangent of the gradient (direction of gradient) is initially quantized into 32 directions and the strength of the gradient is accumulated with each of the quantized direction. Finally, these directions are down sampled using Gaussian filter to get 8 directions from 32 directions. Also 49 x 49 blocks are down sampled using Gaussian filter into 7 x 7 blocks. As a result, we get 7 x 7 x 8 = 392 dimensional feature vector and this feature is fed to the quadratic classifier for recognition.

Rest of the paper is organized as follows. In Section 2 we discuss about Bangla language, its character set and compound character formation, and data collection for the experiment. Feature extraction procedure is presented in Section 3. Section 4 details the classifier used for the recognition. The experimental results are discussed in Section 5. Finally, conclusion on the paper is given in Section 6.

2. Bangla language and data collection

Bangla, the second most popular language in India and the fifth most popular language in the world, is an ancient Indo-Aryans language. More than 200 million people in the eastern part of Indian subcontinent speak in this language. Bangla script alphabet is used in texts of Bangla, Assamese and Manipuri languages. Also, Bangla is the national language of Bangladesh.

The alphabet of the modern Bangla script consists of 11 vowels and 39 consonants. These characters are called as *basic characters*. Basic characters of printed characters are shown in Fig.1. Writing style in Bangla is from left to right and the concept of upper/lower case is absent in this script. It can be seen that most of the characters of Bangla have a horizontal line at the upper part of the character. We call this line as *Matra/Shirorekha*. In Bangla, characters are not alphabetical, like English (Roman) where the characters largely have one-sound one- symbol characteristics. It is a mixture of syllabic and alphabetic characters.

Fig.1. Examples of Bangla basic characters (First 11 are vowels).

In Bangla script a vowel following a consonant takes a modified shape. Depending on the vowel, its modified shape is placed at the left, right, both left and right, or bottom of the consonant. These modified shapes are called *modified characters*.

A consonant or a vowel following a consonant sometimes takes a compound orthographic shape, which we call as *compound character*. Compound characters can be combinations of two consonants as well as a consonant and a vowel. Compounding of three or four characters is also possible in Bangla. There are more than 200 compound characters in Bangla [8]. In this paper we consider the recognition of such complex shaped compound characters, and we consider 138 popular Bangla compound characters for recognition. A detail study of Bangla character occurrence statistics can be found in [17].

Examples of some Bangla compound character formations are shown in Fig.2. Occasionally, in Bangla combination of two basic characters forms a new shape as shown in the first two rows of Fig.2. In the third and fourth rows of Fig.2 one of the constituent character of the compound character retains its shape and the other constituent character reduces its size in the compound character. In the compound character shown in fifth row, two characters sit side by side in compounding, but the size of the first character is slightly reduced. In the compound characters depicted in the sixth and seventh rows are formed by three basic characters where shape of none of its constituent basic characters cannot be found. Since the formation of compound characters is different,

hence it is very difficult to recognize Bangla compound characters.

Constituent basic characters		Compound characters
ক + র	=	ক্র
ঙ + গ	=	ঙ্গ
জ + ব	=	জ্ব
ন + ট	=	ন্ট
চ + ছ	=	চ্ছ
ক + ষ + ম	=	ক্ষ্ম
স + ত + র	=	স্ত্র

Fig.2. Illustration of Bangla compound character formation from its basic characters.

To get an idea of Bangla compound characters and their variability in handwriting, a set of some handwritten Bangla compound characters are shown in Fig.3. Main difficulty of any character recognition system is the shape similarity. It can be noted that because of handwritten style, two different characters in Bangla may look very close to another character. For example see Fig.4 where some similar shaped Bangla compound character pairs are shown. In this figure there are six groups and similar shaped characters are clustered into individual group. From the groups it can be seen that shape of the two compound characters of a group is very similar and such shape similarity makes the recognition system more complex.

Data collection for the present work has been done from different individuals of various professions. We have collected 20,543 samples of Bangla handwritten compound characters and these data are collected from specially designed form. Minimum number of samples of a class was 110. We used a flatbed scanner for digitization. Digitized images are in gray tone with 300 dpi and stored as Tagged Image File (TIF) Format.

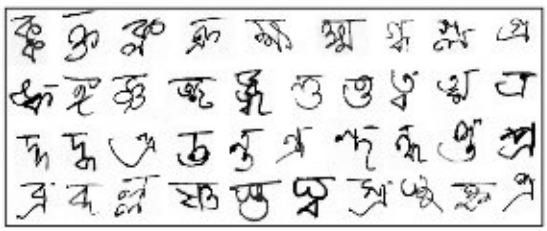


Fig.3. Examples of some Bangla handwritten compound characters.

Fig.4. Examples of some similar shaped Bangla compound characters.

3. Feature extraction

Here we used 392 dimensional feature vector for our experiment and to obtain this feature vector the following steps are executed.

Step 1: Compute the bounding box of the input gray-level image. The gray image portion within the bounding box is used for further processing. To get bounding box information of the grey image we use the binary version (obtained by Otsu method [18]) of the image.

Step 2: Apply 2×2 mean filtering 4 times on this gray image. The filtering is done to get better performance.

Step 3: A non-linear size normalization of the image is done [2]. Here the image is normalized into 148 x 148 pixels. This size is obtained from experiment.

Step 4: Apply again a 3×3 mean filtering 2 times on this normalized gray image.

Step 5: Normalized image is then segmented into 49 x 49 blocks.

Step 6: Apply Roberts filter on the image to obtain gradient image. The arc tangent of the gradient (direction of gradient) is quantized into 32 directions and the strength of the gradient is accumulated with each of the quantized direction. By strength of Gradient ($f(x,y)$) we mean $f(x,y) = \sqrt{(\Delta u)^2 + (\Delta v)^2}$, and by direction of gradient ($\theta(x,y)$) we mean $\theta(x,y) = \tan^{-1}\frac{\Delta v}{\Delta u}$, where $\Delta u = g(x+1, y+1) - g(x,y)$, and $\Delta v = g(x+1, y) - g(x, y+1)$, and $g(x,y)$ is a gray scale at (x, y) point.

Step 7: Histograms of the values of 32 quantized directions are computed in the 49 x 49 blocks. A smoothing filter [1 4 6 4 1] is used to get 16 directions from 32 directions. On this resultant image, another

smoothing filter [1 2 1] is used to get 8 directions from 16 directions. We also use a 31 x 31 two-dimensional Gaussian-like filter (see Fig.5) to get smoothed 7 x 7 blocks from 49 x 49 blocks (shown in Fig.6). Thus, we get 7×7×8 = 392 dimensional feature vector.

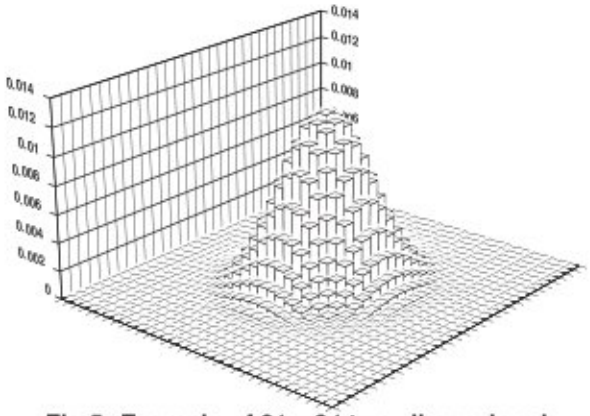


Fig.5: Example of 31 x 31 two-dimensional Gaussian-like filter used for smoothing.

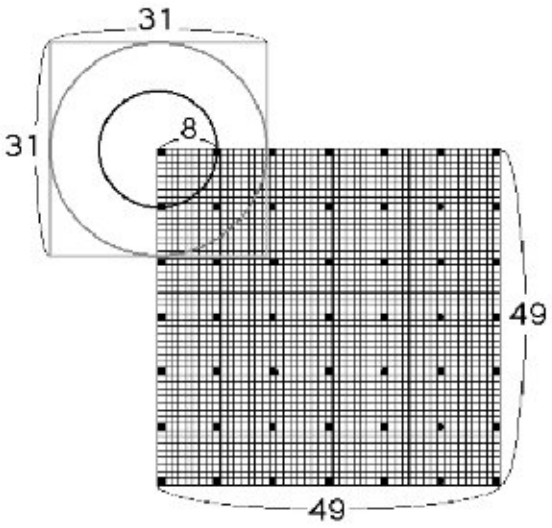


Fig.6: Illustration of getting 7 x 7 blocks from 49 x 49 blocks.

4. Character classifier

Character recognition is carried out using the following quadratic discriminant function [2]. Kimura et al. [19] compared seven statistical classifiers for handwritten zip-code numeral recognition and they obtained best results from quadratic classifier and hence we use this classifier for our experiment.

$$g(X) = (N + N_0 + n - 1)\ln[1 + \frac{1}{N_0\sigma^2}[\|X - M\|^2 - \sum_{i=1}^{k}\frac{\lambda_i}{\lambda_i + \frac{N_0}{N}\sigma^2}\{\Phi_i^T(X - M)\}^2]] + \sum_{i=1}^{k}\ln(\lambda_i + \frac{N_0}{N}\sigma^2)$$

where X is the feature vector of an input character; M is a mean vector of samples; Φ_i^T is the i^{th} eigen vector of the sample covariance matrix; λ_i is the i^{th} eigen value of the sample covariance matrix; k is the number of eigen values considered here, n is the feature size; σ^2 is the initial estimation of a variance; N is the number of learning samples; and N_o is a confidence constant for σ and N_0 is considered as N. We do not use all the eigen values and their respective eigen vectors for the classification. Here, we sort the eigen values in descending order and take first 100 (k=100) eigen values and their respective eigen vectors for classification.

5. Experimental results

Data used for the present work were collected from different individuals of various professions and 20,543 samples of Bangla compound characters are used for experiment. We have used 5-fold cross validation scheme for recognition result computation. Here database is divided into 5 subsets and testing is done on each subset using rest four of the subsets for learning. The recognition rates for all the test subsets are averaged to calculate recognition accuracy.

5.1 Global recognition results

From experiment we noted that the overall compound character recognition accuracy of the proposed scheme using 392 dimensional features was 85.90% when zero percent rejection was considered. 93.01% recognition accuracy was obtained when we considered first two top choices of the recognition result. The detail recognition results obtained from different top choices are given in Table 1.

Table 1: Recognition results based on different choices from top (without any rejection)

Top choices	% of Accuracy
1	85.90%
2	93.01%
3	95.34%
4	96.44%
5	97.16%

We also noted the accuracy of individual compound characters. Maximum accuracy (97.45%) was achieved for the Bangla compound character  . The next highest accuracy (97.12%) was achieved for the compound character  . Accuracies of some of the Bangla compound characters for which we got higher recognition rates are shown in Table 2 (in descending order).

Table 2: Individual accuracy of some Bangla compound characters

Character	Accuracy	Character	Accuracy
	97.45%		97.12%
	96.95%		96.93%
	96.91%		96.71%
	96.57%		96.24%

5.2 Rejection versus error rate computation

We also computed rejection versus error rate of the classifier and the results are presented in Table 3. The table depicts that 11.07% error occurred when rejection rate was 5.00% and only 2.08% error occurred when rejection rate was 30.00%. Rejection criteria of the proposed system was decided mainly based on the difference of 1st and 2nd value of the discriminant function *g(X)* described in Section 4.

Table 3: Rejection versus error rate obtained from the classifier

Rejection (%)	Error (%)
0.00	14.10
5.00	11.07
10.00	8.43
15.00	6.13
20.00	4.38
30.00	2.08
40.09	0.87

5.3 Confusing pair computation

We also notice the main confusing pairs of Bangla compound characters and their error rates are shown in Table 4. The characters  and confused maximum between them and their overall confusion rate was 0.13%. The next most confusing pair was  and , and they confused 0.11% cases. From the experiments we noticed that mainly similar shaped characters confused by the system at higher rate.

Table 4: Main confusion pairs of Bangla compound characters

Confusing character pairs		% of confusion
		0.13%
		0.11%
		0.10%
		0.09%
		0.09%

5.4 Results on poor/noisy image

Since we use statistical classifier and our feature detection technique scheme is not very sensitive to noise, our scheme shows correct results even if the samples are noisy and poor. To get an idea of such samples, some poor images where we get correct results from our system are shown in Fig.7.

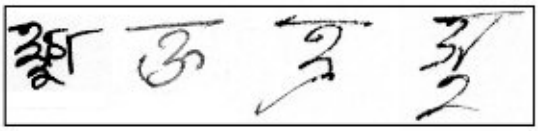
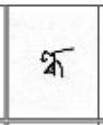
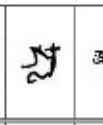
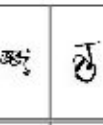
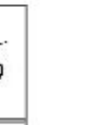
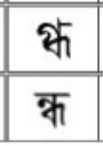
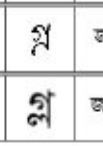
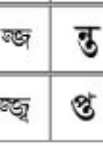
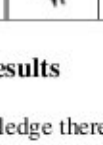
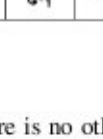


Fig.7: Examples of some poor samples.

5.5 Erroneous results

To get the idea about the samples where our system generates errors, we provide some erroneous samples in Table 5. Actual handwritten compound character samples are shown in the first row of this table and the printed samples of their recognized class are shown in the respective columns of second row. Since the actual handwritten samples and recognized characters are very similar in shape we may think that these samples are recognized correctly. Unfortunately they all mis-recognized. Actual class of each handwritten sample is shown in respective columns of the third row of the table.

Table 5: Examples of some erroneous samples

Actual Samples (Handwritten)				
Recognized as (Printed sample)				
Actual Class (Printed sample)				

5.5 Comparison of results

To the best of our knowledge there is no other published work on off-line handwritten Bangla compound character recognition and hence we cannot compare our recognition results.

6. Conclusions

India is a multi-lingual and multi-script country but not much work has been done towards off-line handwriting recognition of Indian script. In this paper we present a quadratic classifier based system towards the recognition of off-line Bangla handwritten compound characters. To the best of our knowledge there no published work on the compound characters of Bangla. From the experiment of 138-class Bangla compound characters we obtained encouraging results from our system although shapes of compound characters are very complex. The authors think this work will be helpful to the researchers for the handwritten recognition of other Indian scripts. In future we plan to perform different results from various feature sets and classifiers, and to compare such results.

References

[1] R. Plamondon and S. N. Srihari, "On-line and off-line handwritten recognition: A comprehensive survey", IEEE Trans. on PAMI, Vol.22, pp. 62-84, 2000.

[2] F Kimura, K, Takashina, S. Tsuruoka and Y. Miyake, "Modified quadratic discriminant function and the application to Chinese character recognition", IEEE Trans. on PAMI, Vol. 9, pp 149-153, 1987.

[3] J. Cai and Z. Q. Liu, "Integration of structural and statistical information for unconstrained handwritten character recognition", IEEE Trans. on PAMI, Vol.21, pp.263-270, 1999.

[4] H. Byan and S.W. Lee, "A Survey on pattern recognition application of support vector machines", IJPRAI, Vol.17, pp.459-486, 2003.

[5] P. Wunsch and A. F. Laine, "Wavelet descriptors for multi-resolution recognition of dand-printed digits", Pattern Recognition, Vol.28, pp.1237-1249, 1995.

[6] Z.Chin and H. Yan, "A handwritten character recognition using self-organizing maps and fuzzy rules", Pattern Recognition, Vol.22, pp. 923-937, 2000.

[7] K.Kim and S.Y. Bang, "A handwritten character classification using tolerant Rough set", IEEE Trans. on PAMI, Vol.22, pp.923-937, 2000.

[8] U. Pal and B. B. Chaudhuri, "Indian script character recognition: a survey", Pattern Recognition, Vol. 37, pp. 1887-1899, 2004.

[9] K. Roy, U. Pal and F. Kimura, "Bangla handwritten character recognition", International Journal of Tomography & Statistics, Vol. 5, pp. 27-36, 2007

[10] S. Basu, N. Das, R. Sarkar, M. Kundu, M. Nasipuri and D. K. Basu, "Handwritten Bangla alphabet recognition using an MLP based classifier" In Proc. of the 2nd NCCPB, 2005.

[11] U. Bhattacharya, S. K. Parui and B. Shaw, "A hybrid scheme for recognition of handwritten Bangla basic characters based on HMM and MLP classifiers", In Proc. 6th International Conference on Advances in Pattern Recognition, pp. 101-106, 2007.

[12] A. F. R. Rahman, R. Rahman, M. C. Fairhurst, "Recognition of handwritten Bengali characters: a novel multistage approach". Pattern Recognition, Vol. 35, pp. 997-1006, 2002.

[13] U. Pal and B. B. Chaudhuri, "Automatic recognition of unconstrained offline Bangla handwritten characters", Proc. Advances in Multimodal Interfaces, Springer Verlag Lecture Notes on Computer Science (LNCS-1948), pp 371-378, 2000.

[14] U. Pal, K. Roy and F. Kimura, "A lexicon driven method for unconstrained Bangla handwritten word recognition", In Proc. 10th International Workshop on Frontiers in Handwriting Recognition, pp. 601-606, 2006.

[15] G. Srikantan, S. W. Lam and S. N. Srihari, "Gradient-based contour encoding for character recognition", Pattern Recognition, Vol. 29, pp. 1147-1160, 1996.

[16] T. Wakabayashi, S. Tsuruoka, F Kimura and Y. Miyake, "Increasing the feature size in handwritten numeral recognition to improve accuracy", Systems and Computers in Japan, Vol. 26, pp. 2046-2053, 1995.

[17] B. B. Chaudhuri and U. Pal, "Relational studies between phoneme and grapheme statistics in current Bangla", Journal of Acoustical Society of India, vol.-23, pp. 67-77, 1995.

[18] N. Otsu, "A Threshold selection method from grey level histogram", IEEE Trans on SMC, Vol.9, pp.62-66, 1979.

[19] F. Kimura, S. Nishikawa, T. Wakabayashi, Y. Miyake and T. Tsutsumida, "Evaluation and synthesis of feature vectors for handwritten numeral recognition", IEICE Trans. on Information and Systems, Vol. E79-D, 26, pp. 436-442, 1996.