On Recognition of Handwritten Bangla Characters

U. Bhattacharya[1], M. Shridhar[2], and S.K. Parui[1]

[1] Computer Vision and Pattern Recognition Unit,
Indian Statistical Institute, Kolkata, India
{ujjwal, swapan}@isical.ac.in
[2] Department of Electrical and Computer Engineering,
University of Michigan-Dearborn, USA
mals@engin.umd.umich.edu

Abstract. Recently, a few works on recognition of handwritten Bangla characters have been reported in the literature. However, there is scope for further research in this area. In the present article, results of our recent study on recognition of handwritten Bangla basic characters will be reported. This is a 50 class problem since the alphabet of Bangla has 50 basic characters. In this study, features are obtained by computing local chain code histograms of input character shape. Comparative recognition results are obtained between computation of the above feature based on the contour and one-pixel skeletal representations of the input character image. Also, the classification results are obtained after down sampling the histogram feature by applying Gaussian filter in both these cases. Multilayer perceptrons (MLP) trained by backpropagation (BP) algorithm are used as classifiers in the present study. Near exhaustive studies are done for selection of its hidden layer size. An analysis of the misclassified samples shows an interesting error pattern and this has been used for further improvement in the recognition results. Final recognition accuracies on the training and the test sets are respectively 94.65% and 92.14%.

1 Introduction

India is a multilingual country of more than 1 billion population with 18 constitutional languages and 10 different scripts. Bangla is its second most popular script next to Devanagari. It is the script of two other Indian languages, *viz.*, Assamese and Manipuri. On the other hand, Bangla is the official language and script of Bangladesh, a neighbour of India. Thus, handwritten character recognition research for Bangla script has a lot of significance.

Significant research works on optical character recognition (OCR) for printed Indian scripts including Bangla [1] are found in the literature. A survey of Indian script character recognition research can be found in [2]. However, not much research work towards recognition of handwritten characters of Indian scripts is available. The technology of printed OCR cannot unfortunately be extended to

P. Kalra and S. Peleg (Eds.): ICVGIP 2006, LNCS 4338, pp. 817–828, 2006.

recognition of handwritten characters due to the enormous variability in people's handwriting styles. However, if a suitable technology for off-line recognition of handwritten characters of Indian scripts can be developed, automatic processing of hand-filled-in forms can be done in the Indian scenario.

Only a few studies [3,4,5,6] on off-line recognition of handwritten Bangla characters are available in the literature. However, there are several works on off-line recognition of handwritten Bangla numerals which include [7,8,9,10]. Also, there exists research work on recognition of handwritten Devanagari [11], Telugu [12], Tamil [13] and Oriya [14] characters.

Many diverse algorithms/schemes for handwritten character recognition [15,16] exist and each of these has its own merits and demerits. Possibly the most important aspect of a handwriting recognition scheme is the selection of an appropriate feature set which is reasonably invariant with respect to shape variations caused by various writing styles. A large number of feature extraction methods are available in the literature [17].

In the present article, a study on recognition of handwritten Bangla basic characters is presented. This study is based on a large database of real-life handwritten samples. Local chain code histogram features are computed based on both contour and skeletal representations of the input character image. During simulation it has been observed that the chain code histogram feature computed from the character contour provides better recognition results compared to the same corresponding to the character skeleton. This is justified by the fact that contour provides more information about a character shape than its skeleton. Classification results are also obtained after down sampling chain code histogram features in each of the above cases using Gaussian filter. MLP classifiers are used for classification purpose. In each of the above classification attempts, all the fifty classes are considered and not in a single case satisfactory recognition performance is achieved. However, an analysis of the misclassified samples show that most of the misclassifications occur within several subgroups of character classes. So, for each of these subgroups separate classifiers (for fewer classes) are trained and each sample is classified for the second time by a smaller MLP classifier according to the result of initial sub-grouping by the 50 class MLP classifier.

The rest of this article is organized as follows. In Section 2, the database used for training and test of the proposed recognition methods has been described. Recognition methodology is described in Section 3. Some details of our experimental results are provided in Section 4. Section 5 concludes the article.

2 Handwritten Bangla Character Database

All major Indian scripts including Bangla are mixtures of syllabic and alphabetic scripts. They are varied in character and form. Like most of the Indian languages the script of Bangla came from the ancient Indian script, Brahmi. This script runfrom left to right and has no equivalent to capital letters of Latin scripts.

The difficulty in automatic recognition of these handwritten Bangla characters arises from the facts that this is a moderately large symbol set, shapes are usually

extremely cursive even when written separately and there exist quite a few groups of almost similar shape characters in their handwritten forms. Basic characters of Bangla alphabet consist of 11 vowels and 40 consonants. However, the shapes of two consonant characters are the same. Thus, there are 50 different shapes in the Bangla basic character set. Shapes of Bangla basic characters are shown in Fig. 1.

Most of the existing off-line recognition studies on handwritten characters of Indian scripts are based on different databases collected either in laboratory environment or from smaller groups of the concerned population. However, it is an accepted fact that any genuine research work in this area primarily needs at least one representative database. In the present work, we have used a moderately large representative database of handwritten Bangla basic characters.

2.1 Data Collection

Samples of the present database were collected by distributing several standard application forms among different groups of population of the state of West Bengal in India. Subjects were asked to fill-up these forms in Bangla. Since data collected through such forms are not evenly distributed among the character classes, another specially designed form consisting of 2-dimensional array of rectangular boxes had been used for data collection purpose. Subjects were requested to write one single character of Bangla alphabet per box. No other restriction was imposed on the writers. The purpose of data collection was not disclosed to them so that they could produce samples reflecting their natural handwriting styles. In approximately 60% cases, the same subject was asked to write on both types of forms on two different occasions using his/her own writing instrument. In case writing instrument was not available with the subject, it was supplied at random from a set of different types of such instruments. All the above forms were printed on papers of different brands and the samples have been collected over a span of more than two years.

2.2 Data Preparation

The above filled-in forms were scanned at 300 d.p.i. resolution using a state-of-the-art HP flatbed scanner. These are stored as grayscale images using 1 byte per pixel. A software was used for extraction of isolated characters from individual boxes. Since such a software is bound to produce some erroneous results, all the TIF files of isolated character images were checked manually through their thumbnail view and manual extraction (using an image editor) was done whenever certain error in automatic extraction was detected.

2.3 Database Statistics

The present database of Bangla handwritten basic characters consists of 20187 isolated basic character images unevenly distributed over different classes. The main reason of this uneven distribution is that a major part of the data was collected using several standard forms in which entries are proper nouns and there are several characters in the Bangla alphabet which are rarely used in

proper nouns. However, this problem could only be partially tackled by using the specially designed form described above. The distribution of samples in 50 classes of this database is shown in Fig. 1.

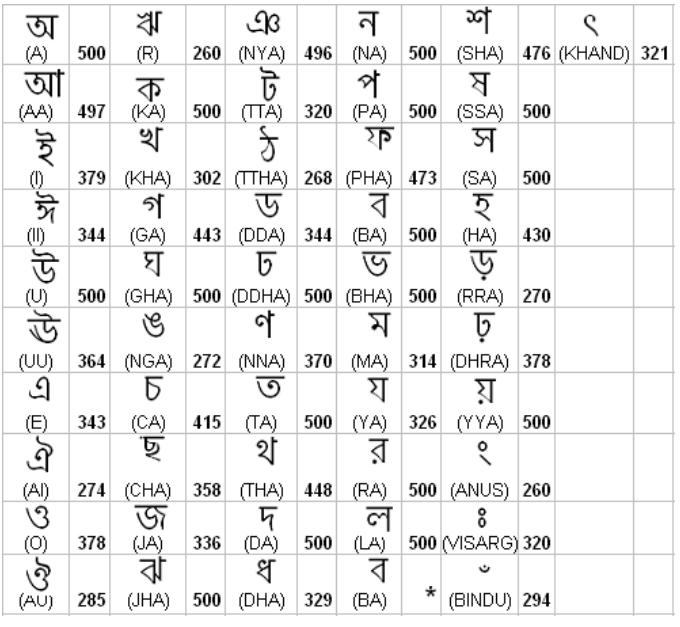


Fig. 1. Number of samples against each basic character shape is shown; within parentheses pronunciations are shown in English; * indicates that this shape occurred before

A small sample set consisting of a few handwritten basic character images from the present database is shown in Fig. 2.

This image database is divided into training and test sets. The training set for Bangla basic characters is composed of 200 samples taken randomly from each of the 50 classes. Thus the total size of the training set for basic characters is 10,000. The remaining 10187 samples form the test set of Bangla basic characters and the minimum number of samples in a class of this test set is 60.

3 Recognition Methodology

3.1 Smoothing and Binarization

The first step of the present recognition scheme is smoothing of the graylevel character image. This is a common preprocessing operation of any character recognition approach for the purpose of removing possible artifacts present in a character image. In the present work, we consider a restricted mean filtering

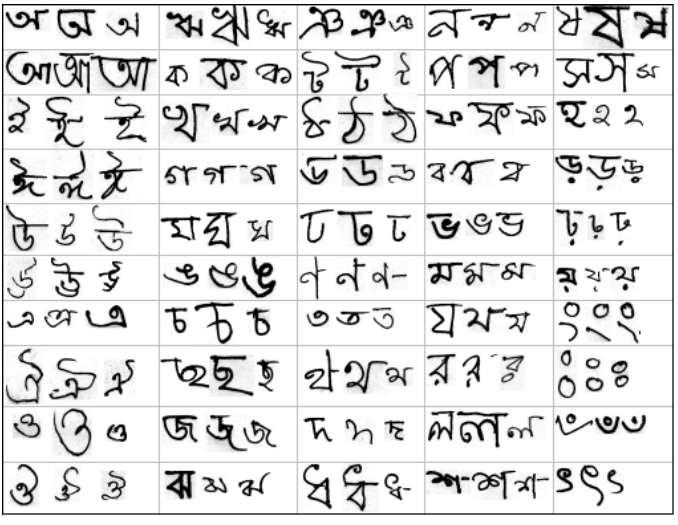


Fig. 2. Samples from the present database of handwritten Bangla characters; three samples in each category are shown

approach for the above purpose. Use of the ordinary mean filter may often connect two disjoint components of a character like ড় causing significant loss in shape information. However, in the restricted mean filter approach, a pixel value is changed by the usual mean provided this does not result in joining two disjoint components existing in the binarized image before smoothing. In Fig. 3, an example of this situation has been shown. After smoothing of the input image it is binarized using Otsu's global thresholding technique [18].

(*a*) (*b*)

Fig. 3. (*a*) Ordinary mean filtering joins two disconnected components; (*b*) Restricted mean filtering does not join originally disconnected components

3.2 Removal of Extra Long Headline

Many Bangla alphabetic characters in their printed/ideal form have a horizontal line (called matra or headline) at the upper part of the symbol; a few characters have a curve-like extension above the headline and several other characters do not have any part above it. Examples of both these cases are shown in Fig. 4.

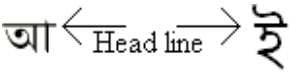
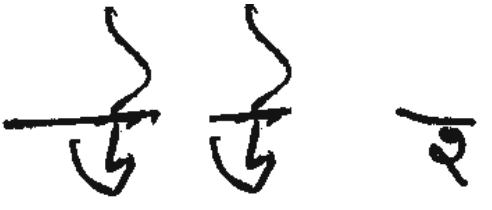


Fig. 4. Headline or matra is shown

Detection of the headline in a handwritten Bangla character is often found crucial whenever it is elongated enough (such as in Fig. 5) increasing the width of the character image substantially. In such situations, normalization of subsequent feature values gets affected. So, before computation of features, headlines are detected using simple heuristics. Example of successful removal of extra long headline according to these heuristics is shown in Fig. 5(a). Also, a situation when our heuristics failed to remove extra long headline is shown in Fig. 5(b).



(a) (b)

Fig. 5. (a) Removal of headline or matra is shown; (b) Headline cannot be removed

(a) (b) (c)

Fig. 6. Shape representations of character image: (a) An input character image (binarized); (b) skeletal representation of the character image; (c) contour representation of the character image

3.3 Feature Extraction

In the present study chain code histogram features of an input character image have been computed from its both the skeletal (thinning as in [19]) and contour representations. The skeletal and contour representations of a character image are shown in Fig. 6. From this example, it is seen that the skeletal representation are often affected by the presence of hair(s) removal of which is usually difficult.

Chaincode Representation of the shape of an input character image is obtained by using Freeman codes [20] while tracing its skeleton or contour. The scheme of Freeman's chain code and the shape representation following this scheme are shown in Fig. 7.

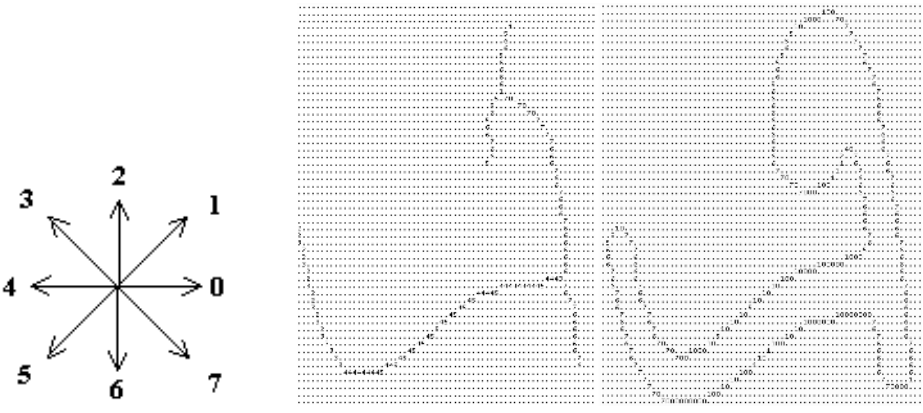


Fig. 7. (*a*) Scheme for chaincodes; (*b*) chain code representation of the shape in Fig. 6(*b*); *c*) chain code representation of the shape in Fig. 6(*c*)

Chaincode Histogram Features [21] are the main features used in the present recognition task and these are obtained as follows. The smallest rectangular frame (bounding box) enclosing the character skeleton or its contour is computed and this is divided into 5 or 7 equal horizontal and vertical strips. If any of the number of rows or columns of the above rectangular frame is not a multiple of 5 or 7, the rightmost vertical or bottommost horizontal strip should have fewer number of rows or columns respectively. Thus, the said frame is divided into 25 or 49 rectangular blocks (Fig. 8) of equal areas save for a few possible extreme blocks with less areas. In each block, a local histogram of the chain codes is calculated. Since the directions along the skeleton or contour should be effectively quantized into one of 4 possible values, *viz.* 0 or 4, 1 or 5, 2 or 6 and 3 or 7, the histogram of each block has four components. The feature vector is composed of these local histograms computed either from the skeletal or contour representation of the input image. Thus, the feature vector has either $4 \times 5 \times 5 = 100$ or $4 \times 7 \times 7 = 196$ components. For size normalization, each component of the feature is divided by the sum of the height and width of the bounding box of the skeleton or contour of input character image depending upon the particular case. In the present study, we also considered feature vector by down sampling the above 7×7 blocks into 4×4 blocks using Gaussian filter.

3.4 Designing Classifier

Multilayer Perceptrons (MLPs) have been chosen as the classifiers of the present study of recognition of handwritten Bangla basic characters. The well known backpropagation (BP) algorithm [22] is used for the training of MLP classifiers. However, in many applications like the present one, the proper training of an MLP largely depends on the choice of the parameter (learning rate and

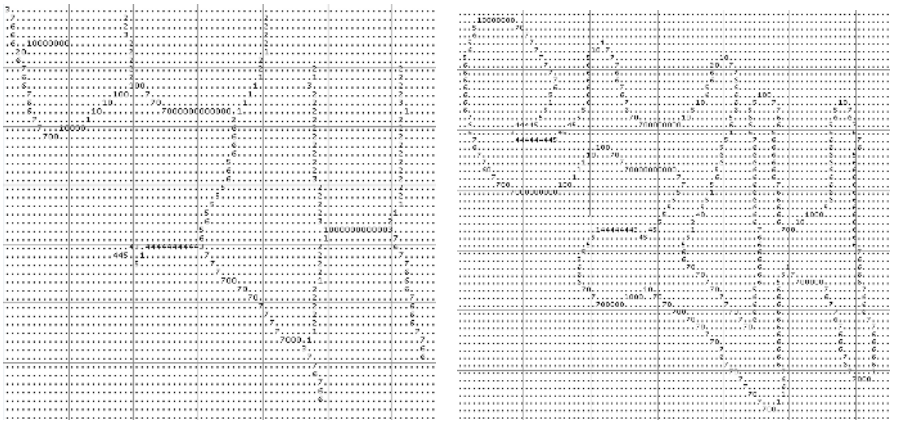


Fig. 8. Division of rectangular frame into 7×7 rectangular zones

momentum factor) values and also it often converges too slowly. There exist a number of modified BP algorithms which take care of these problems of the original BP algorithm. In the present classification task, we considered a modified BP algorithm [23] using self-adaptive learning rate values.

Another issue associated with the use of an MLP classifier is the choice of the size of its hidden layer(s). In fact, it is difficult to get an idea of an optimal size of the hidden layer(s). We experimented with several different choices of hidden layer size in each case and classification results will be reported in the next section corresponding to the best among these choices.

Finally, the strategic selection of the point of termination of the iterative learning of BP algorithm is another important issue. Often a validation set of samples is used to avoid overtraining, and thus a better generalization performance of the network is ensured. Usually, during the initial stages of training of an MLP using BP training algorithm, it gradually decreases the system error [22] on both the training and validation sets. However, after a certain amount of training, this error further decreases on the training set while it starts increasing on the validation set (as shown in Fig.9). The point of time when the error on the validation set increases for at least three consecutive sweeps for the first instance is noted and the weight values before the error starts increasing, are stored.

Since the present database described in the previous section does not exclusively provide any validation set, we have synthetically generated a validation set consisting of 150 samples from each class. These samples have been generated by taking 50 random samples of each class from the training set. These samples are randomly rotated between -10° to $+10^\circ$. Gaussian blurring kernel (standard deviation 2) has been applied on these rotated samples and finally these are binarized using three different threshold values.

An Analysis of Misclassifications after the above first stage of the present recognition scheme, shows that a significant percentage of misclassifications

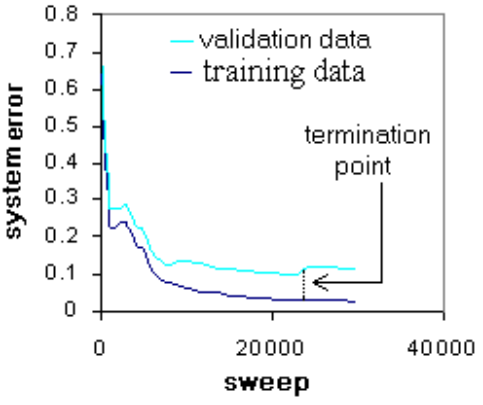


Fig. 9. Increase in error on validation set indicates termination of learning; here, system error is the mean square error between the target and computed output vectors

occurred within several small groups of character classes. In Table 1, these groups of Bangla basic characters corresponding to the recognition results based on chain code histogram features (7×7 blocks down sampled into 4×4 blocks) computed from contour representations of their handwritten shapes are listed. The above situation provided the best recognition performance among our various other choices.

For each of these sub-groups, a distinct MLP classifier with fewer hidden nodes was trained. Samples initially classified as a member of such a group is presented to the relevant smaller MLP architecture for its classification in the second stage. Contour representation based chain code histogram features computed on the 7×7 blocks after their down sampling (with Gaussian filter) to 4×4 blocks had been used in our simulation as the feature vector of the second stage of the present recognition scheme. This second stage of classification, improves the recognition accuracy of the first stage.

In the above context, it is apparent that if an input sample in the first stage of its classification is misclassified in a group other than its own (as shown in Table 1), then the second stage of classification cannot do any help for its correct classification. Only if an input sample is misclassified as a character of its own group, then it gets a second chance of being properly classified by the second stage of classification.

4 Experimental Results

We used six different but related feature sets for the present recognition study of handwritten Bangla basic characters. These are chain code histogram features on 5×5 blocks, 7×7 blocks and histogram on 7×7 blocks down sampled (with Gaussian filter) to 4×4 blocks. Each of the above three sets of features was computed using both skeleton (obtained by thinning) and contour representations of the input character image.

Table 1. Recognition results (best situation) within and outside groups of confusing character shapes after the 1st stage of the present recognition approach

Group No.	Basic characters	Total no. of samples in the Group		Misclassification(%) within group		Misclassification(%) outside the group		Misclassification (%) (Total)	
		Training	Test	Training	Test	Training	Test	Training	Test
1	অ আ ত	600	897	0.45	0.52	0.42	0.41	0.87	0.93
2	ই ঈ	400	323	0.26	0.26	0.28	0.36	0.54	0.62
3	উ ঊ ও ঔ	800	727	0.32	0.50	0.30	0.32	0.62	0.82
4	ক ঝ ধ ফ ব র	1200	1602	0.16	0.31	0.40	0.41	0.56	0.72
5	খ ঘ	400	402	0.70	1.14	0.52	0.76	1.22	1.90
6	গ ন প ল শ	1000	1419	0.25	0.52	0.21	0.33	0.46	0.85
7	ঙ ড ভ ড়	800	586	0.31	0.38	0.47	0.68	0.78	1.06
8	চ ঢ ঢ়	600	693	0.25	0.41	0.33	0.38	0.58	0.79
9	থ য ষ য়	800	974	0.51	0.77	0.51	0.74	1.02	1.51
10	ম ম্	400	414	0.49	0.85	0.39	0.52	0.88	1.37
11	Others (15)	3000	3688	0.14	0.19	0.22	0.29	0.36	0.48
	Total	10000	11725	3.84	5.85	4.05	5.20	7.89	11.05

Table 2. Comparative recognition performance (1st stage) of six feature sets (chain code histogram based) on the present database

Shape representation	Block size	Feature size	No. of hidden nodes	Recognition Accuracy	
				Training	Test
Skeleton	5 X 5	100	55	70.76	68.24
	7 X 7	196	75	77.22	74.65
	7 X 7 down sampled to 4 X 4	64	40	89.42	86.16
Contour	5 X 5	100	55	75.45	71.86
	7 X 7	196	75	80.86	76.12
	7 X 7 down sampled to 4 X 4	64	40	92.11	88.95

Different MLP classifiers were trained using the above six feature vectors. In each case, the near optimal size of the hidden layer was obtained by extensive simulations and these are shown in Table 2. The best results obtained by the first stage of our classification scheme correspond to the feature vector consisting of histogram values on 7×7 blocks computed from the contour representation of an input character followed by its down sampling to 4×4 blocks.

We simulated a second stage of classification for all the eleven groups of characters as shown in Table 1. After the second stage of classification, the final

recognition accuracy corresponding to the best situation of the first stage was 92.14% on the test set and 94.65% on the training set.

5 Conclusions

In the present recognition study, we observed that if chain code histogram features are used for recognition of handwritten basic characters, then acceptable recognition performance may be obtained by computing these features using a division of character contour into 7×7 blocks followed by down sampling the resulting feature components into 4×4 blocks.

Based on the above recognition results, we identified a few groups of characters within which misclassifications are significant. Further classifications within each such group improved the final recognition accuracy.

In future, we plan to study similar recognition performance by using a different feature vector in the second stage for further improvement of the classification accuracy.

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. Bhattacharya, U., Parui, S. K., Sridhar, M., Kimura, F.: Two-stage Recognition of Handwritten Bangla Alphanumeric Characters using Neural Classifiers, CD Proc. IICAI (2005) 1357-1376
4. Bhowmik, T. K., Bhattacharya, U., Parui, S. K.: Recognition of Bangla Handwritten Characters Using an MLP Classifier Based on Stroke Features, Proc. ICONIP, (2004) 814-819.
5. Rahman, F. R., Rahman, R., Fairhurst, M. C.: Recognition of Handwritten Bengali Characters: A Novel Multistage Approach. Pattern Recognition, Vol. 35 (2002) 997-1006
6. Datta, A., Chaudhury, S.: Bengali alpha-numeric character recognition using curvature features. Pattern Recognition, Vol. 26 (1993) 1757-1770.
7. Bhattacharya, U., Chaudhuri, B.B.: Fusion of Combination Rules of an Ensemble of MLP Classifiers for Improved Recognition Accuracy of Handprinted Bangla Numerals, Proc. of ICDAR, Seoul (2005) 322-326.
8. Bhattacharya, U., Das, T. K., Datta, A., Parui, S. K. Chaudhuri, B. B.: A hybrid scheme for handprinted numeral recognition based on a self-organizing network and MLP classifiers. International Journal for Pattern Recognition and Artificial Intelligence, Vol. 16 (2002) 845-864
9. Bhattacharya, U., Das, T. K., Chaudhuri, B. B.: A cascaded scheme for recognition of handprinted numerals. Proceedings of the third Indian Conference on Computer Vision, Graphics and Image Processing, Ahmedabad, India (2002) 137 - 142
10. Pal, U., Chaudhuri, B. B.: Automatic Recognition of unconstrained off-line Bangla hand-written numerals, Advances in Multimodal Interfaces, Springer Verlag Lecture Notes on Computer Science (LNCS-1948), Eds. T. Tan, Y. Shi and W. Gao. (2000) 371-378.

11. Ramakrishnan, K. R., Srinivasan, S. H., Bhagavathy, S. The independent components of characters are 'Strokes', Proc. of the 5th ICDAR (1999) 414-417.
12. Sukhaswami, M. B., Seetharamulu, P., Pujari, A. K.: Recognition of Telugu characters using neural networks, Int. J. Neural Syst., Vol. 6 (1995) 317-357.
13. Suresh, R. M., Ganesan, L.: Recognition of Printed and Handwritten Tamil Characters Using Fuzzy Approach, Proc. of Sixth ICCIMA'05 (2005) 286-291.
14. Mohanti, S.: Pattern recognition in alphabets of Oriya Language using Kohonen Neural Network, IJPRAI, vol. 12 (1998) 1007-1015.
15. Plamondon, R., Srihari, S. N.: On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey. IEEE Trans. Patt. Anal. and Mach. Intell., Vol. 22 (2000) 63-84
16. 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
17. Trier, O. D., Jain, A. K. and Taxt, T.: Feature Extraction Methods for Character Recognition - A Survey. Pattern Recognition, Vol. 29 (1996) 641 - 662
18. Otsu, N.: A Threshold Selection Method from Grey-Level Histograms. IEEE Trans. Systems, Man, and Cybernetics, Vol. 9 (1979) 377-393
19. Datta, A., Parui, S. K.: A robust parallel thinning algorithm for binary images. Pattern Recognition, Vol. 27 (1994) 1181-1192
20. Freeman, H.: Computer processing of Line-drawing Images. ACM Computing Surveys, Vol. 6 (1974) 57-97
21. Kimura, F., Miyake, Y., Sridhar, M.: Handwritten ZIP Code Recognition using Lexicon Free word Recognition Algorithm. Proc. Int. Conf. Document Analysis and Recognition, Vol. II, Motreal, Canada (1995) 906 - 910
22. Rumelhart, D. E., Hinton, G. E., Williams, R. J. Learning internal representations by error propagation. Institute for Cognitive Science Report 8506, San Diego: University of California (1985)
23. Bhattacharya, U., Parui, S. K.: Self-adaptive learning rates in backpropagation algorithm improve its function approximation performance. Proc. of the IEEE International Conference on Neural Networks, Australia (1995) 2784-2788