Effect of Writer Information on Bangla Handwritten Character Recognition

Chayan Halder
Department of Computer Science
West Bengal State University, Barasat,
Kolkata 700126, WB, India
Email: chayan.halderz@gmail.com

Sk Md Obaidullah
Department of Computer Science & Engineering
Aliah University
Kolkata, WB, India
Email: sk.obaidullah@gmail.com

Kaushik Roy
Department of Computer Science
West Bengal State University, Barasat,
Kolkata 700126, WB, India
Email: kaushik.mrg@gmail.com

***Abstract*— Handwritten character recognition has various potential in the field of document image processing. It is one of the important aspects for systems like handwritten optical character recognizer, writer identification/verification, automatic document sorter etc. In Bangla only few attempts are made towards character recognition. In this current study a relatively new attempt is made towards finding the dependency of writer information on character recognition by varying the inputs. This study will provide a better understanding of the input data for character recognition. Also it will help to know the Bangla characters better for writer identification/verification. Here, highest accuracy of 100% is achieved in case of numeral 7 applying LibSVM classifier.**

***Keywords*—Bangla Character Recognition, LibSVM, LibLINEAR, Writer Dependency, Histogram gradient based feature**

I. Introduction

The analysis of handwritten documents is one of the most challenging research task in the area of computer vision and pattern recognition. Since mid 1950's, research on Optical Character Recognition(OCR) is very active and gets most of the attention from the researchers [1]. However, recognition of printed characters are lot easier than the handwritten ones due to the enormous variability in handwriting styles of different persons. Though it is unfortunate that the approaches of printed character recognition can not be directly usable for handwritten character recognition.

In India, Bangla is the second most popular script next to Devanagari and used to represent three languages *viz.*, Bangla, Assamese and Manipuri. It is also the national language of Bangladesh [2]. So, working with Bangla has lots of significance in this respect. But despite its importance very limited works are found on Bangla character recognition as the variation of writing and complex nature of Bangla characters make it more challenging. The Bangla script consists of 50 alphabets (39 consonants + 11 vowels) and 10 numerals. The numbers of alphabets are quite higher compared to Roman script. Also the characters (alphabets + numerals) are structurally complex than the Roman script. The sample view of the complex nature of these characters can be found in Fig. I. Here from the figure differences in printed and handwritten characters are visible. Fig. I (a-c) shows the printed characters and Fig. I (d-f) shows the handwritten characters.

Various works are available on printed character recognition which is evident in [3]. A complete OCR for printed Bangla can be found in [4]. Several off-line handwritten numeral and character recognition works on Indic scripts are available in literature like the work of Bajaj et al., Sharma et al., Bhattacharya et al., Pal et al. on Devanagari script in [5]–[8] and on Kannada numerals in [9] by Sharma et al. Pal et al. have worked on six popular Indian scripts numeral recognition which include Devanagari, Bangla, Telugu, Oriya, Kannada and Tamil scripts [10]. Works on Bangla character recognition can be found in [11], [12]. Having these various pieces of works in Bangla no attempt still made towards finding the effect of writer information on Bangla handwritten character recognition by varying the number of writers. This study not only helps to understand and improve the character recognition in Bangla but also helps to improve the performance of writer identification/verification by distinguishing the characters according to their performance.

In this current study isolated character recognition on Bangla is presented by varying the number of samples from different writers. A database of 45000 characters (37500 alphabets + 7500 numerals) from 150 different writers are considered. The histogram gradient based feature is considered for feature extraction.

The rest of the paper is organized as follows: in Section II, details about the data used and proposed method can be found. Section III presents the results and analysis of the experiment. Conclusion and future scopes are discussed in Section IV.

II. Database and Methodology

A. Data

As stated earlier a database of Bangla isolated characters from 150 writers is considered for the present work. The main concentration of the recognition strategy is on simple characters due to the fact that in a survey on standard Bangla text it is found that almost 97.55% of characters are simple characters. There is no restriction on the type of pen and ink the writers have used. The collected data are scanned and stored in 300dpi gray mode. The database contains 750 samples for each character. A simple histogram based preprocessing technique is used to extract and store the isolated characters in gray format.

978-1-4673-8564-0/15/$31.00

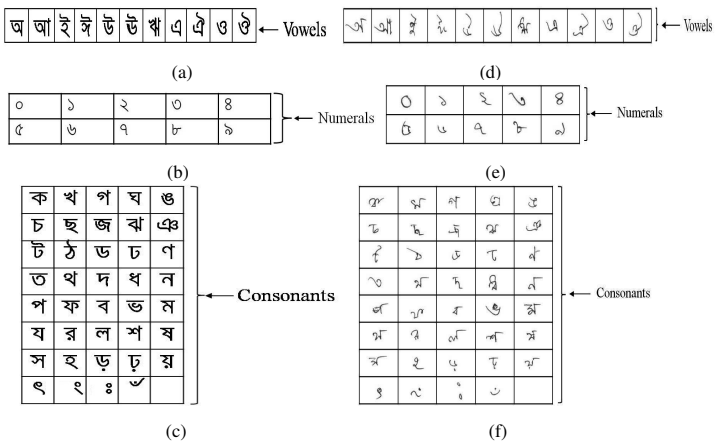


Fig. 1. (a-c) Sample Bangla Printed Characters and (e-f) Sample Handwritten Bangla Characters

B. Experimental Design

The current problem of finding the effect of character recognition due to the variation in writers and writing samples has some challenges. For this present work 14 different experimental scenarios are considered.

1) 750 samples for each characters from 150 writers
2) 500 samples for each characters from 100 writers.
3) 250 samples for each characters from 50 writers.
4) 450 samples for each characters from 150 writers.
5) 300 samples for each characters from 100 writers.
6) 150 samples for each characters from 50 writers.
7) 150 samples for each characters from 150 writers.
8) 100 samples for each characters from 100 writers.
9) 50 samples for each characters from 50 writers.
10) 300 samples for each characters from 20 writers.
11) 200 samples for each characters from 20 writers.
12) 100 samples for each characters from 20 writers.
13) 60 samples for each characters from 20 writers.
14) 20 samples for each characters from 20 writers.

These samples and writers other than the first case are chosen randomly from the pool of all samples from 150 writers.

After selection of data for each case histogram gradient based features is extracted and LibLINEAR and LibSVM classifiers are used for character recognition of these different cases.

1) Feature Extraction: The current experiment is conducted on multi-resolution images of isolated characters by extracting histogram gradient based feature ($f(400)$).

a) Histogram gradient based features ($f(400)$): Algorithm:

Input: Character gray images
Output: Feature set of 400 dimension [$f(d)_1$ to $f(d)_{400}$]
Step1: Binirization and normalization of the input gray image into 73 x 73 pixels.
Step2: Conversion to gray-scale image by applying 2 x 2 mean filter 5 times and then normalization. Now, segment the image to 9x9 blocks.
Step3: Roberts filter is applied to obtain gradient image. The arc tangent of the gradient is quantised into 16 directions and accumulated with each of the quantized direction. The arc tangent is calculated using equation (1).

$$f(x,y) = \sqrt{(\Delta u)^2 + (\Delta v)^2} \tag{1}$$

The direction of gradient is calculated using equation (2).

$$\theta(x,y) = tan^{-1}\Delta v/\Delta u \tag{2}$$

where
$\Delta u = (x+1, y+1) - (x, y)$ and $\Delta v = (x+1, y) - (x, y+1)$
Step4: Now, 9x9 blocks are down sampled to 5x5 by Gaussian filter and histogram of 16 quantized direction values are computed for each block to get 5x5x16 = 400 dimensional feature.
For more details in this see [13].

2) Classification: During this experiment the main goal is to find the effect of character recognition due to variation in the input data set. The LibLINEAR and LibSVM classifiers are chosen to compare and analyse the variations more accurately. The classifiers are chosen over MLP due to the higher computational cost and the problem of over fitting. The total dataset of each character is divided into randomly selected training and testing set in a ratio of 4:1 (5-fold cross-validation) for each experimental case to reduce biased classification results.

a) LibSVM: The Support Vector Machine (SVM) is basically a linear classifier. The LibSVM supports many kernel transformations to turn a non linear problem into linear problem. In LibSVM, the SMO-type algorithm is used which works both for kernel and linear SVM. Though the computational complexity both in terms of space and time is higher compared to LibLINEAR but it produces better results than LibLINEAR for the current scheme. Here C-SVC type SVM is used with Radial basis kernel, having 1 as cost parameter. Details of LibSVM can be found in [14].

b) LibLINEAR: The main drawback of LibSVM is that it has a complexity of $O(n^2)$ or $O(n^3)$ due to the kernel transformations when the problem is not bi-class. On the other hand the LibLINEAR is implemented using the SVM but with some modification like it does not support kernel transformation so the complexity become $O(n)$ (n is the number of samples in the training set). The LibLINEAR is suitable for most cases when the amount of data with instances or features to be classified is large. The SVM Type parameter L2-Loss Support Vector Machine (dual) is used for the current work. Details of LibLINEAR can be found in [15].

III. Result and Analysis

In this section the experimental performance of the proposed method is analysed and discussed to validate the hypothesis that identification of characters are very much dependent on the number of writers and the samples per writer. By varying the number of input data set as stated earlier, it is found that there is some variations in character recognition rate when the number of writers and the number of samples per writer varies. Section III-A discusses about the variation of results due to variation in number of writers. In section III-B the variation in recognition accuracy due to variation in samples per writer is presented.

A. Experimentation with Writer Variation

This section presents the experimental performance that are conducted to verify the hypothesis that character recognition rate is dependent upon the variation of the number of writers. A series of experimental scenarios are applied by using 150, 100, 50, 20 writers for samples 5, 3, 1 per writer to create different train and test samples. Table I, II and III show the variation of character recognition results for different writers having all the samples per writer are same. Here the highest recognition result along with average and lowest accuracies are considered. The observation of the tables reveal that for all different samples average accuracy gradually decreases depending on the reduction of writers and after 50 writers the accuracy increases due to the fact that total number of instances are quite low. For all the scenarios the LibSVM classifier performed better than the LibLINEAR classifier. Top average accuracy of 85.30% and 76.56% is scored by 20 writers when samples are set to 5 each for LibSVM and LibLINEAR classifiers respectively.

TABLE I. RECOGNITION RESULTS FOR VARIATION OF WRITERS HAVING 5 SAMPLES PER WRITER

Writer (Set 5)	LibLINEAR			LibSVM		
	Highest (%)	Average (%)	Lowest (%)	Highest (%)	Average (%)	Lowest (%)
150 (5)	89.87	71.77	23.65	95.51	82.17	56.57
100 (5)	91.42	70.30	38.45	95.40	80.71	51.35
50 (5)	89.96	69.17	28.51	95.58	79.31	45.16
20 (5)	97.78	76.56	46.67	100.00	85.30	57.78

TABLE II. RECOGNITION RESULTS FOR VARIATION OF WRITERS HAVING 3 SAMPLES PER WRITER

Writer (Set 3)	LibLINEAR			LibSVM		
	Highest (%)	Average (%)	Lowest (%)	Highest (%)	Average (%)	Lowest (%)
150 (3)	91.62	69.56	38.06	94.51	80.43	52.53
100 (3)	90.46	68.40	28.52	94.35	79.39	47.08
50 (3)	90.00	66.55	37.58	95.33	76.75	43.62
20 (3)	96.30	73.70	39.62	100.00	81.07	50.00

TABLE III. RECOGNITION RESULTS FOR VARIATION OF WRITERS HAVING 1 SAMPLES PER WRITER

Writer (Set 1)	LibLINEAR			LibSVM		
	Highest (%)	Average (%)	Lowest (%)	Highest (%)	Average (%)	Lowest (%)
150 (1)	86.05	60.54	31.97	91.86	74.72	41.84
100 (1)	93.02	63.41	31.96	94.90	74.03	35.05
50 (1)	84.00	56.21	22.00	92.00	66.57	18.37
20 (1)	95.00	64.64	10.00	95.00	67.40	16.67

B. Experimentation with Samples per Writer Variation

Here another perspective of the experimental scenarios where the number of samples per writers are varied but writers are constant is discussed. For this experiment samples per writers are varied from 15, 10 samples (20 writers only) to 5, 3, 1 sample(s) (150, 100, 50, 20 writers). Fig. 2 (a-d) shows the variation of character recognition results for the current scenarios. Naturally, when the samples are decreased regardless of the total writers, the accuracy also decreases for both LibLINEAR and LibSVM classifiers. Also it is found that when number of samples per writer increases the accuracy gets increased which is evident in various scenarios. But further inspection shows that for some scenarios like in Fig. 2 (b) in case of LibLINEAR classifier with lowest accuracy when the number of sets are reduced from 5 to 3 for 150 writers the accuracy gets increased. The same can be seen in Fig. 2 (d) also. Apart from these it can also be observed that when number of writers is small then the previous assumption of monotonically increasing and decreasing of accuracies depending on the increment and decrement in set numbers respectively, differs mainly for top accuracies. Like for 20 writers when samples are 15, 5 and 3 the top accuracy is constant at 100%.

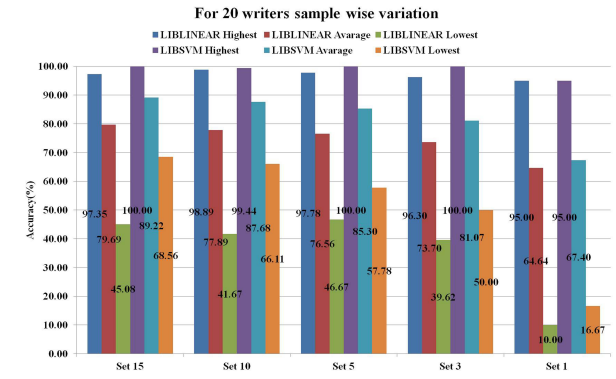
C. Error Analysis

In this current study it is found that the average accuracy is not as much high as in the existing literature. But it should be noted that no attempt is made to improve the quality of the unconstrained characters during the preprocessing stage which is the major reason behind this. Raw gray images of unconstrained Bangla isolated characters are extracted and histogram gradient based 400 dimensional feature is applied. Furthermore alphabets like *MORDHANNA* (ণ) and *DONTANNA* (ন); numeral *TIN* (৩) and alphabet *TA* (ত); numeral *TUI* (২) and alphabet *HA*(হ) etc. are almost structurally same in case of handwriting. These contributes to the low recognition rate of the current study. These similarities can be seen in the Fig. 3.

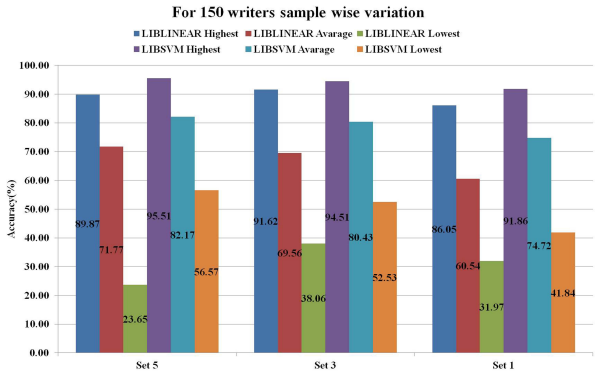
IV. CONCLUSION AND FUTURE SCOPE

The objective of this research is to find the impact of variation in writers and samples per writer for character recognition in Bangla. Total 14 different experimental scenarios are considered on a database of 150 writers to extensively study the current research. To the best of our knowledge, this is the first study of its type attempted till date. A state-of-the-art feature and well established classification techniques are used to verify the hypothesis of writer dependency for character recognition. In context to the existing literature the accuracy is not very high, but it should be considered that no extensive measures are taken to improve the quality of the characters. Study has revealed that reduction in writer numbers do reduces the accuracy accordingly but when samples per writer are increased the accuracy is also increased. Further comparative study on classification techniques shows that the LibSVM classifier performs better than LIBLINEAR for Bangla character recognition. The study also helps to understand the underlying distinguishing features of Bangla characters which intern will help in character recognition and writer identification/verification.

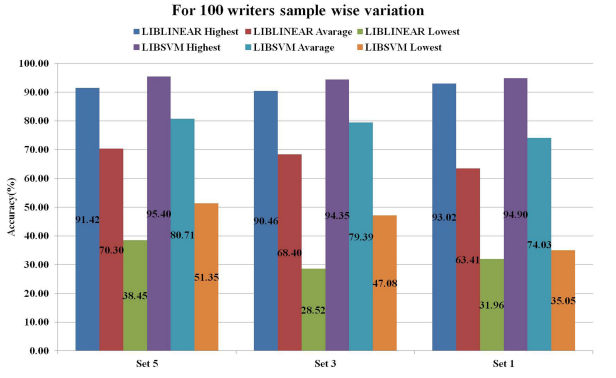
For further extension of this work more feature extraction techniques and combination of features can be introduced to improve the accuracy and also different classifiers can be introduced. Improvement in total writers specially for single sample scenario can be attempted in future. Furthermore a similar study can be made in future to find out whether effect of writer information on character recognition affect the writer identification and verification accuracy or not. Also the same can be applied on other Indic scripts and we are looking forward to these experiments.



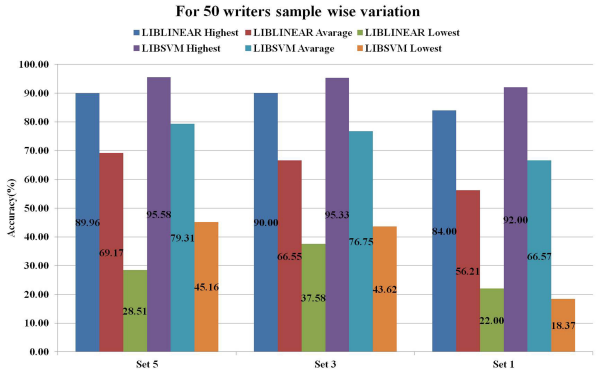
(a)



(b)



(c)



(d)

Fig. 2. Recognition results for variation of samples per writer

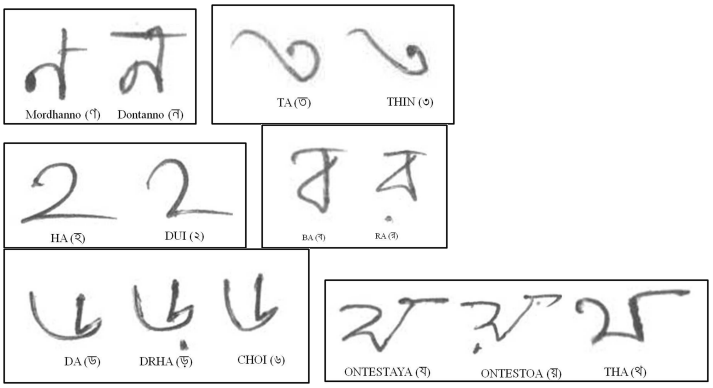


Fig. 3. Some handwritten characters with close structural similarity

ACKNOWLEDGEMENT

One of the authors would like to thank Department of Science and Technology for the support of INSPIRE fellowship.

REFERENCES

[1] O.D. Trier, A.K. Jain, T. Taxt, "Feature Extraction Methods for Character Recognition - A Survey," Pattern Recognition, Vol. 29(4), pp. 641-662, 1996.

[2] U. Bhattacharya, M. Shridhar and S.K. Parui, "On Recognition of Handwritten Bangla Characters," In Proc. of ICVGIP, pp. 817-828, 2006.

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

[4] B. B. Chaudhuri and U. Pal, "A complete printed Bangla OCR system," Pattern Recognition, Vol. 31, pp. 531-549, 1998.

[5] R. Bajaj, L. Dey and S. Chaudhury, "Devnagari numeral recognition by combining decision of multiple connectionist classifiers," Sadhana, Vol. 27(1), pp. 59-72, 2002.

[6] N. Sharma, U. Pal, F. Kimura and S. Pal, "Recognition of Off-Line Handwritten Devnagari Characters Using Quadratic Classifier," In Proc. of ICVGIP, pp. 805-816, 2006.

[7] U. Bhattacharya, S. K. Parui, B. Shaw and K. Bhattacharya, "Neural Combination of ANN and HMM for Handwritten Devnagari Numeral Recognition," In Proc. of 10^{th} IWFHR, pp. 613-618, 2006.

[8] U. Pal, T. Wakabayashi and F. Kimura, "Comparative Study of Devnagari Handwritten Character Recognition using Different Feature and Classifiers," In Proc. of 10^{th} ICDAR, pp. 1111-1115, 2009.

[9] N. Sharma, U. Pal and F. Kimura, Recognition of handwritten Kannada numerals, In Proc. of 9^{th} ICIT, pp. 133-136, 2006.

[10] U. Pal, T. Wakabayashi, N. Sharma1 and F. Kimura, "Handwritten Numeral Recognition of Six Popular Indian Scripts," In Proc. of 9^{th} ICDAR, pp. 749-753, 2007.

[11] K. Roy, U. Pal and F. Kimura, "Recognition of Handwritten Bangla Characters," In Proc. of 2^{nd} ICMI, pp. 480-485, 2005.

[12] N. Das, B. Das, R. Sarkar, S. Basu, M. Kundu and M. Nasipuri, "Handwritten Bangla Basic and Compound character recognition using MLP and SVM classifier," Journal of Computing, Vol. 2(2), pp. 109-115, 2010.

[13] K. Roy and U. Pal, "On the Development of an OCR System for Indian Postal Automation," LAP LAMBERT Academic Publishing, Germany, ISBN: 978-38-443-1403-8, 2011.

[14] C. -C. Chang, C. -J. Lin, "LIBSVM: A library for support vector machines", ACM Transactions on Intelligent Systems and Technology, Vol. 2(3), pp. 1-27, 2011.

[15] R. -E. Fan, K. -W. Chang, C. -J. Hsieh, X. R. Wang, and C. J. Lin, "Liblinear: A library for large linear classification", Journal of Machine Learning Research, Vol. 9, pp. 1871-1874, 2008.