

Text Independent Writer Identification for Bengali Script

Sukalpa Chanda, Katrin Franke
*Department of Computer Science
 and Media Technology,
 Gjøvik University College,
 Norway*

E-mail: -{sukalpa, kyfranke}@ieee.org

Umapada Pal
*Computer Vision and Pattern
 Recognition Unit,
 Indian Statistical Institute,
 India*

E-mail: -umapada@isical.ac.in

Tetsushi Wakabayashi
*Graduate School of
 Engineering,
 Mie University,
 Japan*

E-mail: -waka@hi.info.mie-u.ac.jp

Abstract-Automatic identification of an individual based on his/her handwriting characteristics is an important forensic tool. In a computational forensic scenario, presence of huge amount of text/information in a questioned document cannot be always ensured. Also, compromising in terms of systems reliability under such situation is not desirable. We here propose a system to encounter such adverse situation in the context of Bengali script. Experiments with discrete directional feature and gradient feature are reported here, along with Support Vector Machine (SVM) as classifier. We got promising results of 95.19% writer identification accuracy at first top choice and 99.03% when considering first three top choices.

Keywords- *Text independent writer identification; Bengali script ;Document Analysis; Computational forensics.*

I. INTRODUCTION

Writer identification is a vibrant field of research due to its scope of application as a computational forensic method. There are many pieces of work on writer identification [1-4, 6-10, 12]. Said et al. [8] developed a writer identification system which is text independent, they took a texture analysis based approach. Schomaker and Bulacu [10] proposed an offline writer identification system, using connected-component contours in uppercase handwritten samples. Later Bulacu and Schomaker [6] proposed texture level and allograph level feature based writer identification scheme. Srihari et al. [1] have used a combination of global and local features. Based on analysis of ink texture for each writer, Franke et al.[3] proposed a system for writer ink type identification. Though a large number of people in the world use Indic scripts, to the best of our knowledge, there is only one work on Indic script [12] in the context of writer identification and they proposed an AR co-efficient feature-based writer identification system for 40 Bengali writers. They have used at least 200 words per writer for training and testing their system. But, very often a questioned document is deprived of such huge number of handwritten text words. Hence, to analyze a questioned

document of Indic script with lesser amount of information, a reliable writer identification system is in demand. In order to encounter adversaries like scarcity of data content in questioned documents of Bengali script, we here propose a robust writer identification system.

II. LINE AND CHARACTER SEGMENTATION

For line segmentation, at first, we divide the text into vertical stripes. Stripe width of a document is computed by statistical analysis of text height in the document [13]. Each of those stripes is processed to form Piece Wise Separating Lines (PSL) [13], and joining those PSL's we segmented the text lines. A histogram based approach was used to segment words in each text line. Since most of the characters in a word in Bengali script are connected through headline, for character segmentation we first find individual component and compute their background portion using water reservoir principle [13]. Based on the water reservoir area, the touching characters in a word are segmented into individual character/character allograph. For details about line, word and character segmentation see [13].

III. FEATURE EXTRACTION- DIRECTIONAL FEATURES AND GRADIENT FEATURES

Character allographs of a writer are quite different from other, even when they write same text. Our directional features are good local shape descriptors and hence they are capable of expressing this character allograph level dissimilarity present in the handwritings of different writers. Our gradient feature gives more information in terms of dissimilarity between character allograph of different writers. Dimension of our directional feature and gradient feature are 64 and 400, respectively.

A. Feature computation for directional features

At first we compute the bounding box of a character component. This bounding box is then divided into 7x7 blocks [11]. In each of these blocks the direction chain code for each contour point is noted and frequency of direction codes is computed. Here we use chain code of four directions only: directions 1 (horizontal); 2 (45 degree slanted); 3 (vertical) and 4 (135 degree slanted).

See Fig. 1 for illustration of four chain-code directions. We assume chain code of direction 1 and 5 are same. Also, we assume direction 2 and 6, 3 and 7, 4 and 8 are equivalent, because if we traverse from point 4 to point 8 we will have the same count as point 8 to point 4. Subsequently in each block, we get an array of four integer values representing the frequencies of chain code in these four directions. These frequencies are used as feature. Thus, for 7×7 blocks we get $7 \times 7 \times 4 = 196$ features. In order to reduce the feature dimension, after the histogram calculation in 7×7 blocks, the blocks are down sampled with Gaussian filter into 4×4 blocks. As a result we obtain $4 \times 4 \times 4 = 64$ features for further classification. To normalize the features we determine the maximum value of the histograms from all the blocks and divide each of the above features by this maximum value to get the feature values between 0 and 1. See [11] to get details about this feature.

B. Feature computation for gradient feature

To obtain 400-dimensional gradient features we apply the following steps. (i) The input binary image of each character allograph is converted into a gray-scale image applying a 2×2 mean filtering 5 times. (ii) The gray-scale image is normalized so that the mean gray scale becomes zero with maximum value 1. (iii) Normalized image is then segmented into 9×9 blocks. (iv) A Roberts filter is applied on the image to obtain gradient image. The arc tangent of the gradient (direction of gradient) is quantized into 16 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. (v) Histograms of the values of 16 quantized directions are computed in each of 9×9 blocks. (vi) 9×9 blocks is down sampled into 5×5 by a Gaussian filter. Thus, we get $5 \times 5 \times 16 = 400$ dimensional feature.

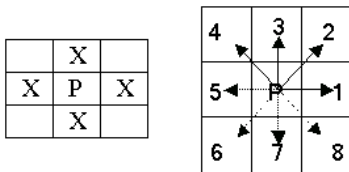


Figure 1. For a point “P” the direction code of its eight neighboring points is shown.

IV. DATASET DETAILS AND EXPERIMENTAL DESIGN

Our dataset consists of two sets of handwriting from each of 104 writers. One set (training set) consist of exactly same piece of text from all writers. The other set (testing set) contains different text with varied number of words, from each writer. (Our training set comprises of 53 Bengali words, and in average our testing dataset consists of 50-60 words per writer). Both of the training and testing data set were scanned to 300 dpi in tiff file format. The printed text used by all writers for writing their respective training files are shown in Fig.2. Character allograph obtained from respective training file of a writer are used in training our classifier for that particular writer.

We mainly performed our experiment to prove the following: (i) Reliability of our proposed scheme when dealing questioned document with relatively low data content. (ii) Robustness of our features to express discriminating characteristics of each individual writer. (iii) How system accuracy is affected when first two top choices and first three top choices are considered instead of only the top choice.

নিজস্ব প্রতিনিধি, কলকাতা: এ রাজ্যে ঘন ঘন বন্ধ নিয়ে শিল্পমহল যতই উদ্বেগ প্রকাশ করুক না এখনও নিজেদের অবস্থানেই অনড় শ্রমিক সংগঠনগুলি। ঘন ঘন বন্ধ রাজ্যের শিল্পের ব্যাপক ক্ষতি করছে বলে সোমবার ইন্ডিয়ান চেম্বার অব কমার্স (আই সি সি)-এর এক সমীক্ষা রিপোর্ট প্রকাশ অনুষ্ঠানে মন্তব্য করেছিলেন আই সি সি প্রেসিডেন্ট হর্ষ কুমার বা।

Figure 2. The text portion used in training file for all writers.

V. CLASSIFIER

We choose Support Vector Machine (SVM) as our classifier. The SVM looks for the optimal hyper-plane which maximizes the distance, the *margin*, between the nearest examples of both classes, named *support vectors* (SVs). In our experiments Gaussian kernel SVM outperformed other non-linear SVM kernels and linear SVM as well, hence we are reporting our recognition results based on Gaussian kernel only. The Gaussian kernel is of the form:

$$[k(x, y) = \exp(-\frac{\|x - y\|^2}{2\sigma^2})].$$

As mentioned earlier, in our experiment for 104 different writers, only about 53 words per writer were used for training. We got best optimized results when gamma parameter ($1/2\sigma^2$) is set to 36.00 and 48.00 for directional feature and gradient feature, respectively. The penalty multiplier parameter is set to 20 for both feature type. Details of SVM can be found in [14] [15].

VI. RESULT AND DISCUSSIONS

A. Writer identification accuracy

Here we show the writer identification accuracy of our scheme, after implementing majority voting technique for all character allograph present in a test image. In Table 1, we report accuracy of our system while using directional feature and gradient feature respectively. For evaluating each test image we did the following: (i) Let in a test image we get 80 character allograph by applying the method as discussed in Section 2. (ii) We extract features from each of them and pass it to the classifier. (iii) The classifier decides the writer for each character allograph. (iv) Majority voting is performed amongst all classified character allograph. (v) Now, if amongst those 80 character allograph, writer 1 gets highest number of character allograph in its favor, we say that test image is written by writer 1. In case of a tie in majority voting we consider that as misclassification.

Results on Directional feature: In our experiment with directional features, there were 6 miss-classifications. We were unable to correctly identify the writer for test images from writer 41, 42, 49, 56, 71, and 83. For test image 41, 42 and 56 we noticed that the second top choice was the original writer whereas for test image 71 the third top choice was the original writer.

Results on Gradient feature: In our experiment with gradient feature, there were 5 miss-classifications. We were unable to correctly identify the writer for test images from writer 21, 41, 42, 56, and 71. Here also, for test image 41, 42 and 56 we noticed that the second top choice was the original writer whereas for test image 71 the third top choice was the original writer. We obtained an accuracy of 94.23% and 95.19% for directional and gradient features, respectively, even when training of our classifier was done with only 53 words per writer. This proves the reliability of our proposed system.

Table I: Writer identification accuracy on 104 test images.

Feature used	Correctly identified	Misclassified
Directional	94.23%	5.77%
Gradient	95.19%	4.81%

B. Distribution of percentage of identified character allograph amongst top two choices

Here we analyze the distribution of percentage of identified character allograph, among top two candidates of majority voting. This is done to give an idea about the

differences between top two choices. To illustrate, say in a test image there are 100 character allographs. Now suppose our classifier model assigns highest number of character allograph (60) to writer ‘‘A’’ and second highest (20) to writer ‘‘B’’. So we can see that 60 % of the total character allograph was assigned to the top choice. The very next second choice is having only 20% of the total character allograph, which is way behind the top choice. This distribution of character allograph between top two choices, are shown with the help of two different curves for all 104 test images. We noticed that for all true hit points, the difference between two curves is quite big. But for all false hit point, we observed very small difference between two curves. By a true (false) hit point, we mean to say the test image with correctly (wrongly) identified writer. From this phenomenon it is clearly evident that for all true hit point there is a big difference of percentage of votes, between top two choices. See Fig.3 and Fig. 4 for a pictorial illustration of such distribution for directional and gradient feature. Please note that in most of the false hit points the difference between two curves is much less.

It can be easily noted that the difference of two curves in Fig.4 is much more than the difference of curves in Fig.3. Hence, we can conclude that our gradient based feature are more robust compared to our directional feature.

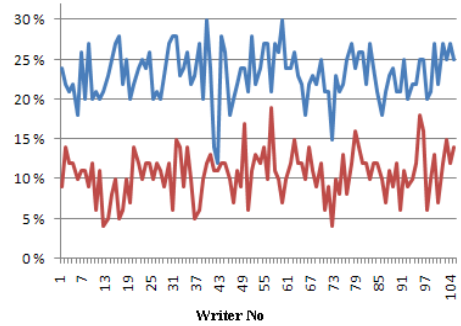


Figure 3. Distribution of percentage of character allograph between top two choices for directional feature. (Top curve-first choice), (Bottom curve-second choice).

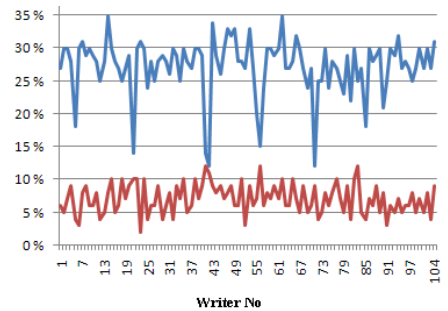


Figure 4. Distribution of percentage of character allograph between top two choices for gradient feature. (Top curve-first choice), (Bottom curve-second choice).

C. Writer Identification accuracy on different top choices

Here we report the accuracy of our writer identification scheme when considering different choices of majority voting. It can be noted that we achieved 99.03% accuracy with gradient feature, when we consider top three choices of our majority voting instead of the top one.

Table II: Writer identification accuracy on different number of top choices of SVM.

No. of Top Choice	Directional Feature	Gradient Feature
1	94.23%	95.19%
2	97.11%	98.07%
3	98.07%	99.03%

D. Error Analysis

In Table 3, we report those writers whose test images are misclassified. We also mention corresponding features responsible for those misclassification. In the Table, ‘‘S’’ signifies that the particular writer was successfully identified by corresponding feature. We noticed that for most of those images there was a marginal win for the erroneous top choice. In most of those cases, after majority voting of character allograph, the original writer were either in the 2nd position or 3rd with very little difference from the top choice. We analyzed the reason and found that sometimes character allographs from two different classes were visually very similar.

Table III: List of misclassified writers.

Original Writer	Identified writer	
	Directional	Gradient
21	S	46
41	37	37
42	38	38
49	75	S
56	84	33
71	41	100
83	90	S

E. Comparison with similar other works

Though there are many pieces of work on writer identification for non Indic scripts, only one work [12] has been reported in the context of Indic script. Garain and Paquet [12] developed a writer identification system and evaluated their scheme on Roman and Bengali script. For Bengali script, they used a dataset of 40 writers, where each writer contributed two samples. One sample was used for training and other for testing. On an average, number of words in each of their sample was 200 or more. On Bengali script, they got 75% accuracy on first top choice amongst 40 writers. From our scheme, we obtained 95.19% accuracy when 104 writers are considered and number of word in each sample is much less (only about

50-60 words) compared to that of the number of words in samples of [12] (200 or more words in each sample).

VII. CONCLUSION

Here we propose a system for Bengali text independent writer identification using directional chain-code and gradient-based features. From the experiment on 104 writers we got promising results of 95.19% writer identification accuracy. We did not impose any rejection criteria in our classifier. We plan to do so in our future work.

REFERENCES

- [1] S. N. Srihari, M. Beal, K. Bandi, V. Shah and P. Krishnamurthy, ‘‘A Statistical Model for Writer Verification’’, In Proc. 8th International Conf. on Document Analysis and Recognition, pp. 1105-1109, 2005.
- [2] Lambert Schomaker, Katrin Franke, Marius Bulacu, ‘‘Using codebooks of fragmented connected-component contours in forensic and historic writer identification’’, Pattern Recognition Letters 28(6), 2007, pp. 719-727.
- [3] K. Franke, O. Bünemeyer, and T. Sy, ‘‘Writer identification using ink texture analysis’’, In Proc. 8th Workshop on Frontiers of Handwriting Recognition, 2002, pp. 268–273.
- [4] A. Schlapbach and H. Bunke, ‘‘Using HMM-Based Recognizers for Writer Identification and Verification’’, In Proc. 9th International Workshop on Frontiers of Handwriting Recognition, 2004, pp.167-172.
- [5] M. Shi, Y. Fujisawa, T. Wakabayashi, and F. Kimura, ‘‘Handwritten numeral recognition using gradient and curvature of grayscale image’’, Pattern Recognition, Vol.35, 2000, pp.2051-2059.
- [6] M. Bulacu and L. Schomaker, ‘‘Text-Independent Writer Identification and Verification Using Textural and Allographic Features’’ IEEE Trans. on PAMI, vol. 29, 2007, pp. 701-718.
- [7] U.-V. Marti, R. Messerli, and H. Bunke, ‘‘Writer Identification Using Text Line Based Features,’’ In Proc. 6th International Conf. on Document Analysis and Recognition, 2001, pp. 101-105.
- [8] H. Said, T. Tan, and K. Baker, ‘‘Personal Identification Based on Handwriting,’’ Pattern Recognition, vol. 33, no. 1, 2000, pp. 149-160.
- [9] S. Srihari, S. Cha, H. Arora, and S. Lee, ‘‘Individuality of Handwriting,’’ J. Forensic Sciences, vol. 47, no. 4, 2002, pp. 1-17.
- [10] L. Schomaker and M. Bulacu, ‘‘Automatic Writer Identification Using Connected-Component Contours and Edge-Based Features of Uppercase Western Script,’’ IEEE Trans. on PAMI, vol. 26, no. 6, 2004, pp. 787-798.
- [11] U. Pal, N. Sharma, T. Wakabayashi and F. Kimura, ‘‘Handwritten Numeral Recognition of Six Popular Indian Scripts’’, In Proc. 9th International Conf. on Document Analysis and Recognition, pp.749-753,2007.
- [12] U. Garain and T. Paquet, ‘‘Off-Line Multi-Script Writer Identification Using AR Coefficients’’, In Proc. 10th International Conf. on Document Analysis and Recognition, 2009, pp. 991-995.
- [13] U. Pal and S. Datta, ‘‘Segmentation of Bangla unconstrained Handwritten Text’’, In Proc. 7th International Conf. on Document Analysis and Recognition, 2003, pp. 1128-1132.
- [14] V. Vapnik, ‘‘The Nature of Statistical Learning Theory’’ Springer Verlag, 1995.
- [15] C. Burges, ‘‘A Tutorial on support Vector machines for pattern recognition’’ Data mining and knowledge discovery, vol.2,1998,pp.1-43.