Content Independent Writer Identification using Occurrences of Writing Styles for *Bangla* Handwritings

Samit Biswas
Department of Computer Science and Engineering
Bengal Institute of Technology
Kolkata, India
E-mail: samitbiet@yahoo.com

Amit Kumar Das
Department of Computer Science and Technology
Bengal Engineering and Science University
Howrah, India
E-mail: amit@cs.becs.ac.in

***Abstract*—In most forensic cases handwritten documents are useful for person identification. It is not easy to identify a writer or narrow down a writer from a group of writings. Each person has his or her own individual handwriting styles. This paper presents an approach to extract unique features based on writers' writing styles from the handwritten *Bangla* scripts. The computed features are used to identify the writer from a group of writings. An identification strategy using those features and rejection criteria are shown. The proposed approach is applied on a collected dataset and it gives us an encouraging results.**

***Keywords*-Handwritten document image, Connected component (CC) and Correlation similarity.**

I. Introduction

Analysis of handwritten documents is important for determining the writer and for few cases writer identification is necessary for Criminal Justice System. Handwriting identification is the study for identifying or verifying the writer of a given document. Writer recognition is considered as a difficult problem to solve due to variations found in the writing, even from the same writer [1].

Handwritten words carry more individuality than handwritten allographs. The handwriting features are the cornerstone in the identification process, the classification accuracy is sensitive in terms of how the writers are scored based on these features [2].

This communication intended to choose only the useful writing styles as feature for identification. The useful styles of writing are those which the writers draw in their writings unconsciously. Feature extraction, identification strategy and rejection criteria are also presented in this report.

A. Bangla Language, Script and Handwriting

Bangla is the national language of Bangladesh and the second most popular language in India. Writing style of *Bangla* [3] is from left to right in a horizontal manner. The concept of upper or lower case is absent in *Bangla*. The basic character set comprises of 11 vowels, 40 consonants and 10 numerals.

Many *Bangla* characters have a horizontal line at the upper part called 'matra' or headline. In *Bangla* successive characters in a word touch the matra. The characters in a word usually reside in between the matra and the base line. A vowel following a consonant sometimes takes a modified (allographic) shape, and is called a vowel modifier. Depending on the vowel, the allograph is placed at the left, right (or both) or bottom of the consonant below the base line. Some part of the allograph may also be present above the matra. There can also be some compound character being combination of consonant with consonant as well as consonant with vowel.

In case of handwritten text sometimes 'matra' may be absent for some characters and modifiers may not touch the characters. It may vary from one writer to another; however, a writer always inadvertantly uses some 'matra' less or disconnected modifiers in his/her writing. This 'matra - less' disconnected modifiers may be used for identification which would substantially minimize the space and time complexity. An example for *Bangla* writing is shown in Figure 1a; here the topic for the writing is so chosen that most of the basic *Bangla* characters are present.

II. Related Work

Many well established writer identification strategies for non-Indic scripts are reported in the literature [1-9]. Marti et al. [4] computed twelve features based on visible characteristics of the writings. Using k-nearest-neighbour classifier and Feed forward Neural network identification rates 87.8% and 90.7% respectively obtained on a subset of the IAM database with 20 writers, five handwritten pages per writer.

Srihari et.al [5], [6] obtained the features from the entire document or from each paragraph, word or even a single character and constructed the features vector. Two types of features was considered Conventional features and Computational features. It required some form of detailed and elaborate user interaction.

In [7] the writing was divided into a large number of small sub images and the sub images which were morphologically similar were grouped together into same class. Authors used co-relation similarity for clustering the small sub-images.

978-0-7695-4599-8/11 $26.00 © 2011 IEEE
DOI 10.1109/NCVPRIPG.2011.40

The pattern which occurs frequently was extracted. The authors used Bayesian classifier for identification.

In [8] the authors introduced a set of features that was extracted from the contour of hand written document images at different observation levels, i.e., global and local. For the global level features the authors extracted the histograms of chain code, the first and second order differential chain codes and the the histogram of the curvature index at each point of the contour of handwriting. At the local level, the handwritten text was divided into a large number of small adaptive windows and within each window the contribution of each of the eight directions (and their differentials) is counted in the corresponding histograms. Identification was performed by computing the distance between the query image and all the images in the dataset.

Though a large number of people in the world use Indic scripts, to the best of our knowledge, there are only two work on Indic script [9], [10] which have been reported in the context of writer identification. Garain and Paquet [9] proposed an AR co-efficient feature based writer identification system for 40 *Bangla* writers. They have used at least 200 words per writer for training and testing their systems.

Chanda et al. [10] proposed a text independent writer identification system for *Bangla* script for 104 writers where discrete directional features and gradient features were used for identification. Fifty to sixty words per writer were considered for training and testing their system. Each character (symbol and modifier) of the word was segmented into an individual character/character allograph. In both the work [9], [10] the authors have not tested their system with false data; i.e., they did not impose any rejection criterion.

We intend to analyse a handwritten Indic script document with lesser amount of information by utilising only a useful set of styles of writings which are considered as features for identification.

III. Proposed Approach

We have started with the gray images of the handwritten documents. Next it is converted to binary image by a well-known global thresholding algorithm [12]. The following subsections describe similarity Measure between two images, extraction of features from handwritten documents, writer identification of handwritten documents and the rejection scheme.

A. Similarity Measure between two images

Two images X and Y of same size, can be compared by the following co-relation similarity measure [13]: $S(X, Y) = \frac{p_{11}p_{00} - p_{10}p_{01}}{[(p_{11}+p_{10})(p_{01}+p_{00})(p_{11}+p_{01})(p_{10}+p_{00})]^{1/2}}$

With p_{ij} being the number of pixels for which the two images X and Y have the values i and j respectively, at the corresponding pixel positions. When X and Y are same then $S(X, Y)$ will be 1. Here we have used a threshold for the correlation similarity, $S(X, Y) = 0.65$

B. Extraction of Features From Handwritten Documents

The extracted features are used as the reference base for the identification. The features are specific for every writer. We have developed an approach for highlighting the individual details from each handwriting. We are finding useful symbols from writings as feature and include the symbols to the dictionary. Also compute the occurrences of probabilities of each symbol of the dictionary for every writing. In brief the feature extraction procedure (dictionary formation) for the reference base is as follows:

Step 1: Consider the total number of writer for the document image in reference base is m and sample handwritten document per writer is n. If Γ is the set of images of reference base then, $\Gamma = \{I_1, I_2, I_3, ..., I_{m\times n}\}$. Perform the following *Step-2* for each of the sample images.

Step 2: Compute the connected components for the binarized image. Compute mean width, b_w of the connected components. Choose those set of connected components whose $width \leq b_w$ and $width \geq T_h$. In our case the threshold T_h is chosen as 15. This threshold (T_h) is chosen to eliminate the punctuations (too small symbols like period, comma, semi-colon etc.) because their appearance may not vary with different writers and hence not very useful for the purpose of identification.

So, after *Step-2*, if modified set of images is Γ_{mod}, then IM_i are modified images where $i = 1, 2, 3, ...m\times n$. $\Gamma_{mod} = \{IM_1, IM_2, IM_3, ..., IM_{m\times n}\}$.

Step 3: Compute the Maximum Width ($MaxWidth$) and Maximum Height ($MaxHeight$) from those set of connected components of all the modified writings (Γ_{mod}).

Step 4: Construct the dictionary of symbols of writing styles for each of the writings. The size of each of the elements of the dictionary will be $MaxHeight \times MaxWidth$. Consider each of the connected components of sample images and put it inside the window of size, $MaxHeight \times MaxWidth$ from left to right and top to bottom. (See Figure 2). Include the window to the dictionary if the similar window is not already included. The similarity measure is done by the correlation similarity; $S(X, Y)$.

Step 5: Compute the probabilities of occurrences of each symbol of the dictionary for each of the modified sample writings (Γ_{mod}) for all the writers. If Γ_{prob} is the set of probabilities of occurrences of symbols of dictionary for each sample of every writer then: $\Gamma_{prob} = \{p_1, p_2, p_3, ..., p_{m\times n}\}$. Here p_1 is a column vector, probabilities of occurrences of symbols of dictionary for modified sample image IM_1, and similarly p_2 for IM_2 and so on.

Step 6: Store the dictionary, $MaxWidth$, $MaxHeight$ and Γ_{prob} in database for further use. Here the dictionary and Γ_{prob} is used as features for writer identification.

C. Writer Identification Of Handwritten Documents

In brief, the identification procedure for a test handwritten document image consists of the following steps:

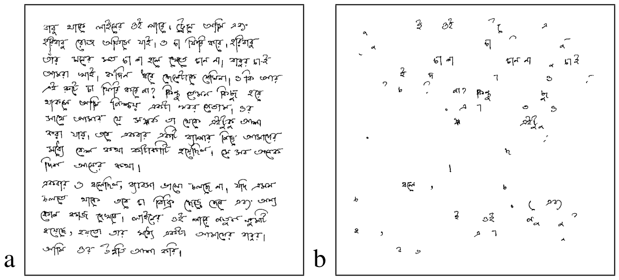

Figure 1. a) Preprocessed binarized image. b) Features: useful styles of writing considered for the writing shown in (a).

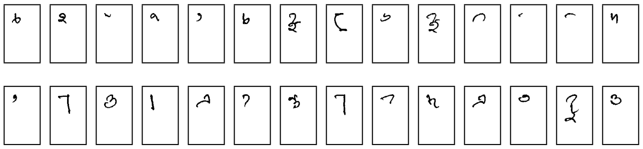

Figure 2. Few examples of the symbols included in the Dictionary.

Step 1: Compute the connected components for the binarized test image. Choose those set of connected components whose $width \leq MaxWidth$ and $width \geq T_h$ and $height \leq MaxHeight$. In our case the threshold T_h is chosen as 15. This threshold (T_h) is chosen for eliminating the too small symbols like dots, stop of words etc. Consider the modified test image is I_T.

Step 2: Choose each of the connected components from the modified test image I_T. Put the component inside the window of size $MaxHeight \times MaxWidth$, from left to right and top to bottom and find the similar symbol from the dictionary using correlation similarity measure (See Section-III-A). Compute the probabilities of occurrences of each symbol of the dictionary for the modified test image I_T based on the similarity. If Π_{prob} is the set of probabilities of occurrences of symbols of dictionary for test image I_T, then: $\Pi_{prob} = \{\pi_1, \pi_2, \pi_3, \ldots\}$ Where $\pi_i, i = 1, 2, 3, \ldots$ is the occurrences of probabilities of corresponding symbol of the dictionary.

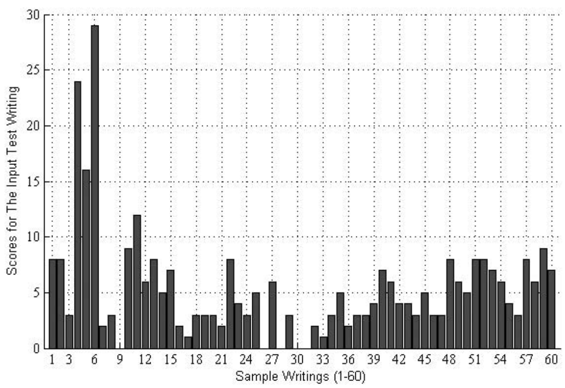

Figure 3. Plots for a test document (True) of writer-2, where x-axis represents the sample writings for all the writers in reference base and y-axis denotes the scores for the test document (True) against each writing.

Step 3: Consider the set Γ_{prob} and Π_{prob}. Γ_{prob} is already calculated in previous subsection. Find indexes of non-zero elements from $p_i, i = 1, 2, 3, \ldots, m \times n$ of Γ_{prob}. If the set of indexes are Γ_{index} then $\Gamma_{index} = \{x_1, x_2, x_3, \ldots, x_{m \times n}\}$ Where $x_i, i = 1, 2, 3, \ldots, m \times n$ is the set of indexes of non-zero elements in corresponding $p_i, i = 1, 2, 3, \ldots, m \times n$.

Compute Γ_{score} for the test document image against each of the elements in Γ_{prob} such as $\Gamma_{score} = \{s_1, s_2, s_3, \ldots, s_{m \times n}\}$ where $s_i, i = 1, 2, 3, \ldots, m \times n$ and s_i will be caclulated as follows:

For each element of x_i in Γ_{index} Compute s_i as follows:

$$t = p_i(x_i(j)) \times \Pi_{prob}(x_i(j))$$

$$s_i = \begin{cases} s_i + 1 & \text{If } t > 0 \\ s_i & \text{Otherwise} \end{cases}$$

Here $x_i(j)$ means the $j - th$ element in x_i, $p_i(x_i(j))$ is the probability of occurrences of $x_i(j)th$ dictionary symbol for the $i - th$ writing in reference base and $\Pi_{prob}(x_i(j))$ is the probability of occurrences of $x_i(j)th$ dictionary symbol for the test document image.

Step 4: Compute the maximum score from Γ_{score}, $maxScore = max(\Gamma_{score}) = max\{s_1, s_2, s_3, \ldots, s_{m \times n}\}$ Compute the index, $Sample_{id}$ of $maxScore$ from the set Γ_{score} and compute the $Writer_{id}$ as follows: $Writer_{id} = \lceil \frac{Sample_{id}}{n} \rceil$

As an example see the histogram of Γ_{score}, (Figure 3) for a test document of writer-2, where x-axis represents the sample writings for all the writers and y-axis denotes the scores for the test document against each sample writing. Here, sample per writer is considered as three. Sample-6 have the maximum score; so the writer of the test document is the writer of Sample-6.

D. Rejection Scheme

The goal of using the rejection strategy is to improve the accuracy of the proposed approach. For identification, a score is calculated for the test document against each samples of the reference base. Find all the local maxima. Compute the score difference(WS_{th}) between the maximum score ($maxScore$) of local maxima and mean (A_m) of all remaining local maxima. $WS_{th} = maxScore - A_m$.

If the score difference, WS_{th} is greater than a predefined threshold, T_r the test document is written by the writer of the sample document. Otherwise the test document does not match with any of the sample documents of the reference base. For Experimentation we have chosen T_r as 5.

IV. Experimental Result

Since there is no standard benchmark *Bangla* handwritten database for writer identification, we have created our own

Table I
WRITER IDENTIFICATION RATE (TRUE) FOR BANGLA HANDWRITINGS

Top Choices	TAR(%)	TRR(%)
1	86.67	13.33
2	90.00	10.00
3	90.00	10.00

Table II
WRITER IDENTIFICATION RATE (FALSE) FOR BANGLA HANDWRITINGS

FAR(%)	FRR(%)
29.17	70.83

Bangla hand written database (*BESUS Database*). This database consists of images of writings of 50 writers. Every writer has four samples on two different topics. Here the topics for the writings is so chosen that most of the basic *Bangla* characters are present in the writings. One of the samples is shown in Figure 1a. 50% of the total writer of the database are female and the remaining 50% writer are male with the age group varying from 21 to 23 years. We have taken 2 or 3 samples per writer in the reference base and another one is used as test document. The content of test document is not same as the sample document. Half of the writings of the database is used for the reference base and remaining is used for test document image.

We have tested the proposed method by considering the sample writings on one topic and test writings on another topic. The accuracy of the proposed approach have measured on the basis of *True Acceptance Rate (TAR)*, *True Rejection Rate (TRR)*, *False Acceptance Rate (FAR) and False Rejection Rate (FRR)*.

For BESUS database Table I shows the writer identification rate (true) of the proposed approach for *Bangla* handwritings and Table II shows the writer identification rate (false) of the proposed approach for *Bangla* handwritings. Table III shows the comparison result. Note that Chanda et al. [10] did not use false data to test the performance and TAR only may not reflect the true performance index in the absence of FRR.

V. CONCLUSION

Experimental results have demonstrated that our method is meaningful for writer identification for *Bangla* handwritten scripts. The decision values for the thresholds used in the proposed approach are chosen by testing over some sample experimental dataset. The proposed method works well but there are few limitations; if all the symbols/words of writing are connected then the proposed approach will fail and if the total number of writers or documents in reference base is increased then memory requirement for the dictionary is also increased. We plan to solve these shortcomings in future and also to increase the number of writers for experimentation. Possible extension of this work may include the identification of small sized handwritten documents (ransom notes, threatening letters etc.).

Table III
COMPARISON RESULT (TOP-1 CHOICE)

Method	TAR (%)	FRR(%)
Garain and Paquet [9]	75	–
Chanda et al. [10]	95.19	–
Proposed Method (BESUS database)	86.67	70.83

REFERENCES

[1] A. Imdad, S. Bres, V. Eglin, C. Rivero-Moreno, and H. Emptoz, “Writer identification using steered hermite features and svm,” in *ICDAR*, 2007, pp. 839–843.

[2] K. M. B. Abdl and S. Z. M. Hashim, “Swarm-based feature selection for handwriting identification,” *Journal of Computer Science*, vol. 6, no. 1, pp. 80–86, 2010.

[3] A. Bishnu and B. B. Chaudhuri, “Segmentation of bangla handwritten text into characters by recursive contour following,” in *ICDAR*, 1999, pp. 402–405.

[4] U.-V. Marti, R. Messerli, and H. Bunke, “Writer identification using text line based features,” in *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*, 2001, pp. 101 –105.

[5] S. N. Srihari, S.-H. Cha, H. Arora, and S. Lee, “Individuality of handwriting,” *Journal of Forensic Sciences*, vol. 47, no. 4, pp. 1–17, 2002.

[6] S. N. Srihari, M. J. Beal, K. Bandi, and V. Shah, “A statistical model for writer verification,” in *ICDAR*, 2005, pp. 1105–1109.

[7] I. Siddiqi and N. Vincent, “Writer identification in handwritten documents,” in *ICDAR*, 2007, pp. 108–112.

[8] ——, “A set of chain code based features for writer recognition,” in *ICDAR*, 2009, pp. 981–985.

[9] U. Garain and T. Paquet, “Off-line multi-script writer identification using ar coefficients,” in *ICDAR*, 2009, pp. 991–995.

[10] S. Chanda, K. Franke, U. Pal, and T. Wakabayashi, “Text independent writer identification for bengali script,” in *ICPR*, 2010, pp. 2005–2008.

[11] P. Purkait, R. Kumar, and B. Chanda, “Writer identification for handwritten telugu documents using directional morphological features,” in *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, nov. 2010, pp. 658 –663.

[12] B.Chanda and D. D. Majumder, “Digital image processing and analysis,” 2000, iSBN: 81-203-1618-5.

[13] A. Bensefia, A. Nosary, T. Paquet, and L. Heutte, “Writer identification by writer’s invariants,” in *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on*, 2002, pp. 274 – 279.