2012 10th IAPR International Workshop on Document Analysis Systems

Writer Identification of *Bangla* handwritings by Radon Transform Projection Profile

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*—Writer identification is the task of determining the person whose handwritten sample is available in a set of writings, collected from multitude of writers. This has useful applications in many areas, notably in forensic analysis. The task of writer identification is quite difficult due to minimal variations found in different handwritten samples from same person/writer. Several identification algorithms have been proposed so far which are mostly for non-Indic writings. This paper presents a new approach for extracting two different sets of components (essentially fragments of characters); namely *fragment set-A* and *fragment set-B*. Features are extracted from each element of these two sets to identify the writing style of a particular person. The features are computed based on Radon transform projection profile. The proposed approach uses lesser amount of information from the handwritten samples; thus saving computation time as well as memory requirement. The condition to determine that the writer is unknown (i.e., there is no handwritten sample from that writer in reference base) is also proposed. The approach is tested on a collected dataset of *Bangla* writings and the experimental results are encouraging.**

***Keywords*-Handwritten document image, *Bangla* Language, Allograph, Writing style and Radon transform.**

I. Introduction

Writer identification for a handwritten script is a crucial task. It can be used in criminal justice system. Handwriting identification is the study for identifying or verifying the writer of a given document. Research on handwritten document analysis continued from last few decades. Most of the research work carried out for writer identification/verification is based on non-Indic script.

Most characters of words may be written using more than one different shape or model. These models are called allographs [1]. For example, the letter 'a' can be written in several different ways, such as an upper-case, a block printed, or a cursive variant. Those variants are different character allographs for the character 'a'; visually different but indicating the same thing. So, handwritten words carry more individuality than handwritten allograph. The handwriting features are the cornerstone in the identification process and the classification accuracy is sensitive in terms of how the writers are scored based on the features [2]. Features for identification may be visible characteristic of writing, for example; width, slant and height of the three main writing zones or texture based features i.e., contour based, run length based and autocorrelation etc.

The originality of this work is the extraction of two different sets of components (essentially fragments of characters); namely *fragment set-A* and *fragment set-B* and extraction of features from those based on Radon transformation. The fragments in *set-A* are those which writer draws in his/her writings unconsciously. These fragments are unique for every writer and whenever a writer writes something these components will be present there. The fragments in *set-B* are extracted by analysing the words (where most of the symbols are connected). These components are language dependent. We have considered here *Bangla* writings. A basis for *Bangla* language, handwritings is presented in the next subsection.

A. Bangla Language, Script and Handwriting:

Bangla is the national language of Bangladesh and the second most popular language in India. All major Indian scripts including *Bangla* are mixtures of syllabic and alphabetic scripts. 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 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 vivid description of the characteristics of Bangla script is available in [4]. 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 inadvertently uses some 'matra less or disconnected modifiers in his/her writing. This matra-less

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.98

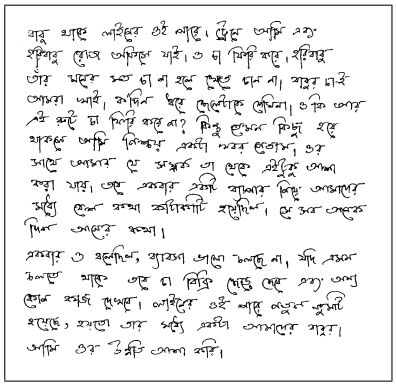

Figure 1. Pre-processed Binarized Image

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 1; Here the topics for the writings are so chosen that most of the basic *Bangla* characters are present.

II. Related Work

Various schemes for writer identification and verification have been proposed by many researchers. Many well established writer identification strategies for non-Indic scripts are reported in [2, 4-15]. Marti et al. [5] computed twelve features based on visible characteristics of the writing; for example width, slant and height of the three main writing zones. Using k-nearest neighbour classifier and feed forward neural network identification rates of 87.8% and 90.7% were obtained respectively in tests on a subset of the IAM database with 20 writers and five handwritten pages per writer.

Srihari et al. [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 were considered – conventional features and computational features. It required some form of detailed and elaborate user interaction for separating the characters from words.

In [7], [8] the writer was considered to be characterized by a stochastic pattern generator producing a family of character fragments (fraglets). The authors formed an independent training set using a codebook of such fraglets. Kohonen self-organised map was used for forming the codebook. Identification rate is 83% on dataset of 210 writers for upper case texts.

Schlapbach and Bunke [9] computed nine features; three global and six local. The global features were the fractions of black pixels in the window, the center of gravity and 2nd order moment. The local features were the positions of upper and lower most pixels, the number of black to white transitions in the window and the fraction of black pixels between the upper and lower most black pixels. Using HMM based recogniser identification rate achieved was 95.6% on a subset of the IAM database with 100 writers.

Bulacu et al. [10], [11], [12] used edge-direction distribution, edge-hinge distribution, run length distribution and their combinations as features for writer identification. By combining directional features, grapheme, and run-length information, yielding, on the different data set, Top-1 writer identification rate varies from 29% to 89% percent.

In [13] the writing was divided into a large number of small sub images and the sub images that were morphologically similar were grouped together into same class. Authors used co-relation similarity for clustering the small sub-images. The pattern which occurs frequently was extracted. The authors [13] used Bayesian classifier to identify the author of unknown writing.

In [14] 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 author extracted the histograms of chain code, the first and second order differential chain codes and the the histogram of the curvature indices 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 is only two published research work on Indic script [15], [16] in the context of writer identification. Garain and Paquet [15] 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 system. But very often a questioned document is deprived of such huge number of hand written words.

Chanda et al. [16] proposed an text independent writer identification system for *Bangla* script. Discrete directional features and gradient features were used for writer identification. They have used at least 50-60 words per writer for training and testing their system for 104 writers. Each of the characters (symbols and modifiers) of the word was segmented into an individual character/character allograph. That means they have to consider too many character/character allograph. In both of the works [15], [16] authors have not also tested their system with unknown data, i.e., they did not define any condition to determine that a given writing is of unknown writer.

We intend to analyse a handwritten Indic (*Bangla*) script with lesser amount of information. Here we extract Radon transform projection profile from two types of fragments (*fragment set-A* and *fragment set-B*). This projection profile is used for identification/verification.

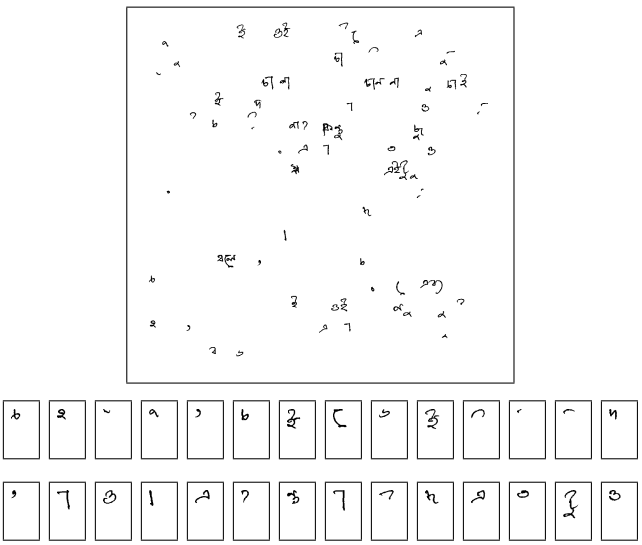

Figure 2. Fragment set-A extraction from handwritings.

III. Proposed Approach

We have started with the gray image of the handwritten documents. Next it is converted to binary image, (I) by the well-known global thresholding algorithm [17]. The following subsections describe the extraction of the *fragments* from the writings, extraction of features from each *fragment*, identification strategy and verification strategy.

A. Extract the fragments from the writings

1) Fragment set-A: Compute the connected components of the binarized image, (I). Compute mean width, b_w of the connected components and keep those set of components whose *width* $\leq b_w$ and *width* $\geq T_h$. Here 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 useful for the purpose of identification. Each of the detected symbols has to be resized to $m \times n$. We have used the size 60×40. See Figure 2, here *fragment set-A* is extracted from the hand written documents of Figure 1.

2) Fragment set-B: Compute connected components of the binarized image, (I). Remove those set of components whose *width*$\leq$ T_h. As explained in subsection-A(1) to eliminate too small symbols, T_h is chosen to be 15. Now remove *fragment set-A* from it and consider each of the connected components. Find the vertical projection profile. Cut the components vertically from the local minimas of the vertical projection profile. After cutting, remove the components whose width is less than or equal to T_h. See Figure 3, *fragment set-B* is extracted from the handwritten words of Figure 1.

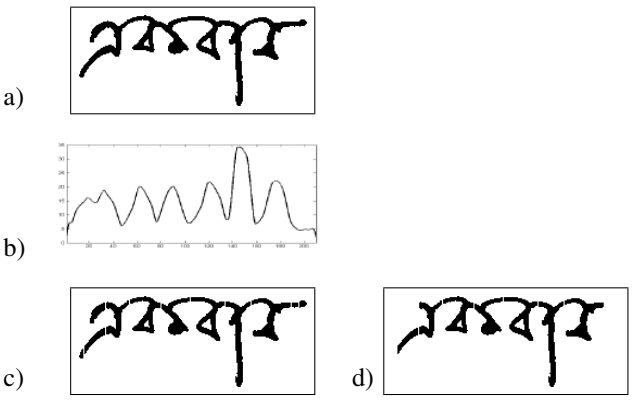

Figure 3. Fragment set-B Extraction from writings: a) Original component b) Vertical projection profile c) Cutting vertically according the local minima's of the vertical projection profile d)Fragment set-B Extracted.

B. Feature Extraction

Consider each of the extracted *fragments* and resize it to $m \times n$. Here we have resized each of the *fragments* to 60×40. The size of the feature vector for a $m \times n$ image will be the length of its diagonal. Compute Radon transform projection profile of the resized *fragments* for the orientations $0^0, 45^0, 90^0, 135^0$. A projection of a two-dimensional function $f(x, y)$ is a set of line integrals. The radon transform [18], [19] computes the line integrals from multiple sources along parallel paths, or beams, in a certain direction. The beams are spaced 1 pixel apart. To represent an image, the radon function takes multiple, parallel-beam projections of the image from different angles by rotating the source around the center of the image. For example, the line integral of $f(x, y)$ in the vertical direction is the projection of $f(x, y)$ onto the x-axis; the line integral in the horizontal direction is the projection of $f(x, y)$ onto the y-axis.

Projections can be computed along any angle, θ. In general, the Radon transform, R_θ of $f(x, y)$ is the line integral of f parallel to the y'-axis. Size of the feature vector for each orientation will be the size of diagonal.

$$R_\theta(x) = \int_{-\infty}^{\infty} f(x' \cos(\theta) - y' \sin(\theta), x' \sin(\theta) + y' \cos(\theta)) dy' \tag{1}$$

where

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

C. Identification Strategy

For the task of writer identification we calculate score of written document image of an unknown writer against each of the written document of the reference base. Extract *fragment set-A* and *fragment set-B* from test handwritings (see subsection III-A) and extract features from each of the fragments. Compute Euclidean distance from each of the fragments to the reference base.

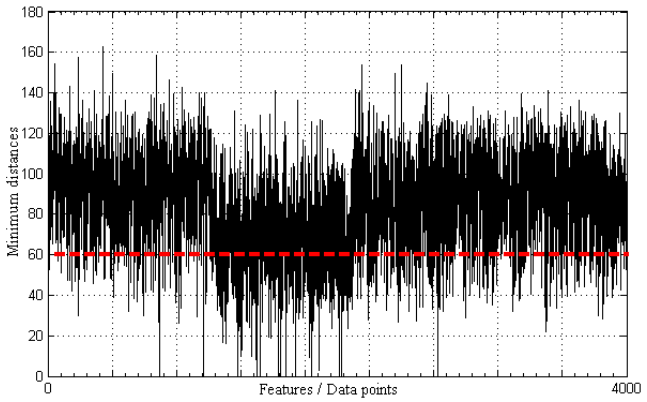

Figure 4. Portion of row vector R, horizontal axis represents features/datapoints for each fragment of reference base and vertical axis represents minimum distances with respect to test writings.

Compute a distance matrix, M for each feature vector of unknown writer to the feature vectors of reference base. Here rows in M are the data points for the test document image and columns are the data points for the reference base images and the cells represents the distance between the test data point and reference base data points. W_{id} is a 1D vector, size of this vector is same as of the total number of features in the reference base. These values are used as an index of writer corresponding to the extracted features in reference base. The score for the test document image against each document image in reference base is calculated as follows:

Step 1: Compute a row vector, R from the M by choosing the minimum values from each column of M. See Figure 4, where horizontal axis represents the data points in reference base and vertical axis represents the minimum value (distance) of each column of distance matrix.

Step 2: Choose those set of indexes, I_{dx} from R where $R \leq T_{h1}$. See next subsection for threshold (T_{h1}) selection.

Step 3: Choose those set of values from W_{id}, for indexes I_{dx}. Compute the histogram of the chosen values. The corresponding index of the peak in the histogram from W_{id} will be the writer-id for the test document.

D. Verification Strategy

The goal of verification strategy is to improve the accuracy of the proposed approach. The purpose of verification is to ascertain if the two writing is written by same writer or not. Extract the features i.e., *fragment set-A* and *fragment set-B* from both the writings. Compute the distance between the features of two writings. If the computed distance is less than a predefined threshold then the writer is present in the reference base. The threshold values are computed as follows:

Step 1: Compute the histogram for the vector, R. Find the index, I_{dx} of peak from the histogram. The threshold is computed as follows: $T_{h1} = I_{dx}/2$.

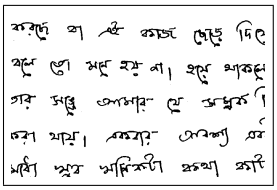

Figure 5. A portion from handwritten script - considered lesser information

Step 2: Compute the histogram of W_{id} and compute the average number of occurrences, W_{avg} for an element (writer) within W_{id}.

Step 3: Choose those set of indexes, I_{fid} from R, where distances is less than or equal to T_{h1}. Compute the histogram for $W_{id}(I_{fid})$. If the peak value for this histogram is greater than or equal to W_{avg} then T_{h1} is used as threshold. If threshold calculation is not possible then the writer for the test writing is not present in the reference base.

IV. Dataset Details and Experimental result

Since there is no standard benchmark *Bangla* handwritten database for writer identification, we have created our own *Bangla* hand written database (*BESUS Database*).

The *BESUS* database consists of images of writings of 55 writers. Every writer has four samples on two different topics. Here the topic 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 1. Fifty percent of the total writers of the database are female and the remaining writers are male with the age group varying from 21 years to 23 years. We have taken one sample per writer in the reference base and remaining samples are used as test document. The content of test document is not same as of sample document and it consists of approx 90-110 words. Table I shows the performance result. Later, we have considered two or three sentences consist of 20-30 words from a test handwritten script (see Figure 5 as an example). Table II shows the performance result for the lesser amount of information. For verification purpose half of the writings of the database is used for the reference base and remaining is used for test document image.

Though there are many research work on writer identification for non Indic scripts, only two work [15], [16] has been reported in the context of Indic scripts. Both of the work has not defined any condition for the unknown writings. Here *true acceptance rate (TAR)* and *false rejection rate (FRR)* is used for comparing the proposed approach with others. Table III shows the comparison results.

V. Conclusion

Experimental results have demonstrated that our method is meaningful for writer identification for *Bangla* handwritten scripts. The method works well but few limitations of this method is that if the total number of writer or documents is

Table I
WRITER IDENTIFICATION RATE OF THE PROPOSED APPROACH
(TESTING SCRIPT CONSIST OF 90 - 110 WORDS)

Fragments	Top 1	Top 2	Top 3
Fragment set-A	54.54	70.9	70.9
Fragment set-B	74.5	87.27	87.27
set-A and set-B	83.63	92.72	92.72

Table II
WRITER IDENTIFICATION RATE OF THE PROPOSED APPROACH
(TESTING SCRIPT CONSIST OF 20 - 30 WORDS)

Fragments	Top 1	Top 2	Top 3
Fragment set-A	21	30	30
Fragment set-B	58	74	74
set-A and set-B	61.8	80	80

Table III
COMPARISON RESULT (UPTO TOP-3 CHOICE)

Method	TAR (%)	FRR(%)	Word(#)
Garain and Paquet [15]	82.5	not imposed	200
Chanda et al. [16]	95.19	not imposed	50-60
Ours (BESUS Database)	92.72	84	90-110
	80	72	20-30

Word(#): Number of words considered, per test handwritten script.

increased then memory requirement for the reference base is also increased. We plan to solve above limitations in our future work. We are also trying to increase the number of writers for experimentation. Proposed approach is tested for the scripts of 90-110 words and 20-30 words separately and identification results achieved are 92.72% and 80% respectively. Possible extensions of this work may include the identification of small sized handwritten documents (ransom notes, threatening letters etc.). The authors are currently working in this direction.

REFERENCES

[1] M. Parizeau and R. Plamondon, "A fuzzy-syntactic approach to allograph modeling for cursive script recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, no. 7, pp. 702 –712, jul 1995.

[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] B. Chaudhuri and U. Pal, "An ocr system to read two indian language scripts: Bangla and devnagari (hindi)," in *Document Analysis and Recognition, 1997., Proceedings of the Fourth International Conference on*, vol. 2, aug 1997, pp. 1011 – 1015 vol.2.

[5] 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.

[6] 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.

[7] L. Schomaker, K. Franke, and M. Bulacu, "Using codebooks of fragmented connected-component contours in forensic and historic writer identification," *Pattern Recognition Letters*, vol. 28, no. 6, pp. 719 – 727, 2007.

[8] L. Schomaker and M. Bulacu, "Automatic writer identification using connected-component contours and edge-based features of uppercase western script," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 787–798, 2004.

[9] A. Schlapbach and H. Bunke, "Using hmm based recognizers for writer identification and verification," in *Frontiers in Handwriting Recognition, 2004. IWFHR-9 2004. Ninth International Workshop on*, oct. 2004, pp. 167 – 172.

[10] M. Bulacu and L. Schomaker, "Text-independent writer identification and verification using textural and allographic features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 701–717, 2007.

[11] ——, "Combining multiple features for text-independent writer identification and verification," in *In Proc. of 10th IWFHR*, 2006, pp. 281–286.

[12] M. B. Lambert and L. Schomaker, "Writer identification using edge-based directional features," in *in Proc. of ICDAR 2003*. IEEE Computer Society, 2003, pp. 937–941.

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

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

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

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

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

[18] M. Barva, J. Kybic, J.-M. Mari, and C. Cachard, "Radial Radon transform dedicated to micro-object localization from radio frequency ultrasound signal," in *IEEE International Frequency Control Symposium and Exposition*. IEEE, Aug. 2004, pp. 1836–1839.

[19] M. R. Hejazi, G. Shevlyakov, and Y.-S. Ho, "Modified discrete radon transforms and their application to rotation-invariant image analysis," in *Proc. IEEE 8th Workshop on Multimedia Signal Processing*, 2006, pp. 429–434.