Text Independent Writer Identification for Oriya Script

Sukalpa Chanda
Department of Computer Science and Media Technology, Gjøvik University College, Norway
E-mail:- sukalpa@ieee.org

Katrin Franke
Department of Computer Science and Media Technology, Gjøvik University College, Norway
E-mail:- kyfranke@ieee.org

Umapada Pal
Computer Vision and Pattern Recognition Unit, Indian Statistical Institute India
E-mail:-umapada@isical.ac.in

***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 ensured. Lack of data threatens system reliability in such cases. We here propose a writer identification system for Oriya script which is capable of performing reasonably well even with small amount of text. Experiments with curvature feature are reported here, using Support Vector Machine (SVM) as classifier. We got promising results of 94.00% writer identification accuracy at first top choice and 99% when considering first three top choices.**

Keywords-: Writer Identification; Oriya Script; Curvature Feature; SVM.

I. INTRODUCTION

Writer identification utility could be an important tool in any computational forensic system. There are many pieces of work on writer identification [1, 2, 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. Though a large number of people in the world use Indic scripts, to the best of our knowledge, there are very few pieces of work on Indic scripts [12, 16, 17] in the context of writer identification. Garain et al. [12] 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. In [16] a Gradient feature-based writer identification system is proposed for Bengali script which can perform well even when there are 50-70 words per writer. A writer identification system for Telegu script is proposed in [17]. In [17] the authors considered 5 samples from each of 22 writers; there they used structural information based features.

In order to encounter adversary, like scarcity of data content in questioned documents Oriya script, we here propose a robust writer identification system.

II. LINE , WORD 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. For word segmentation from a line, we compute vertical histogram of the line. In general the distance between two consecutive words of a line is bigger than the distance between two consecutive characters in a word. Taking the vertical histogram of the line and using a distance criteria [13] we segment words from lines.

In principle, when two or more characters in Oriya get connected one of the four following situations happens in most of the cases: (a) two consecutive characters create a large bottom reservoir; (b) the number of reservoirs and loops in a connected component will be greater than that of an isolated component; (c) two consecutive characters create a small top reservoir near mean line (d) the shape of the touching character will be more complex than isolated characters, (for details please see [13]). Computing different features obtained by the above observations we identify isolated and touching characters. If a component is detected as touching by the above algorithm then we segment the connected pattern to get its individual characters. For the segmentation of a touching pattern at first, the touching position is found. Next, based on the touching position, reservoir base-area points, topological and structural features the component is segmented to generate character allograph. Details about the method can be found in [13].

III. FEATURE EXTRACTION

Oriya handwritten text is characterized by mainly round shaped characters/character allograph. But roundness in character allograph varies amongst different writers, even when they write the same text. Our curvature features are used as a descriptor to express this character allograph level dissimilarity present in the handwritings of different writers. Dimension of our curvature feature is 1176 which is later reduced to 392 using PCA.

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.86

A. Feature computation for Curvature feature

Curvature feature used in this paper has been calculated using bi-quadratic interpolation method as described in [5] and the procedure is as follows:

The curvature c at x_0 in a gray scale image is defined by

$$c = \frac{y''}{\sqrt{(1+y'^2)^3}} \tag{1}$$

where $y = g(x)$ is the equi-gray scale curve passing though x_0, (x, y) is the spatial co-ordinates of x_0, y' and y'' are the first and second order derivative of y, respectively. The derivatives y' and y'' are derived from bi-quadratic interpolating surface for the gray scale values in the 8-neighbourhood of x_0. (The eight neighborhood of x_0 is shown in Fig.1. The pixel value of x_k is denoted by f_k. The bi-quadratic interpolated surface is given by

$$z = [1 \quad x \quad x^2]\begin{bmatrix} a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} \\ a_{20} & a_{21} & a_{22} \end{bmatrix}\begin{bmatrix} 1 \\ y \\ y^2 \end{bmatrix} \tag{2}$$

Then the equi-gray curve passing through x_0 is given by

$$(a_{22}x^2 + a_{12}x + a_{02})y^2 + (a_{21}x^2 + a_{11}x + a_{01})y + a_{02}x^2 + a_{10}x + a_{00} - f = 0 \tag{3}$$

Differentiation of both sides of Eq. (3) by x we get

$$y' = \frac{-\{(2a_{22}x + a_{12}x)y^2 + (2a_{21}x + a_{11})y + 2a_{20}x + a_{10}x\}}{2y(a_{22}x^2 + a_{12}x + a_{02}) + a_{21}x^2 + a_{11}x + a_{01}} \tag{4}$$

Substituting the co-ordinates (0,0) of x_0 to (4), the value of y' at x_0 is given by

$$y' = -a_{10}/a_{01} \tag{5}$$

Similarly, the value of y'' at x_0 is given by

$$y'' = -2(a_{10}^2 a_{02} - a_{01}a_{10}a_{11} + a_{01}^2 a_{20})/a_{01}^3 \tag{6}$$

Solving the simultaneous liner equations (2) holding for 8-neighbour of x_0, the coefficients of the bi-quadratic surface are given by

$$a_{10} = (f_1 - f_5)/2,\ a_{20} = (f_1 + f_5 - 2f_0)/2$$

$$a_{01} = (f_3 - f_7)/2,\ a_{02} = (f_3 + f_7 - 2f_0)/2$$

$$a_{11} = (f_2 - f_8) - (f_4 - f_6)/4 \tag{7}$$

The coefficients a_{10} and a_{20} are respectively, the first and the second order partial derivatives of $f(x, y)$ with respect to x, a_{01} and a_{02} are similar partial derivatives with respect to y, and a_{11} is the derivative obtained with respect to x and y. Substituting Eqs. (5) and (6) to Eq. (1), the curvature is given by

$$c = -2(a_{10}^2 a_{02} - a_{01}a_{10}a_{11} + a_{01}^2 a_{20})/(a_{10}^2 + a_{01}^2)^{\frac{3}{2}} \tag{8}$$

By definition (8), the curvature is indefinite if $a_{10} = a_{01} = 0$. When such situation occurs then we assume curvature is zero in our algorithm.

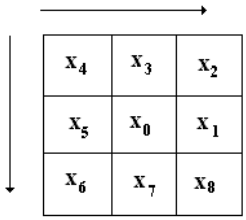

Figure 1. Neighborhood of a pixel, X_0

To get the curvature feature the following steps are applied.

Step 1: The direction of gradient is quantized to 32 levels with $\pi/16$ intervals.

Step 2: The curvature c computed by the above formula (8) is quantized into 3 levels using a threshold t (for concave, linear and convex regions). For concave region $c \le -t$, for linear region $-t < c < t$ and for convex region $c \ge t$. We assume t as 0.12 in our experiment.

Step 3: The strength of the gradient is accumulated in each of the 32 directions and in each of the 3 curvatures levels of each block to get 49x49 local joint spectra of directions and curvatures.

Step 4: A spatial and directional resolution is made as follows. A smoothing filter [1 4 6 4 1] is used to get 16 directions from 32 directions. On this resultant image, another smoothing filter [1 2 1] is used to get 8 directions from 16 directions. Further more, we use a 31 x 31 two-dimensional Gaussian-like filter (See Fig.2) to get smoothed 7 × 7 blocks from 49 x 49 blocks (shown in Fig.3). So, we get 7×7×8 = 392 dimensional feature vector. Using curvature feature in 3 levels we get 392 × 3 =1176 dimensional features.

Step 5: Using principal component analysis we reduce 1176 dimensional feature vector to 392 dimensional feature vector and we fed this 392 dimensional feature vector to our classifier.

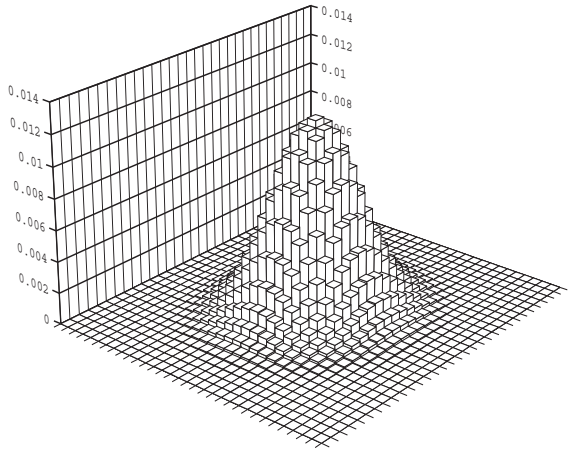

Figure 2. Example 31 x 31 two-dimensional Gaussian-like filter used for smoothing.

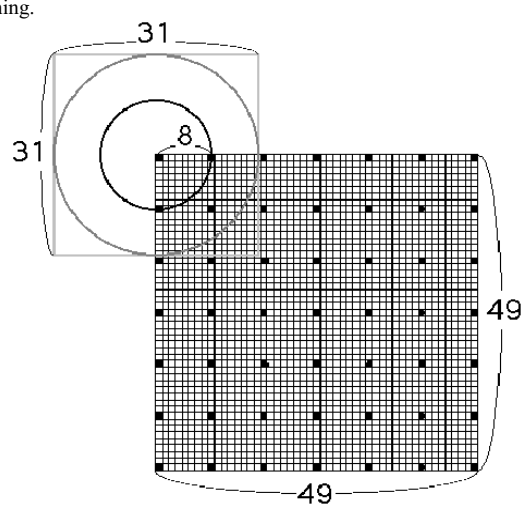

Figure 3. Illustration of getting 7 x 7 blocks from 49 x 49 blocks.

IV. CLASSIFIER AND EXPERIMENTAL DESIGN

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 100 different writers, in an average about 60-80 words per writer were used for training. We got best optimized results when gamma parameter ($1/2\sigma^2$) is set to 0.05. The penalty multiplier parameter is set to 1. Details of SVM can be found in [14] [15].

For evaluating each test image we did the following: (i) Let in a test image we get N character allograph by applying the method as discussed in Section II. (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)If amongst those N character allographs, 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 a rejection.

V. DATASET DETAILS

Our dataset consists of two sets of handwriting from each of 100 writers. One set is used for training and other set for testing. Both set contains different text with varied number of words, from each writer. Our training (testing) dataset comprises of 60-80 (60-80) Oriya words in average. All data were scanned to 300 dpi in tiff file format. We mainly performed our experiment to investigate the following: (i) Robustness of curvature features to express discriminating characteristics of each individual writer. (ii) To identify dissimilar character allograph shapes amongst 100 different writers.(iii) To identify most similar character allograph shapes amongst 100 different writers.

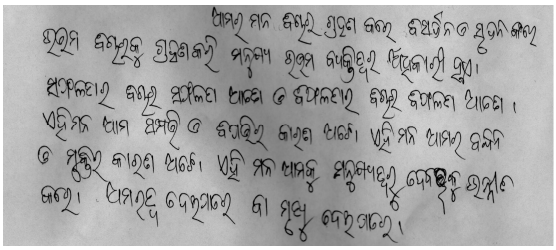

Figure 4. Example of a test file from one of our writers.

VI. RESULTS AND DISCUSSION

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 I, we report accuracy of our system using curvature feature. From the experiment on 100 writers, there were 5 misclassifications and one rejection.

B. Most dissimilar character allographs amongst 100 writers

Here we tried to investigate some character allograph shapes which actually help us in differentiating our 100 writers. We identified such character allograph based on two criteria, (i) We considered character allograph with a high confidence

score [11] (with confidence score of at least 0.7) that were assigned to the right writer, we call them C_{hp} (ii) By looking for most frequent character shapes, those were correctly assigned to the right writers. We call them C_{cs} . We can conclude that those C_{hp} and C_{cs} largely contributed in discriminating Oriya writers. We noted that C_{hp} type character allograph not necessarily belongs to C_{cs} type character allograph or vice-versa. There was a lot of C_{cs} type character allograph those had a top choice confidence score of even less than 0.5. It is not mandatory that character allograph shapes once identified as C_{hp} type character allograph always had a high confidence score value. In figure 5 we show few examples of common character allograph shapes that helped us in discriminating amongst 100 writers.

C. *Characteristics of similar shaped character allographs amongst 100 writers*

We were also interested to find out characteristics of similar shaped character allograph generated by 100 writers that actually reduces our system accuracy. We can conclude that those shapes were most difficult to be assigned to the correct class/writers. We noticed that majority of character allograph those got wrongly assigned to incorrect writers, had a very low confidence score on the top choice. Figure 6 is a graph which shows the distribution of confidence score value amongst all such character allograph those were wrongly assigned to incorrect writers. We can see about 60% of wrongly assigned character allographs had a confidence score of 0.3 or less, whereas about only 5% of those wrongly assigned character allographs obtained a confidence score of 0.6 or more. In the error analysis section we will see why those few erroneously assigned character allographs had such high confidence score value.

Figure 5. Some common character allograph shapes which highly contributed in discriminating different writers

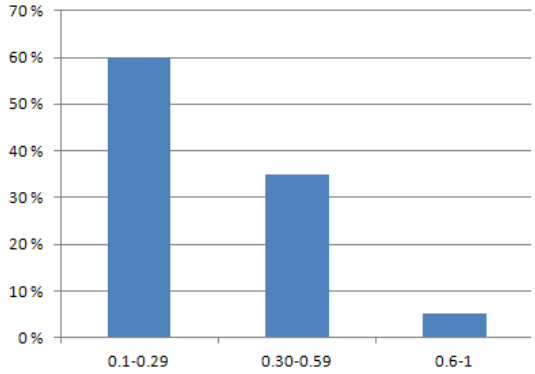

Figure 6. Confidence score distribution amongst erroneously classified character allograph.

D. *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% accuracy with curvature feature, when we consider top three choices of our majority voting instead of the top one.

TABLE I. WRITER IDENTIFICATION ACCURACY ON DIFFERENT NUMBER OF TOP CHOICES OF SVM.

No. of Top Choice	Accuracy
1	94.00%
2	97.00%
3	99.00%

E. *Error Analysis*

We considered those test images which are misclassified for further analysis. 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 or 3rd position. We analyzed the reason and found that character allograph were assigned to wrong classes due to mainly two reasons, (i) Sometimes character allograph from two different writers were visually very similar. (ii) Deformed character allograph was formed due to erroneous segmentation, where three or more character/character allograph forms a single character-component. With the help of figure 7 we show examples where two different writers produce very similar character allograph. Here the left character allographs in figure 7 is from the test image of writer 15 and were assigned to writer 17 with a confidence score of 0.8. We were surprised to see such high confidence score on erroneous classifications. We analyzed all the character allograph generated from the training file of writer 15 and writer 17. Unfortunately there were no similar character allographs in the training file of writer 15. But in the training file of writer 17 we found very similar characters (for an example please look into right hand character allographs in figure 7). So we can conclude that if

our training process encounters character allograph of very similar shapes from different writers. During testing character allograph of those writers might get misclassified with high confidence score.

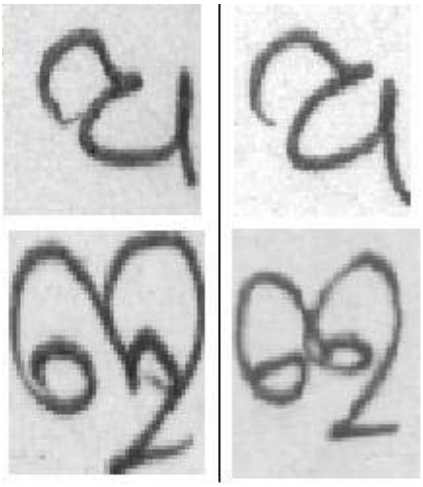

Figure7. Four visually similar shaped character allograph: (left) from test image of writer 15, (right) from training image of writer 17.

F. Comparison with similar other works

Though there are many pieces of work on writer identification for non-Indic scripts, only few pieces of work [12, 16, 17] have been reported in the context of Indic scripts. 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. Another work [16] on same Bengali script reports an accuracy of 95.19% with 104 writers in a constrained environment of only 50-60 words per writer. There gradient features along with SVM classifier were used. A system for Telegu text independent writer identification is proposed in [17]. They have achieved an accuracy of 98% but they have considered only 22 writers. Here we have considered 100 writers and obtained an accuracy of 94% considering only the top choice of our classifier.

VII . CONCLUSION

Here we propose a system for Oriya text independent writer identification using directional chain-code and curvature-based features. From the experiment on 100 writers we got promising results of 94% writer identification accuracy. In future we would like to implement a L1-norm based Multiple Kernel SVM for its inherent feature selection/dimensionality reduction and classification capability, and compare with our present technique.

REFERENCES

[1] S. N. Srihari, M. Beal, K. Bandi, V. Shah and P. Krishnamurthy, "A Statistical Model for Writer Verification", Proc. 8th International Conf. on Document Analysis and Recognition, 2005, pp.1105-1109.

[2] Lambert Schomaker, Katrin Franke, Marius Bulacu, "Using codebooks of fragmented connected-component contours in forensic and historic writer identification", Pattern Recognition Letters, vol.28(6), 2007, pp. 719-727.

[3] K. Franke, O. Bünnemeyer, and T. Sy, "Writer identification using ink texture analysis", 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", 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," 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] T.-F. Wu, C.-J. Lin, and R. C. Weng. "Probability Estimates for Multi-class Classification by Pair wise Coupling". Journal of Machine Learning Research, 5, 2004, pp. 975-1005.

[12] U. Garain and T. Paquet, "Off-Line Multi-Script Writer Identification Using AR Coefficients", Proc. 10th International Conf. on Document Analysis and Recognition, 2009, pp. 991-995.

[13] N. Tripathy and U. Pal, "Handwriting Segmentation of Unconstrained Oriya Text", Proc. 9th International Workshop on Frontiers in Handwriting Recognition (IWFHR), 2004, pp.306-311.

[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.

[16] Sukalpa Chanda, Katrin Franke, Umapada Pal and Tetsushi Wakabayashi, "Text Independent Writer Identification for Bengali Script", Proc. 20th International Conference on Pattern Recognition, 2010, pp.2005-2008.

[17] Pulak Purkait, Rajesh Kumar, Bhabatosh Chanda, "Writer Identification for Handwritten Telugu Documents Using Directional Morphological Features", In Proc. International Conference on Frontiers of Handwriting Recognition, 2010, pp.658-663.