Writer Identification for Handwritten Telugu Documents using Directional Morphological Features

Pulak Purkait[1], Rajesh Kumar[2] and Bhabatosh Chanda[1]
[1] *ECSU, Indian Statistical Institute, Kolkata, India*
[2] *Directorate of Forensic Science, MHA, GOI, New Delhi, India*
$\{pulak\_r, chanda\}$*@isical.ac.in and rajesh.forensic@gmail.com*

Abstract

Linking a person based on handwritten documents is one of the oldest techniques that is used by crime investigators and forensic scientists. The importance of writer recognition in anthrax letter cases has made this examination popular in recent years. In this paper we propose four feature set namely directional opening, directional closing, direction erosion and k-curvature features for writer recognition on Telugu handwritten documents. Each of the features is extracted from the words after dividing them into a number of cells and then subjected to a nearest neighbor classifier for writer recognition. Although the results of each of the feature set is quite encouraging, the directional opening feature outperforms other feature sets.

Keywords: Writer identification, Telugu handwritten documents, directional morphological features, k-curvature .

1 Introduction

Linking a person based on handwritten documents is one of the most oldest techniques that is used by crime investigators and forensic scientists. It is evident from the literature [10] that Osborne(1929) [6] was one of the earliest researcher who has given a systematic approach for questioned document examination and writer recognition.

Writer identification is based on the hypothesis that every individual has its own writing habits and that is sufficiently different from others. This hypothesis was validated by different researchers starting from Osborne who has used Newcomb rule of probability to give a statistical basis for document examination [4]. In recent years, Srihari [13] has validated the hypothesis by doing an extensive experiment on the handwriting of 1500 individuals of different age groups, sex and educational background using image processing and pattern recognition techniques.

Writer identification methods can be divided into two major categories: based on text independent and text dependent strategies. In text independent strategies a code book of graphemes is generated from the text block and the features are extracted from these graphemes [1]. It is text independent in the sense that the features don't depend on the semantic content of the text. On the other hand text dependent approach is based on comparison of characters, words and/or line based on their semantic content. The text dependent approach is similar to a forensic document examiner approach and in the current study we follow the same.

In text dependent approach words are looking more powerful than characters as far as writer recognition is concerned, more discriminatory power of words than characters is also evident from the literatures [14, 15]. A writer may have the habit to write a character differently at different position in a word. To illustrate this phenomenon we take an example of English words. In Figure 1, one can see the difference in shape and design of 'a', 'r' and 't' in three different words although the writer for all the three words are the same. Variations in the design of these characters are the part of natural variation (intra-writer variation) of that particular writer. Thus if a writer recognition system is based on character features only, these words may be recognized as written by different writers and gives enough scope of error. Similar kind of thing is applicable for Telugu script also. Above argument gives sufficient reason to choose words against characters for the writer identification in the current study.

Research on writer identification is as old as handwriting itself. As far as automatic writer identification is concerned a comprehensive review of work done before 1989 is given in [8]. A survey of recent research in this field can be found in [1]. Although lots of research

978-0-7695-4221-8/10 $26.00 © 2010 IEEE
DOI 10.1109/ICFHR.2010.108

(a) (b) (c)

Figure 1. An illustration of variations of characters in different words written by same writer (a): 'art' (b): 'rat' and (c)'tar'

is going on in this field [11, 3, 16], there are two major issues that make this problem challenging : *natural variation* and *disguise*. Natural variation is variation in handwriting of an individual writer. The disguise is a kind of auto forgery by a writer to deny his or her writing later on for some benefits. In this paper we are not addressing any kind of forgeries but writer recognition based on the most natural writing habits of the writers.

With best of our knowledge, we didn't find any systematic work on Telugu script as far as writer recognition is concerned. That motivated us to work on Telugu script. Current study aims to give a systematic approach to the writer identification for Telugu script using morphological directional features and k-curvature features. After pre-processing for the scanned image of Telugu handwritten document, features are extracted from each word. Segmented words of similar content are size normalized and divided into a number of cells of fixed size. Features are extracted from each cell and the feature vectors of different words due to different writer is subjected to nearest neighbor classifier for identification purposes.

The rest of the paper is organized as follows. In Section 2, we describe about the handwritten Telugu data set and their pre-processing. Recognition scheme is proposed in Section 3. Results and discussions on the proposed method is placed in Section 4 followed by the concluding remarks and future directions of the present work in Section 5.

2 Data acquisition and pre-processing

An exemplar of Telugu data is prepared such that it contains almost every character of Telugu alphabet and also some frequently used words. 110 handwritten documents from 22 writers, 5 from each, is prepared. All the five documents of a single writer are written in his/her most natural writing habits.

Each document is digitized at 300 dpi using a flatbed scanner as a gray-scale image. The gray-scale document image is converted into binary using Otsu's

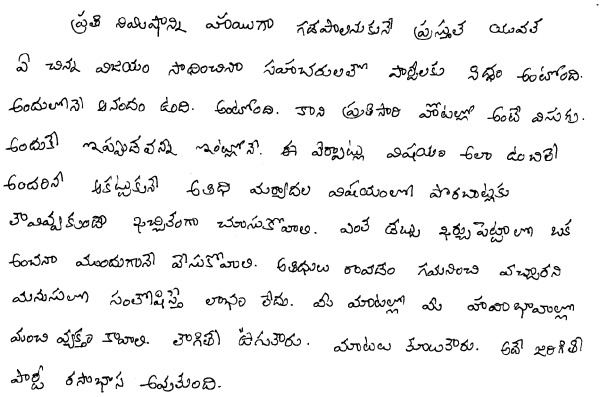


Figure 2. An exemplar of handwritten Telugu document

method [7]. Morphological filter [12] (opening followed by closing) is used for noise removal. While examining the document it is not necessary to take each and every word into consideration, even forensic document examiners search for the most frequent words with peculiarities. For this particular problem, we take first 10 words for further analysis. Each word image is thinned and then skeleton is dilated with 'square' structuring element of fixed radius to normalize the stroke width. Although we starts with 10 words, analysis and results for individual word too are given in subsequent sections.

3 Identification Framework

3.1 Feature Extraction

Line quality is one of the most important characteristics used by forensic document examiners to distinguish between different writers. Good line quality means smooth and round curve throughout the writing while the poor line quality is having tremors in strokes and rough curves throughout the handwriting. So line quality encompasses many features that stem from the dynamic processes used to guide the writing instrument as it moves across the paper. These elements include pen pressure, speed of execution, pen lifts, consistency and uniformity of the writing, rhythm and writing skills [5]. Writings that are exhibiting poor line quality is indicative of an inexperience writer or one who is suffering from physical and mental disability.

To capture the line quality (structural information) we propose four sets of features namely directional opening, direction closing, directional erosion and k-curvature [9, 2] based features. Since one can consider

that a curve is made up of small straight lines, thus morphological features basically give the information of type of curves which is involved to build a stroke of a particular word.

Arrangement of characters in a word is as important as other individuality like structure and design of writing strokes. The deviation of different characters from baseline may be considered as one of the most important features. To get the deviation of strokes from baseline we draw a compact bounding box of the word and divide it into a number of cells. The size and number of cell may vary according to the length of the word. The features are extracted from each cell of a word to capture the arrangement of strokes in a word. In this investigation we have tried different sizes of cell; however a cell of size 18×18 gives the best result among them. So, for the entire investigation we have taken cells of the size 18×18.

3.1.1 Directional Opening

The pre-processed image is opened in four direction using a 'line' structuring element of a size of twice the average width of the strokes and the normalized mass of the opened image in each cell is taken as a feature. Let I be the pre-processed word image and $[S^1, S^2, S^3, S^4]$ be the four 'line' structuring element in four direction (horizontal, vertical and two diagonal directions). Let N be the number of cells in which a particular word has been divided and ${m_j}^i$ be the mass of a word in j^{th} cell after opening I with the structuring element S^i. The normalized mass ${M_j}^i$ of that particular word in j^{th} cell after opening with S^i is given as $M_j^i = \frac{m_j^i}{\max_{j=1,\cdots,N}(m_j^i)}$. Thus for a word with number of cells N, we have a directional opening feature vector of length 4N.

3.1.2 Directional Closing

Instead of opening mentioned above, the pre-processed word image is closed with the same 'line' structuring elements and normalized mass of the closed image in each of the cell is taken as a feature. Normalization is also done in the same way. Thus for a word with number of cells N, a directional closing feature vector is also of length 4N.

3.1.3 Directional Erosion

Another morphological feature is obtained from the pre-processed skeleton image X. Skeleton image is eroded by the diagonal structuring elements shown in figure 3(a)-3(d), to get the number of co-occurrences of skeletal pixels in four different direction. Here structuring elements are $S^1 = \{(0,0),(1,1)\}$, $S^2 = \{(1,0),(0,1)\}$, $S^3 = \{(0,0),(1,0)\}$, and $S^4 = \{(0,0),(0,1)\}$. Each image $\hat{X}_i = X \Theta S^i, i = 1, ..., 4$ contains the pixels at which a particular directional co-occurrences on skeleton occur, where Θ is the morphological erosion operation. Then we divide each image $\hat{X}_i$ into N cells and make a normalized histogram of directional co-occurrences $(s_i^j, j = 1, ..., N)$ of portion of skeleton in each cell. In this case also we get a feature vector of length 4N.

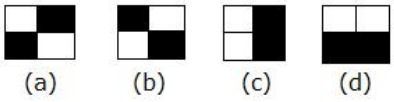


Figure 3. Four structuring element (a)-(d) with center as the first element used to find the images containing the portion of skeleton in the respective direction.

3.1.4 k-curvature features

The slope of chain code at any point is a multiple of $45°$, and the slope of a crack code is a multiple of $90°$. In order to arrest finer variation in slope, we use some type of smoothing over, say, k such chain code to define k-curvature.

Here left and right k-slopes at a point P on a curve are defined as the slopes of the line joining P to the points k steps away along the curve on each side of P, and the k-curvature of P as the difference between its left and right k-slopes. In this definition it is assumed that the curve from P to the fixed k-steps away can be approximated by a straight line segment. This assumption is better satisfied if k is small. On the other hand, in the case of small k, the slope may be influenced by small perturbation due to noise. So value of k is selected compromising these two conflicting characteristics of slope measure.

We calculate k-slopes at each point of the morphological skeleton of the object as the angle with respect to horizontal axis, and hence the k-curvature as acute angle between left and right segment at P.

We take $k = 4$ and observe that the value of k-curvature in most of the pixels lies between $90°$ and $180°$, with a bias towards $180°$. This suggests to finer bias in the interval after $90°$ at the time of quantization. So we quantize the interval $[0°, 180°]$ into five unequal bins as follows : $0° \leq \theta < 90°$, $90° \leq \theta < 120°$,

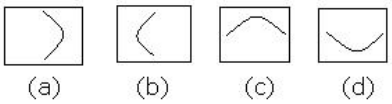


Figure 4. Four possibilities (a)-(d) where histogram of k-curvature are identical.

$120° \leq \theta < 140°$, $140° \leq \theta < 160°$ and $160° \leq \theta \leq 180°$. We make histogram on the basis of this distribution. Now there are four possibility as shown in the figure 4(a)-(d) which are different in terms of character shape but their histograms of k-curvature are nearly identical.

So we need to distinguish among the histogram of curve segments shown in Figure 4. The first two i.e. (a) and (b) and the rest two i.e. (c) and (d) can be separated by the difference of the row number and column number of the end points of the curve segments. Then they may be further distinguished by checking the location of mid-point with respect to the straight line joining the end points of the curve segments. Then we divide the image into cells of size 18×18 and calculate the normalized histogram having (5×4) 20 bins for each cell to get 20N dimension feature vector, where N is the number of cells into which the word image is divided.

3.2 Writer Identification

Writer Identification is a one to many comparison between a query document and the database documents associated with different writers. Features from a document is compared to features from the document of all the writers in the database and the inference is made against the comparisons. For writer identification we have followed 'Leave-one out' strategy. Out of 5 documents from each writer, one document is kept out for testing and remaining 4 documents are used as training data. All the combinations of 1 document out was tried and an average accuracy is reported in results and discussions section.

For writer identification nearest neighbor looks an obvious choice since the number of writers (class) may vary. We use Euclidean distance as similarity measure between different words. The writer of the query or test document is identified if the distance to the nearest neighbor is less than a predefined threshold.

3.3 Combining word-level performance for a document

Forensic document examiners, while examining the documents for writer identification look for individuality of different words and make a decision based on the combined performance (individuality) of different words in a document. Following a similar strategy to link a writer from a document we also combine the performances of different words in a document.

Let ${p_i}^j$ be the probability of a word w_j of a particular document A to be of written by the writer i. The probability of a document A having words $w_j, j = 1, 2, \cdots, k$ to be of writer i can be given as

$$P_i(A) = \prod_{j=1}^{k} {p_i}^j \tag{1}$$

The authorship of the document A is decided based on the maximum probability $P_i(A), i = 1, 2, \cdots, W$.

4 Results and Discussions

From Table 1, one can see that the directional opening feature set outperforms other feature sets. A maximum accuracy of 82.70% is obtained for word with word index 6. For that particular word (word index 6) every feature set gives the best performance in their respective class.

From the table, it is also evident that peculiarities of a word is more important than the length of the word. The longest words with word indices 5 and 2 could not perform well in comparison to the words with word indices 6 and 1 having more peculiar characters (first character i.e. 'pra' of these words) with lots of variations within writers. Table 3 shows inter writer variation of character 'pra' by six different writers. Variations shown in Table 3 support the higher accuracy of word with word indices 6 and 1.

Table 2 depicts the performance of nearest neighbor when the different number of words is combined to take decision on a document according to Section 3.4. Each of the four feature sets links more than 98% documents with their respective writers correctly when ten words is taken into account. Directional opening feature is the highest performer over all the four feature sets that links all the documents with their respective writers. Merely 5 words in combination gives an accuracy of more than 90% with each of the four feature sets.

Table 1. Identification results on individual words

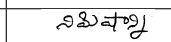
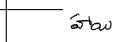
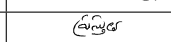
word index	word	Directional opening	Directional closing	Directional erosion	k-curvature
1		75.00	67.47	69.09	61.82
2		68.18	65.45	46.36	63.64
3		78.18	76.36	53.63	62.72
4		67.27	69.09	60.00	57.27
5		60.90	60.90	52.72	72.73
6		82.72	82.72	72.78	72.73
7		69.09	64.55	60.90	68.18
8		72.73	68.18	67.27	65.45
9		72.73	70.00	44.54	56.36
10		70.00	60.90	53.63	57.27

Table 2. writer identification on document level by combining the performance of individual word

No. of words	Directional opening	Directional closing	Directional erosion	k-curvature
1	71.73	68.55	58.09	63.82
2	86.89	84.32	72.18	79.49
3	93.12	91.73	81.53	87.65
5	98.21	97.38	90.91	94.71
7	99.29	98.67	95.03	97.64
10	100.00	99.09	98.18	99.09

Table 3. Inter writer variation of characters 'pra' by different writers

					
(a)	(b)	(c)	(d)	(e)	(f)

5 Conclusion and Future works

In this paper, we do a study on writer recognition for Telugu script. We propose four sets of feature namely directional opening, directional closing, directional erosion and k-curvature features. After noise removal, 10 words are segmented from each document. Segmented words of similar content are size normalized and divided into a number of cells of fixed size. All the feature sets are extracted from 10 different words of each of the document in the data set. Although the size of database, we are using, is small but it fulfills the objective of the study; to give a systematic approach to writer identification for Telugu script. The proposed features give an encouraging results on both word and document level but the directional opening outperforms other feature set.

Since we evaluate the proposed feature sets for a smaller database, the obvious future work may be evaluation of performance of proposed feature sets on a larger database. Moreover, we will continue our investigation to see the performance after combining all the feature set.

Acknowledgements

We would like to acknowledge Venkatappaiah Kurapati, Indian Statistical Institute, Kolkata for collection and preparation of digital Telugu handwritten documents.

References

[1] M. Bulacu and L. Schomaker. Text-independent writer identification and verification using textural and allographic features. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(4):701–717, 2007.

[2] B. Chanda and D. D. Majumdar. *Digital Image Processing and Analysis*. Prentiece Hall of India, New Delhi, 2005.

[3] K. Franke and M. Koppen. A computer-based system to support forensic studies on handwritten documents. *Pattern Recognition*, 33(1):133–148, 2000.

[4] R. A. Huber and A. M. Headrick. *Handwriting Identification: facts and fundamentals*. CRC Press, 1999.

[5] J. S. Kelly and B. S. Lindblom. *Scientific Examination of Questioned Documents*. Taylor and Francis, New York, 2006.

[6] A. S. Osborne. *Questioned Documents*. Boyd Printing Co., New York, 1929.

[7] N. Otsu. A threshold selection method from gray-level histogram. *IEEE Trans. System, Man and Cybernatics*, 6:62–66, 1979.

[8] R. Plamondon and G.Lorrete. Automatic signature verification and writer identification -the state of art. *Pattern Recognition*, 22(2):107–131, 1989.

[9] A. Rosenfeld and A. C. Kak. *Digital Picture Processing, Volume 2*. Academic Press, London, 1982.

[10] R. E. Saferstien. *Criminalistics: An Introduction To Forensic Science*. Prentice Hall, 2006.

[11] L. Schomaker and M. Bulacu. Automatic writer identification using connected-component contours and edge-based features of uppercase western script. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(6):787–798, 2004.

[12] P. Soille. *Morphological Image Analysis*. Springer Verlag, Berlin, 1999.

[13] S. N. Srihari, S. H. Cha, H. Arora, and S. Lee. Individuality of handwriting. *Journal of Forensic Sciences*, 47(4):1–17, 2002.

[14] C. I. Tomai, B. Zhang, and S.N.Srihari. Discriminatory power of handwritten words for writer recognition. In *Seventeenth International Conference on Pattern Recognition, Cambridge*, pages 638–641, 2004.

[15] B. Zhang and S.N.Srihari. Analysis of handwriting individuality using word features. In *Seventh International Conference on Document Analysis and Recognition, Edinburgh*, pages 1142–1146, 2003.

[16] E. N. Zois and V. Anastassopoulos. Morphological waveform coding for writer identification. *Pattern Recognition*, 33(3):385–398, 2000.