Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition, Oct. 2006, La Baule

Text Independent Writer Identification from Online Handwriting

Anoop Namboodiri and Sachin Gupta
International Institute of Information Technology
Gachibowli, Hyderabad, India
anoop@iiit.ac.in, sachin_g@students.iiit.ac.in

Abstract

Automatic identification of the author of a document has a variety of applications for both online and offline handwritten data such as facilitating the use of writer-dependent recognizers, verification of claimed identity for security, enabling personalized HCI and countering repudiations for legal purposes. Most of the existing writer identification techniques require the data to be from a specific text or a recognizer be available, which is not always feasible. Text-independent approaches often require large amount of data to be confident of good results. In this work, we propose a text-independent writer identification framework that uses a specified set of primitives of online handwritten data to ascertain the identity of the writer. The framework allows us to learn the properties of the script and the writers simultaneously and hence can be used with multiple languages or scripts. We demonstrate the applicability of our framework by choosing shapes of curves as primitives and show results on five different scripts and on different data sets.

Keywords: Biometrics, Individuality, Online Handwriting, Clustering

1. Introduction

With the increase in use of computers in every aspect of life, secure and automatic person identification is becoming an important problem. A robust method that can verify the identity of a person can help deter crime and fraud, while saving critical resources. Biometrics aims at automatic person identification based on the physiological and behavioral traits of a person, such as fingerprint, face, iris pattern, hand geometry, speech, etc. Person identification using handwriting, a behavioral biometric, is based on the hypothesis that people write uniquely and can be characterized based on the information present in their handwriting.

Automatic writer identification systems can be useful in a variety of applications including banks, criminal justice systems, determining the authenticity of handwritten emails, etc. Person identification based on handwriting is the natural choice to establish authorship of a handwritten document, which is either paper-based or electronic in origin. Traditionally, signature verification has been the most popular variant of writer identification, where the handwriting is restricted to a specific word, namely the signature of the writer. In contrast, generic writer identification algorithms try to establish the identity of the writer using either pre-specified or arbitrary text, provided by a writer.

The major challenge in any writer identification system arises from the variability in style, shape, size and consistency of the allographs written by a person. The variability between handwriting samples of a particular writer can increase further as the writing surfaces and conditions change. The problem is compounded by the fact that handwritings of different persons can look alike as they are essentially trying to write the same characters of the alphabet.

Handwritten documents can be primarily divided into two classes: offline and online documents. Offline documents are scanned images of handwritten documents. With the ease of availability and increase in use of Pocket PCs, Tablet PCs, and other pen enabled input devices, online documents are gaining popularity. Online documents contain the temporal information of the handwriting process in addition to the coordinates of pen movements, along with pen-up and pen-down events. Online handwriting allows us to use velocity, pressure and spatial information, which are not available with offline data. Online identification systems can be used in automated identity verification systems, such as ATMs and secure data access devices, where user can be authenticated based on a signature, name or password written using a stylus.

1.1. Background

Individuality information can be present at various levels in handwritten data. Huber and Headrick [14] identify five different levels of individuality information present in handwriting: i) subcharacter-level information embedded in the construction, design, and shape of allographs, ii) character-level information like slant and slope, iii) word-level information such as connection between characters, iv) line-level information such as arrangement of characters and, v) paragraph-level information contained in the organization of lines.

Character level information such as Gradient, Structural and Concavity (GSC) features have been used by Zhang et al. [1] based on the information about the character being examined. Tomai et al. extended this work

Table 1. Analysis of Previous work

Work	Data type	Features	Pros	Cons
Zhang et al. [1]	Offline	Character level GSC features	Accurate and robust	Required OCR to recognize characters, Text independent
Tomai et al. [3]	Offline	Character level(GSC), Word level features	Fast, Needs only few words	Requires OCR, Text dependent
Zuo et al. [7]	Offline	PCA based method	Fast and accurate	Requires OCR, Text dependent
Said et al. [5], He. et al. [6]	Offline	Multi resolution Gabor filter, Text as texture	Text independent, Accurate	Low level information lost, Requires large data set
Schlapbach et al. [8]	Offline	Text line based features, HMMs as classifier	Text independent, Not sensitive to noise and small shape variations	Large volume of data needed
Schomaker and Bulacu [11]	Offline	Connected Contour and edge based directional features	Text independent, Fast	Requires large data set
Pitak et al. [12]	Online	Velocity of Bary center, Fourier transform based approach	Text independent	Slow, Sensitive to Noise
Yashushi et al. [13]	Online	HMM based approach	Unaffected by small shape changes, Invariant to noise	Text dependant
Proposed work	Online	Allograph-level shape-based features	Text independent, Fast, Requires less test data	Only low level features used

to include shape curvature and shape context features that are computed at the word-level. However, the word-level features are often not sufficient to discriminate the writers, especially in offline documents.

A second approach to writer identification is to treat the image region containing the text as a texture and use the texture characteristics to identify the writer. Typical approaches from texture classification and object recognition domains, such asmultichannel Gabor filter based features [5, 6] and PCA-based feature extraction (at word-level) [7] have been used for this purpose. The texture-based approaches typically work at the paragraph level as it requires a region of the image for computation of features and classification.

Table 1 gives an overview of the previous work done in writer identification and its comparison to the current work.

The problem of writer identification can be divided into text dependent and text independent approaches. Text dependent approaches requires handwriting based on a specific text, or assume the availability of a handwriting recognizer for testing the authenticity of writer. Person identification using signature is most popular instance of these kind of approaches. The advantage of text dependent approaches is that they can use the knowledge of the content of the data to separate style from content. This increases accuracy of text dependent systems. The major problem with text dependent systems is that they are not applicable to cases where the text is not available, such as in criminal justice systems when text documents with different content need to be compared. Secondly, text dependent systems are more prone to forgery (such as replay attack) as same data is presented for testing. These type of systems can be implemented in the cooperative environment, where accuracy is the major concern and writer can be asked to write specific text to prove his identity.

On the other hand, text independent writer identification systems model the style information, independent of the content and can identify the writer based on any given text. This usually requires the use of statistics of features computed from a large quantity of data to avoid anomalies due to specific text. We propose a framework in which one can determine the identity of a writer based on small amount of data by using the distribution of basic primitives of writing that are learned from the training data.

The paper is organised as follows. Section 2 describes full framework for writer identification problem. Section 3 describes an example for extraction, representation of primitives with details of unsupervised clustering and classification algorithms. Section following this section contains details of all the experiments performed and results on different data sets.

2. Proposed framework

In this paper, we propose a generic text independent framework for writer identification. The basic idea of our framework is to automatically identify repeatitive primitives in handwritten data of a particular script, and then use the variations in those primitives to identify the writer. Hence our framework is applicable to any script where the specified primitives are present. The identification framework consists of the following components:

1. **Defining Primitives:** The primitives that are used for person identification could be any repeatitive structure or property of handwriting. Examples include shape primitives, allograph types, relationship between parts of handwriting, etc.

2. **Extraction and Representation:** Once the primitives to be used are defined, a mechanism to extract them from the data are identified. A consistent mechanism for primitive extraction is essential for ensuring consistency in the statistics derived from them, and hence the overall accuracy of the system. A clear definition of the primitive will lead to a simple extraction scheme. The primitives then need to be represented using a set of features for comparison.

3. **Similarity Measure:** A similarity or distance measure between two primitives need to be defined for comparison, based on the representation chosen in the previous step.

4. **Identifying Consistent Primitives:** Depending on the script under consideration, certain primitives repeat more often in various allographs. These primitives are identified using any unsupervised clustering algorithm. Pair-wise distances are computed for each pair of primitives extracted from the training samples and a distance-based clustering method, such as k-means clustering is used. The clusters thus formed are referred to as consistent primitives of the script.

5. **Writer Identification:** The final step is to use the between-writer variations within consistent primitives to determine the identity of the writer. This involves the design of a classifier for each of the consistent primitives (clusters) and then combining the results to get the most likely identity.

   Let S_j be the j^{th} primitive that was extracted from the data and C_k be the k^{th} cluster in the script. The data likelihood of the primitive S_j, given a particular writer W_i can be computed as:

$$p(S_j/W_i) = \sum_{k=1}^{N} p(S_j/W_i, C_k) * \alpha_{i,k}, \quad (1)$$

   where $\alpha_{i,k}$ is the weight of the kth cluster for the ith writer that quantifies the discriminability of the kth cluster for the ith writer.

   Now the complete data likelihood (for the document D), given writer W_i, can be computed from equation 1 as:

$$p(D/W_i) = \prod_{S_j \in D} p(S_j/W_i), \quad (2)$$

   The probability that the given document belongs to a writer can now be computed using Bayes rule from equation 2.

$$P(W_i/D) = \frac{p(D/W_i) * P(W_i)}{\sum_{i=1}^{n} p(D/W_i) * P(W_i)}. \quad (3)$$

   Equal prior probabilities are assigned to all writers.

3. Shape based Curve Extraction and Representation

As discussed earlier, five different levels of individuality information: subcharacter-level, character-level, word-level, line-level and paragraph-level, are present in handwritten data. Subcharacter level information includes design, construction and spatial distribution of curves present in the script. As subcharacter level information is text independent, we can use different subcharacter level information like size, shape, style for extracting the writer information. We demonstrate the effectiveness of the proposed framework using curve shapes as the basic primitive. The curve shape and size captures only part of the individuality information present at sub-character level. However, the results suggest that even the partial information can effectively distinguish between writers. Our framework allows for extension to multiple primitives for writer identification.

For extraction of shape primitive, velocity profile of the pen movement is used. According to kinematic theory of human movements, presented by Plamandon [16], human movements are combination of different forces and transition between these forces. In case of handwriting, a single force corresponds to the equi-curvature portion of handwritten stroke, between two minimum velocity points. Figure 1 shows dominent (maximum and minimum velocity) points, extracted using velocity profile of the stroke shown in Figure 2. Portion of the stroke between two consecutive minimum velocity points is shape curve. Two consecutive shape curves are used as basic primitive to exploit the individuality information present between transition from one shape curve to another.

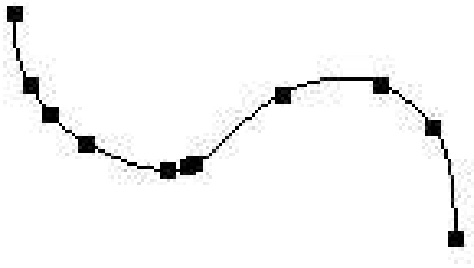

Figure 1. Dominant Points of Stroke

The third step is to devise consistent representation for shape primitive. A curve of constant curvature can be uniquely represented using three parameters: the incident direction, the curvature and size or length of the

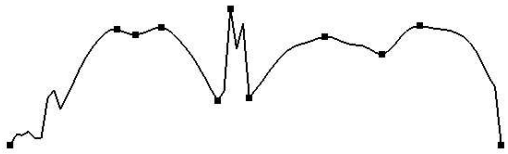

Figure 2. Velocity Profile of Stroke

curve [17]. Based on this principle, curve shapes are represented using angle of incidence, angle between corresponding vectors and size of the vectors. Figure 3 shows all the elements, used for representation of a particular shape-based primitive curve. Features 1–4 represent the incident angles and the curvature of each portion of the curve, while the other features represent the length of the curve. Thus each shape primitive is represented using an 8-dimensional feature vector. The representation constitutes an abstraction of the curve that is both direction and scale dependent.

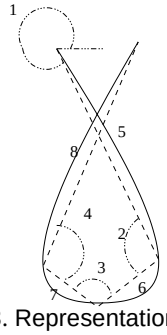

Figure 3. Representation of Curve

Since the representation of the curve is a fixed sized feature vector, a distance measure between two curves can be defined using Euclidean distance. To account for the variations in scales of the angular features and the length feature, we use a weighted Euclidean distance.

The distribution of shape primitive curves varies in different scripts. To identify repeatitive shape primitives present in the script, unsupervised k-means clustering algorithm is used. Ratio of within-cluster variance to between-cluster variance is used as cluster validation criteria. Figure 4 shows six primitive shape clusters extracted from Devanagari script.

To calculate the between-writer variation for consistent primitives, we design a classifier using the labeled training samples that falls in each of the k clusters. In this experiment, we have used Neural Network based classifier for classifying each curve primitive. The output of the classifier for each of the classes is used as the probability of observation of the curve, given the cluster and the writer. For each consistent primitive cluster different classifier is used. Equation 1 and equation 2 are used to calculate the log-likelihood of shape based primitives. Equation 3 is used to find out the probability of the writer given

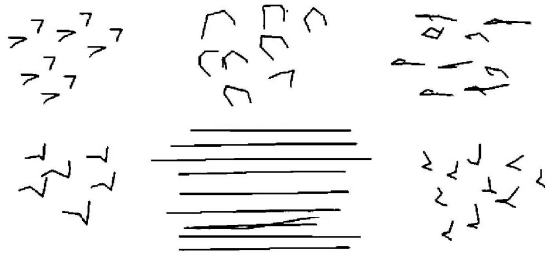

Figure 4. Different Clusters Extracted from Devanagari Script

the document. One could replace the classifier in each node with any other technique such as Gaussian models or kNN, as long as the classifier returns a confidence measure for the given curve.

4. Experiments and Results

Experiments were performed on 5 different scripts; Devanagari, English, Cyrillic, Arabic and Hebrew. For each script, experiments were performed for 10 to 12 writers. Data was collected using IBM CrossPad. Each user was asked to write out any text in particular script on a letter sized paper, that was captured electronically by the CrossPad. Data was divided randomly into four parts and at every step, three parts of the data were used for training and the remaining part for testing.

The data was smoothened using a Gaussian low-pass filter prior to training and testing, to remove any noise added due to pen vibration. Around 700 instances of basic shape primitives are extracted from the training data of each writer.

Three different sets of experiments were performed to determine the variation in accuracy of the identification scheme: i) variation as data size varies ii) variation as number of writers increase and iii) variation with different scripts under consideration. First two set of experiments were performed only for Devanagari script as we had more data available for it.

For the first two experiments, around 700 curves were extracted from Devanagari data collected from 10 different writers. The data was clustered into 16 clusters (experimentally chosen) and the classifiers were trained on each of these clusters. Ratio of within cluster variance to between cluster variance was used as cluster validity criterion. Data was varied starting from 10 curves (approx. 1 word) to 300 curves (approx. 25 words). With around 200 curves, accuracy of 80% is achieved. Experiments are performed 30 times for each data size with different set of data. figure 5 shows the accuracy variation with variation in test data size.

As seen from the figure 5 accuracy of classifier increases, as more test data is available. Only $20-30$ curves (approx. $2-3$ words) are required to identify 60% of the

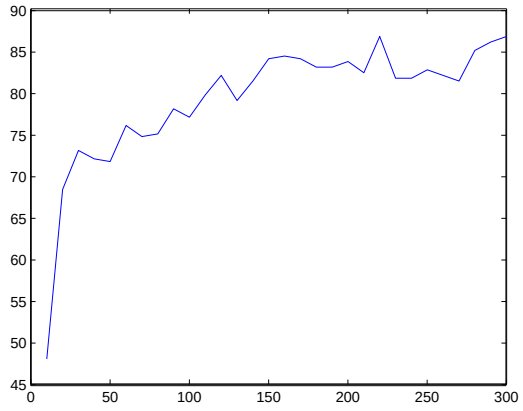

Figure 5. Test data Vs Accuracy

samples correctly, and with 220 curves (approximately 12 words) probability of correct classification increased up to 87%. Accuracy of 87% is reported using only single primitive, as more and more primitives will be used accuracy will increase.

Second set of experiments were performed to determine the effectiveness of our algorithm to classify handwritten data from multiple scripts. For each script, the set of primitive clusters are usually different and hence need to be trained separately. However, the overall procedure remains the same for all scripts. Table 2 shows the accuracy of each script with number of writer. For all the scripts, the Top-2 accuracy approaches 100%. For all the scripts other than Roman approximate 700-800 curves are used for training and approx 100 curves (approx. 10 words) for testing. For Roman script around 400 curves are used for training.

Table 2. Script Vs Accuracy

Script	No. of Writer	Top-1 Accuracy	Top-2 Accuracy
Devanagari	10	87	100
Roman	6	83	88
Cyrillic	10	80	100
Arabic	15	85	97
Hebrew	10	90	100

Third set of experiments were performed to determine the effect of the different number of writers on accuracy. This set of experiments were also performed using Devanagari data set. Table 3 shows the variation in accuracy as number of writers increased from 2 to 10. From each writer approx. 700 curves were taken as training data and 100 (approx. 10 words) were taken as testing data.

Shape based primitive proved a bad choice for Chinese Script, as most of the shapes extracted were straight lines, that do not contain much individuality information. Only 50% accuracy could be acheived. However, one could identify a different set of primitives and different representation scheme to rectify this problem. For Chinese and Roman scripts, in which lots of shape based primitives are straight lines, size based primitive, like ratio of size between consecutive primitive curves can be used. Experiments are being performed to check the confidence measure of size based primitives for different scripts.

Table 3. Accuracy Vs Different Number of Writers

No of Writers	Accuracy
2	100
3	99.8
4	99.6
5	97.0
6	98.0
7	98.3
8	97.0
9	92
10	87

5. Conclusion

In this paper we proposed a text independent writer Identification system for online documents. Advantage of this method include the need of small amount of test data in addition to being text independent. The classification is fast and we can improve our confidence in the results as the data size increases (evidence accumulation). Even with one line of data we can get a high confidence about the identity of the writer. We have used sub character level features for writer identification.

To improve on the accuracy and robustness of the system, for the script like Chinese and Roman, in future we can use other high level primitives based on character,word, line and paragraph. Different primitives like shape, size and other higher level features can also be used in combination to improve the system. More robust representation, like spline, can be used for shape primitives. More Robust cluster validity criteria can be used for cluster validity.

References

[1] Bin Zhang, Sargur N. Srihari and Sangjik Lee, "Individuality of Handwritten Characters", *Proceedings of the 7th International Conference on Document Analysis and Recognition*, Edinburgh, Scotland, August 3-6, 2003, pp 1086-1090.

[2] Bin Zhang and Sargur N. Srihari, "Analysis of Handwriting Individuality Using Word Features", *Proceedings of the 7th International Conference on Document Analysis and Recognition*, Edinburgh, Scotland, August 3-6, 2003, pp 1142-1146.

[3] Catalin I. Tomai, Bin Zhang and Sargur N. Srihari, "Discriminatory Power of Handwritten Words for Writer Recognition", *Proceedings of the 17th International Conference on Pattern Recognition*, Cambridge, UK, 2004, Vol. 2, pp 638-641.

[4] Sargur Srihari, Matthew J. Beal, Karthik Bandi, Vivek Shah and Praveen Krishnamurthy, "A Statistical Model For Writer Verification", *Proceedings of International Conference of Document Analysis and Recognition*, Seoul, Korea, August 2005, pp 1105-1109.

[5] H. E. S. Said, G. S. Peake, T. N. Tan and K. D. Baker, "Writer Identification from Non-uniformly Skewed Handwriting Images", *Proceedings of British Machine Vision Conference*, Southampton, UK, 1998, pp 478-487.

[6] Z.Y. He and Y.Y. Tang, "Chinese Handwriting Based Writer Identification by Texture Analysis", *Proceedings of the 3rd International Conference on Machine Learning and Cybernetics*, Shanghai, 26-29 August, 2004, pp 3488-3491.

[7] Long Zuo, Yunhong Wang and Tieniu Tan, "Personal Handwriting Identification Based on PCA", *Proceedings of SPIE Second International Conference on Image and Graphics*, July 2002, pp 766-771.

[8] Andreas Schlapbach and Horst Bunke, "Off-line Handwriting Identification Using HMM Based Recognizers", *Proceedings of the 17th International Conference on Pattern Recognition*, Cambridge, UK, 2004, pp 654-658.

[9] Andreas Schlapbach and Horst Bunke, "Using HMM Based Recognizers for Writer Identification and Verification", *Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition* , Kokubunji, Tokyo, Japan, October, 2004, pp 167–172.

[10] Marius Bulacu, Lambert Schomaker and Louis Vuurpijl "Writer Identification Using Edge-Based Directional Features", *Proceedings of the 7th International Conference on Document Analysis and Recognition* , Edinburgh, Scotland, August 3-6, 2003,pp 937-941.

[11] Lambert Schomaker and Marius 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, No. 6, June 2004 pp 787-798.

[12] Pitak Thumwarin and Takenobu Matsuura "On-line Writer Recognition for Thai Based on Velocity of Barycenter of Pen-point Movement", *Proceedings of IEEE International Conference on Image Processing*, Singapore, October 24-27, 2004, pp 889-892.

[13] Yasushi Yamazaki,Tsuyoshi Nagao and Naohisa Komatsu "Text-indicated Writer Verification Using Hidden Markov Models", *Proceedings of the 7th International Conference on Document Analysis and Recognition*, Edinburgh, Scotland, 2003, pp 329-332.

[14] Huber R.A and Headrick A.M., *Handwriting Identification: Facts and Fundamentals*, Boca Roton, CRC Press,1999.

[15] R. jean Plamondon, "A kinematic theory of rapid human movements. Part-I Movement representation and generation", *Biological Cybernetics*, 1995.

[16] R. jean Plamondon, "A kinematic theory of rapid human movements. Part-II. Movement time and control", *Biological Cybernatics* , 1995.

[17] R. jean Plamondon and Sargur N. Srihari, "On-Line and Off-Line Handwriting Recognition:A Comprehensive Survey", *IEEE Transactions on Pattern Analysis and Machine intelligence*, VOL. 22, NO. 1, Amsterdam, Sep 11-13 2000, pp 333-342.

[18] Kun Yu, Yunhong Wang and Tieniu Tan, "Writer Authentication Based on the Analysis of Strokes", *Proceedings of SPIE: Biometric Technology for Human Identification*, Vol. 5404, pp. 215–224.