Individuality of Handwritten Characters

Bin Zhang Sargur N. Srihari Sangjik Lee
CEDAR, Computer Science and Engineering Department
State University of New York at Buffalo, Buffalo, NY 14228
Email: {binzhang, srihari, sjl1}@cedar.buffalo.edu

Abstract

Analysis of handwritten characters (allographs) plays an important role in forensic document examination. However, so far there lacks a comprehensive and quantitative study on individuality of handwritten characters. Based on a large number of handwritten characters extracted from handwriting samples of 1000 individuals in US, the individuality of handwritten characters has been quantitatively measured through identification and verification models. Our study shows that in general alphabetic characters bear more individuality than numerals and use of a certain number of characters will significantly outperform the global features of handwriting samples in handwriting identification and verification. Moreover, the quantitative measurement of discriminative powers of characters offers a general guidance for selecting most-informative characters in examining forensic documents.

1 Introduction

Analysis of handwritten characters (allographs) plays an important role in forensic document examination [3]. In the past twenty years, many efforts have been made to employ handwritten characters for writer identification and verification, referred as text-sensitive writer authentication by Plamondon and Lorrette [7].

The efforts using characters for authentication can be grouped into two categories depending on feature extraction methods, i.e., transform based approach and structural approach. While Mihelic et al. took as features the coefficients of Walsh-Hadamard transform [5] and Kuckuck [4] analyzed the Fourier spectrum of handwriting images, Naske [6] used deformation analysis and a whitening transformation to extract features from characters. These methods belong to transform based approach. Under the frame of structural approach, Dinstein [1] used a trigonometric-representation based method to describe the line tracing results on a character, Yoshimura et al. [10] extracted direction and arc features from characters for identifying Japanese writing, and Kuckuck [4] used the curvature statistics as features.

Recent comprehensive research in handwriting individuality has also shown the high discriminative power of handwritten characters [8]. In [8], a handwritten document is characterized by twelve floating macro-features for parameterizing against document and paragraph levels, and 5120 binary micro-features extracted from ten characters (eight characters in a word "referred" plus 'b' and 'h'). Based on testings over handwriting samples of 1500 individuals in U.S., use of ten characters alone gives 82% identify accuracy and 91% verification accuracy when 900 writers were considered.

As a continuation of the research in [8], we have segmented 62 characters from each handwritten document for studying handwriting individuality. Based on a large number of handwritten characters extracted from more than 3000 handwriting samples of over 1000 writers, the individuality of handwritten characters has been quantitatively measured through identification and verification models. The quantitative evaluation of individuality of handwritten characters from this research provides a general guidance for selecting most-informative characters in examining forensic documents.

The rest of the paper is organized as follows. In Section 2, we introduce the binary micro-features for handwritten characters. In Section 3, we define similarity measures for k-nearest neighbor classification. In Section 4, we describe experimental settings for writer identification and verification. In Section 5, we present the experimental results and analysis. We draw conclusions in Section 6.

2 Feature Extraction

Given a handwriting sample, a set of characters are first segmented, then for each isolated character, the so-called micro-features are extracted. Therefore, each handwriting sample is characterized by a number of micro-feature vectors corresponding to the characters available from the sam-

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
0-7695-1960-1/03 $17.00 © 2003 **IEEE**

ple.

Micro-features have been successfully used for recognizing handwritten characters [2] and analyzing handwriting individuality [8]. For an individual character, the micro-features consist of 512 bits corresponding to gradient (192 bits),structural (192 bits), and concavity (128 bits) features. Each of these three sets of features rely on dividing the scanned image of the allograph (character or combination of characters) into a 4 x 4 region. The gradient features capture the frequency of the direction of the gradient, as obtained by convolving the image with a Sobel edge operator, in each of 12 directions and then thresholding the resultant values to yield a 192-bit vector. The structural features capture, in the gradient image, the presence of corners, diagonal lines, and vertical and horizontal lines, as determined by 12 rules. The concavity features capture, in the binary image, major topological and geometrical features including direction of bays, presence of holes, and large vertical and horizontal strokes.

3 Classifier Design

Two different models, identification and verification, are used to study the individuality of handwritten characters. Writer identification is a task of determining the writership of a handwriting sample, and writer verification concerns about whether two handwriting samples were written by the same writer or by two different writers.

In identification model, a number of 512-dimensional binary micro-feature vectors are associated with each handwritten document, whereas, in verification model, a real-valued distance vector (each component represents the distance between two characters each featured by a 512-dimensional binary vector) is used to describe the difference between a pair of documents. Thus, it is critical to adopt proper similarity measures in order to maximize the performance of handwriting analysis. We differentiate two types of features, binary micro-features and real-valued distance features. In the following, we define the similarity functions for the two types of features and for the combinations of these features.

3.1 Similarity Measure for Binary Vectors

Let Ω be the set of all n-dimensional binary vectors. To measure the similarity between two binary vectors, we use the Correlation measure, one of the eight measures summarized by Tubbs [9]. This measure has shown very good performance in handwriting identification using binary features [11].

Let S_{ij} $(i, j \in \{0, 1\})$ be the number of occurrences of matches with i in the first pattern and j in the second pattern at the corresponding positions. Given two binary feature vectors $X \in \Omega$ and $Y \in \Omega$, each similarity measure $S(X, Y)$ above uses all or some of the four possible values, i.e., S_{00}, S_{01}, S_{10} and S_{11}. We define a dissimilarity measure $D^b(X, Y)$ corresponding to the Correlation measure as:

$$D^b(X,Y) = \frac{1}{2} - \frac{S_{11}S_{00} - S_{10}S_{01}}{2(S_{10}+S_{11})(S_{01}+S_{00})(S_{11}+S_{01})(S_{00}+S_{10})^{1/2}} \quad (1)$$

3.2 Similarity Measure Functions for Heterogeneous Features

For the identification model and verification model, we use different similarity functions to combine feature vectors from characters and distance vectors from character pairs.

For the identification model, given two documents, A and B, with l pairs of same-class characters available, the distances between the character pairs, $d_i^b, i = 1, 2, ..., l$, are calculated according to (2). The combined distance is used to characterize the difference between A and B, given by

$$D(A,B) = \frac{1}{l}\sum_{i=1}^{l} d_i^b \quad (2)$$

In the verification model, we use the Euclidean metric weighted by deviations of features to measure the distance between two distance vectors.

3.3 Classification Techniques

A two-stage k-nearest-neighbor search and an artificial neural network were implemented for handwriting identification and verification respectively [8]. However, our recent experiments revealed that the two-stage k-nearest-neighbor search doesn't necessarily outperform the single-stage nearest-neighbor search. As neural network based classification requires the fixed number of features from all patterns, it is not applicable in the case of missing features. However, the reality for handwritten documents is that some features or some characters are often unavailable, thus for both handwriting identification and verification, we proposed the k-nearest neighbor classification by using the aforementioned similarity functions. Specifically, for the identification model we use nearest neighbor classification based on the similarity function (1) and for the verification model we employ distance-weighted k-nearest-neighbor classification based on the 2-nd order weighted Minkowski metric.

4 Experimental Settings

Handwriting identification was performed on 3081 documents written by 1027 writers in US. Each writer copied three times a source document specially designed by CEDAR [8]. The testing set consists of 875 randomly selected documents written by 875 writers randomly chosen from 1027 writers, the training set includes the remaining 2206 documents.

As the verification model is to verify whether two documents were written by the same writer or two different writers, the testing and training sets consist of within-writer and between-writer distance vectors. The handwriting verification was tested on 3000 documents written by 1000 writers in US. The 1000 writers are partitioned into two groups, each with 500 writers. Each group has 1500 documents. From each group, we choose a number of document pairs written by the same writers and different writers to constitute either a testing set or a training set, shown as follows.

For a group with 1500 documents written by 500 writers, let Φ be the set of 1500 pairs of documents written by the same writers and Θ be $C^1_{500}C^1_3C^1_{499}C^1_3/2 = 1,122,750$ pairs of documents by different writers. A verification testing set consists of all elements in Φ and 1500 elements randomly chosen from Θ; A verification training set consists of all elements in Φ and 1500 elements randomly chosen from Θ.

For each document, 62 characters, ('0'$\sim$'9', 'a'$\sim$'z', and 'A'$\sim$'Z'), are used as sample allographs. Micro-features are extracted from these characters. Therefore, two document pairs may have different number of same-class character pairs, the aforementioned similarity functions are specially designed to handle the variation of same-class character pairs.

In the next section, we will examine the identification and verification performance under following scenarios:

- Individual characters.
- Accumulated characters: start from the numeral '0', add one more character at a time in the order of ('0'$\sim$'9', 'a'$\sim$'z', 'A'$\sim$'Z').
- Four combination schemes of numerals and alphabet: C1 for using all 62 characters, C2 for 10 numerals ('0'$\sim$'9'), C3 for 8 numeral ('2'$\sim$'9'), C4 for 52 alphabetic characters.

5 Handwriting Identification and Verification

We present experimental results with regard to the aforementioned scenarios, followed by analysis and discussion.

5.1 Identification Results

Figure 1(a) shows the identification performance of individual characters. It's easy to find that in general handwrit-

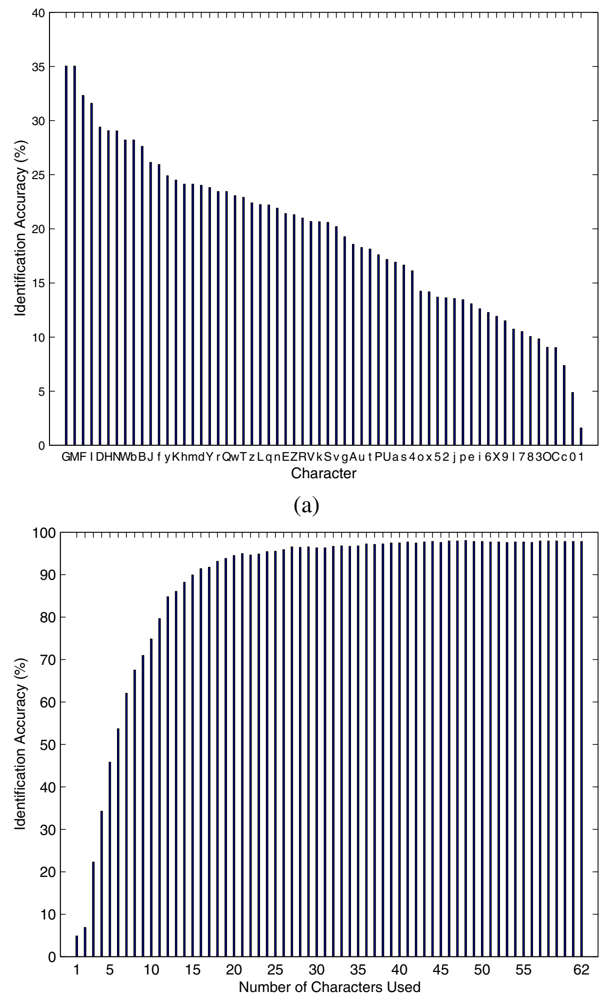

Figure 1. Handwriting identification performance using characters: (a) individual characters; (b) accumulated characters in the order of '0'$\sim$'9', 'a'$\sim$'z', 'A'$\sim$'Z'.

ten numerals have lower identification power than handwritten alphabetic characters. The numeral '1' has the least performance of 1.6% and the numeral '0' comes as the second worst (4.88%). As expected, some simples alphabetic characters like 'C' and 'X' also present low individuality. Different characters have very different discriminative power of handwriting individuality.

Figure 1(b) shows the identification performance of accumulated characters. A combination of the 10 numerals ('0'$\sim$'9') can only differentiate less than 74.86% of writers. However, a combination of the 10 numerals and 10 alphabetic characters ('a'$\sim$'j') can correctly identify more than 94.5% writers. Use of all 62 characters leads to a high

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
0-7695-1960-1/03 $17.00 © 2003 IEEE

identification accuracy of 97.83%.

Among the four combination scenarios, the scenario C1 (all 62 characters are employed for identification) has the best performance of 97.83%. The identification accuracy with the numerals (the scenarios C2 and C3) is less than 75%. A high accuracy of 97.71% is achieved with use of 52 alphabetic characters (C4). However, the early work [8], where each handwritten document is characterized by twelve document-level features and ten characters, can correctly identify only 87.3% writers when 900 writers are considered.

With the same training and testing sets, eleven global macro-features described in [8] can correctly identify only 65.94% of 875 writers. Thus, handwritten characters show much higher discriminative power than those global features.

5.2 Verification Results

Figure 2(a) gives the verification performance of individual characters. The observation from Figure 2(a) leads to the similar conclusions drawn from 1: (i) the handwritten numerals have lower identification power than handwritten alphabetic characters, (ii) the numeral '1' has the least performance of 61.79% and the numeral '0' with a verification accuracy 66.55% comes as the second worst, (iii) some simple alphabetic characters like 'C' and 'X' also present low individuality.

Figure 2(b) shows that the accumulation of handwritten numerals has relatively low verification accuracy (less than 89.46%). Use of alphabetic characters greatly improves the verification performance (with an accuracy 96.10%). The verification accuracy with all 62 characters can top 96.40%.

The experiments with different combination scenarios show that the scenarios C1 (96.40%) and C4 (96.10%) have similar performance and the combinations of numerals have low verification performance (89.46% and 89.05% for the scenarios C2 and C3 respectively). This means that the handwritten numerals does not help much for writer verification when there are sufficient number of alphabetic characters.

5.3 Disriminability Measure of Individuality of Characters

Certainly, discriminability measure of individuality of each character can be computed from the corresponding identification and verification accuracies, but this method is classifier-dependent, thus lacks inferrability. We propose a non-parametric and classifier-independent method to measure the discriminability of each character by using the associated receiver operating characteristic (ROC) curve.

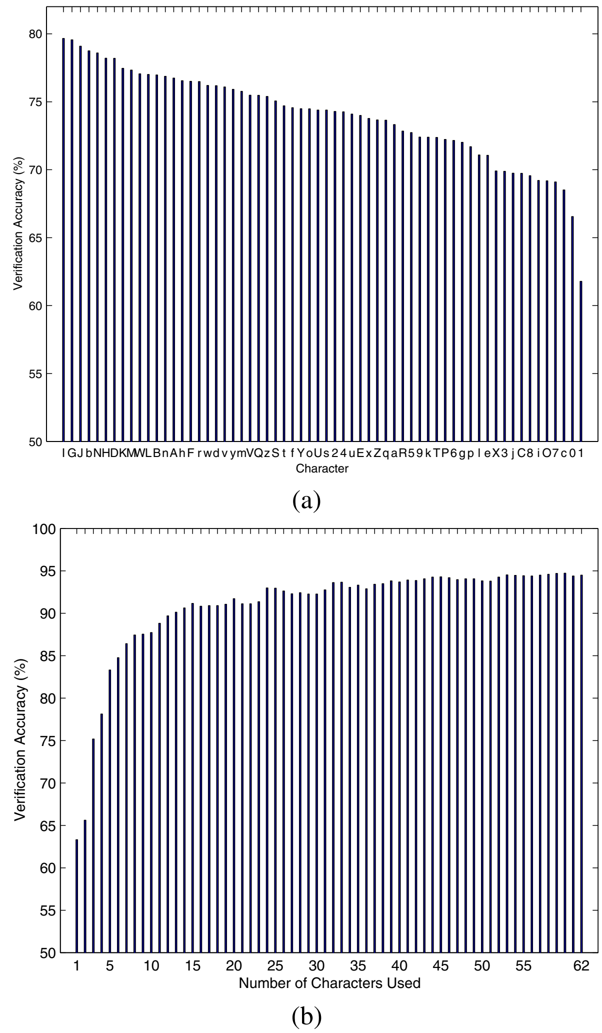

Figure 2. Handwriting verification performance using characters: (a) individual characters; (b) accumulated characters in the order of '0'$\sim$'9', 'a'$\sim$'z', 'A'$\sim$'Z'.

The discriminability measure $\wp$ is obtained by (i) computing the probability distributions of the distance between two character samples conditioned on whether the samples belong to the same individual or to different individuals, and (ii) constructing the ROC curve for each character from the two distributions, and (iii) determining the area under each ROC curve for each character.

Figure 3 shows the ranking order of 62 characters by discriminability. The top ten characters with least discriminative power (less than 0.35) are '1', '0', 'X', 'x', 'C', 'O', 'c', 'i', '3', and '7'. The eight most informative characters (with $\wp \leq 0.41$) are 'G', 'b', 'N', 'I', 'K', 'J', 'W', 'D', 'h', and 'F'. As expected, '0' (with $\wp = 0.305$) and '1' (with $\wp = 0.265$) are least discriminative.

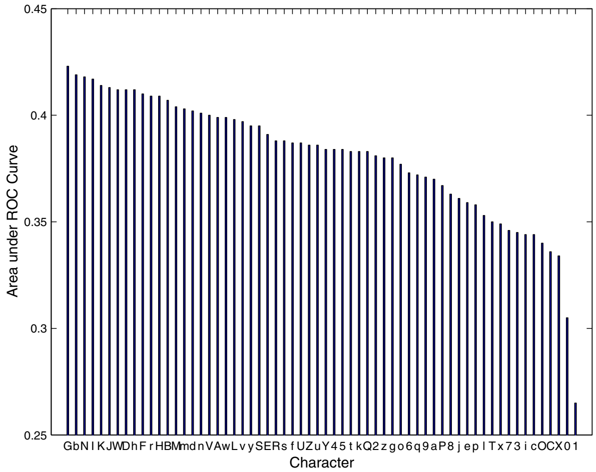

Figure 3. Ranking characters by discriminability measure based on area under ROC curve.

Several general conclusions can be made from the discussions and observations:

(a) Use a large number of handwritten characters presents higher discriminative power of handwriting individuality than document-level features.

(b) Different handwritten characters have different power of discriminating handwriting individuality.

(c) Handwritten alphabetic characters are much more powerful in discriminating handwriting individuality than handwritten numerals.

(d) Handwritten numerals don't tell much about handwriting individuality, and the most frequently used numerals, '0' and '1', have the least individuality.

(e) Handwritten characters are ranked in descending order of individual discriminative power: 'G'; 'b'; 'N'; 'I'; 'K'; 'J'; 'W'; 'D'; 'h'; 'F'; 'r'; 'H'; 'B'; 'M'; 'm'; 'd'; 'n'; 'V'; 'A'; 'w'; 'L'; 'v'; 'y'; 'S'; 'E'; 'R'; 's'; 'f'; 'U'; 'Z'; 'u'; 'Y'; '4'; '5'; 't'; 'k'; 'Q'; '2'; 'z'; 'g'; 'o'; '6'; 'q'; '9'; 'a'; 'P'; '8'; 'j'; 'e'; 'p'; 'l'; 'T'; 'x'; '7'; '3'; 'i'; 'c'; 'O'; 'C'; 'X'; '0'; '1'.

Notice that in this paper all character images belonging to the same class were segmented from the word images of the same content and they have the same relative position in the word images and handwriting samples. However, it has been shown that the discriminative power of a character may vary with its position in word and content of word [11]. As the exhaustive exploration of so many variations requires huge effort, if not impossible, we can only use limited experiments to estimate discriminative power of characters. From this respective, the conclusions above are significant.

6 Conclusions

The power of sixty-two handwritten characters to distinguish between individuals has been quantitatively measured through writer identification and verification on a large number of handwriting samples from representatives of the U.S. population. The ranking order of characters in discriminative power of individuality, for the first time, provides a general guidance for selecting most-informative characters in examining forensic documents.

Acknowledgments

This work was supported in part by the Department of Justice, National Insititute of Justice grant 2002-LT-BX-K007."

References

[1] I. Dinstein and Y. Shapira. Ancient hebraic handwriting identification with run-length histograms. *IEEE Trans. Syst. Man Cyber.*, SMC-12:405–409, 1982.

[2] J. T. Favata and G. Srikantan. A multiple feature/resolution approach to handprinted digit and character recognition. *International Journal of Imaging Systems and Technology*, 7:304–311, 1996.

[3] R. A. Huber and A. M. Headrick. *Handwriting Identification: Facts and Fundamentals*. CRC Press LLC, Boca Raton, FL, 1999.

[4] W. Kuckuck. Writer recognition by spectrum analysis. In *Proc. 1980 Int. Conf. Security through Sci. Engin.*, pages 1–3, West Berlin, 1980.

[5] F. Mihelic, N. Pavesic, and L. Gyergyek. Recognition of writer of handwritten texts. In *Proc. 1977 Int. Conf. on Crime Countermeasures - Sci. Engin.*, pages 237–240, University of Kentuky, Lexington, 1977.

[6] R. D. Naske. Writer recognition by prototype related deformation of handprinted characters. In *Proc. 6th Int. Conf. on Pattern Recognition*, pages 819–822, Munich, 1982.

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

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

[9] J. D. Tubbs. A note on binary template matching. *Pattern Recognition*, 22(4):359–365, 1989.

[10] I. Yoshimura and M. Yoshimura. Off-line writer verification using ordinary characters as the object. *Pattern Recognition*, 24(9):909–915, 1991.

[11] B. Zhang and S. N. Srihari. Binary vector dissimilarities for handwriting identification. In *Proceedings SPIE, Document Recognition and Retrieval IX*, pages 155–166, Santa Clara, CA, January 2003.

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
0-7695-1960-1/03 $17.00 © 2003 **IEEE**

IEEE COMPUTER SOCIETY