Accepted by the 7th International Conference on Document Analysis and Recognition, Edinburgh, Scotland, August 3-6, 2003. (Paper ID: 526)

Analysis of Handwriting Individuality Using Word Features

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

Abstract

Analysis of allographs (characters) and allograph combinations (words) is the key for obtaining the discriminating elements of handwriting. While allographs usually inhabit in words and segregation of a word into allographs is more subjective than objective, especially for cursive writing, analysis of handwritten words is a natural and better option. In this study, a handwritten word image is characterized by gradient, structural, and concavity features, and individuality of handwritten words is experimented through writership identification and verification on over 12,000 word images extracted from 3000 handwriting samples of 1000 individuals in U.S.. Experimental results show that handwritten words are very effective in differentiating handwriting and carry more individuality than most handwritten characters (allographs).

1 Introduction

Analysis of allographs (characters) and allograph combinations (words) is the key for obtaining the discriminating elements of handwriting [3]. For computerized writer authentication, many approaches have been proposed to extract features from handwritten characters and words.

In the approaches using handwritten characters, there are mainly two types of features, i.e., transform-based features [4, 5, 6], and structural features [1, 4, 8, 11]. As characters usually inhabit in words and segregation of a word into allographs is more subjective than objective, especially for cursive writing, use of handwritten words for studying handwriting individuality is a natural choice. Individuality analysis of handwritten words is similar to signature verification. Key techniques for off-line signature verification have been summarized in [7]. Recently, Zuo et al. [12] proposed a writer-identification algorithm based on Principal Component Analysis (PCA) of a set of characteristic Chinese words. This method directly works on gray-scale word images and demonstrates high identification accuracy with 400 handwriting samples from 40 writers. In a more comprehensive research on handwriting individuality [8], very high identification and verification performance is promised over a large number of handwriting samples by 1500 individuals, representative of the U.S. population, when the eleven global macro-features from handwriting samples and the local micro-features from characters are both employed, however, the macro-features from words presented very low identification rate. So far, there lacks an effective and comprehensive work on examining the discriminative power of handwritten words.

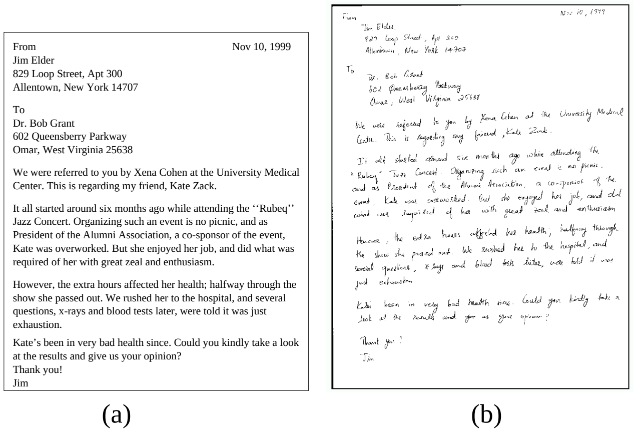
This study is a first step towards establishing objective measurement of individuality of handwritten words, as well as a complement to the work in [8]. In the study, we present an effective algorithm to extract the gradient, structural, and concavity features from handwritten word images and examining the individuality of four words through writership identification and verification on over 12,000 word images, extracted from 3081 handwriting samples of 1027 individuals in U.S.

The rest of the paper is organized as follows. In Section 2, we describe data collection. In Section 3, we present the algorithm for extracting handwritten word features. In Section 4, we define similarity measures for k-nearest neighbor classification. In Section 5, we describe experimental settings for writer identification and verification. In Section 6, we present the experimental results and analysis. We draw conclusions in Section 7.

2 Handwriting Word Samples

A source document in English, which was to be copied by each writer, was designed for the purpose of this study (Figure 1). It is concise (156 words) and complete in that it captures all characters (alphabets and numerals) and certain character combinations of interest. Each participant (writer) was required to copy-out the source document three times in his/her most natural handwriting, using plain, unruled sheets, and a medium black ballpoint pen provided by us. The repetition was to determine, for each writer, the variation of handwriting from one writing occasion to the next.

The study focuses on four characteristic words, “been”, “Cohen”, “Medical”, and “referred”. For each handwriting sample, four images, corresponding to the four characteristic words, are manually segmented. Figure 2 shows the copies of four characteristic words provided by three writ-



From Nov 10, 1999
Jim Elder
829 Loop Street, Apt 300
Allentown, New York 14707

To
Dr. Bob Grant
602 Queensberry Parkway
Omar, West Virginia 25638

We were referred to you by Xena Cohen at the University Medical Center. This is regarding my friend, Kate Zack.

It all started around six months ago while attending the ‘‘Rubeq’’ Jazz Concert. Organizing such an event is no picnic, and as President of the Alumni Association, a co-sponsor of the event, Kate was overworked. But she enjoyed her job, and did what was required of her with great zeal and enthusiasm.

However, the extra hours affected her health; halfway through the show she passed out. We rushed her to the hospital, and several questions, x-rays and blood tests later, were told it was just exhaustion.

Kate’s been in very bad health since. Could you kindly take a look at the results and give us your opinion?
Thank you!
Jim

Figure 1. Handwriting Exemplar: (a) source document to be copied by writers, and (b) a digitally scanned handwritten sample provided by writer.

ers, each of whom wrote the same word three times.

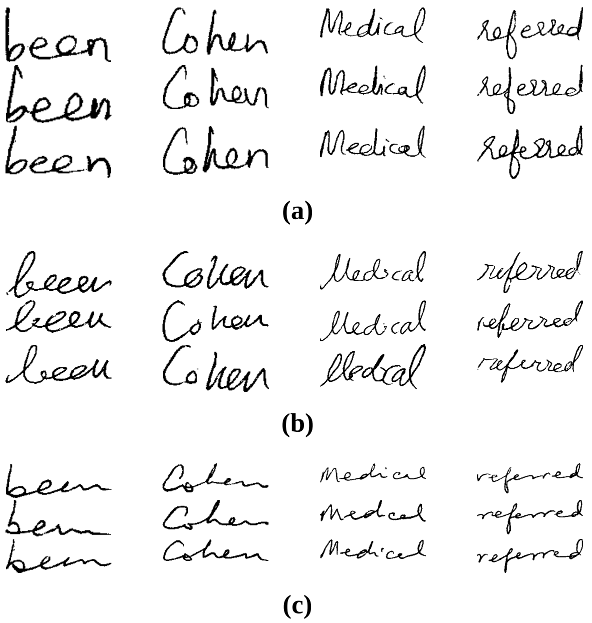


Figure 2. Samples of four characteristic words by three writers, each of whom wrote the same word three times.

3 Word Feature Extraction

Our word feature extraction algorithm originated from the so-called GSC algorithm for recognizing handwritten characters [2].

In [2], the features for a handwritten character consist of 512 bits corresponding to gradient (192 bits),structural (192 bits), and concavity (128 bits) features. This algorithm is frequently called GSC algorithm. Each of these three sets of features rely on dividing the scanned image of the character 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.

The GSC algorithm was modified to recognize handwritten digit pairs [10] by dividing a digit-pair image into a 4 x 6 region. Inspired by the success of digit pair recognition [10], we attempted to adjust the division of a word image to fit the length of the word content.

Suppose that an image I is to be divided into n x m subregions, the division works as following. At first, the image I is divided into n sub-regions along the vertical direction such that each sub-region contains the same number of black pixels; then I is divided into m sub-regions along the horizontal direction in the similar way. Therefore, I is divided into n x m sub-regions. Then, for each region, we use the aforementioned method to extract 12-bit gradient features, 12-bit structural features, and 8 concavity features. The n x m division results in a binary feature vector of n x m x 32 dimensions.

In this study, we tried five divisions for images of all four characteristic words, 4 x 4, 4 x 6, 4 x 8, 4 x 8, 4 x 10, and 4 x 12. From the experiments with these divisions, it is expected that a general rule for dividing a word image can be discovered. Figure 3 provides three examples of 4 x 8 division and the corresponding 1024-bit feature vectors.

4 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 the identification model, a number of binary feature vectors from word images are associated with each hand-

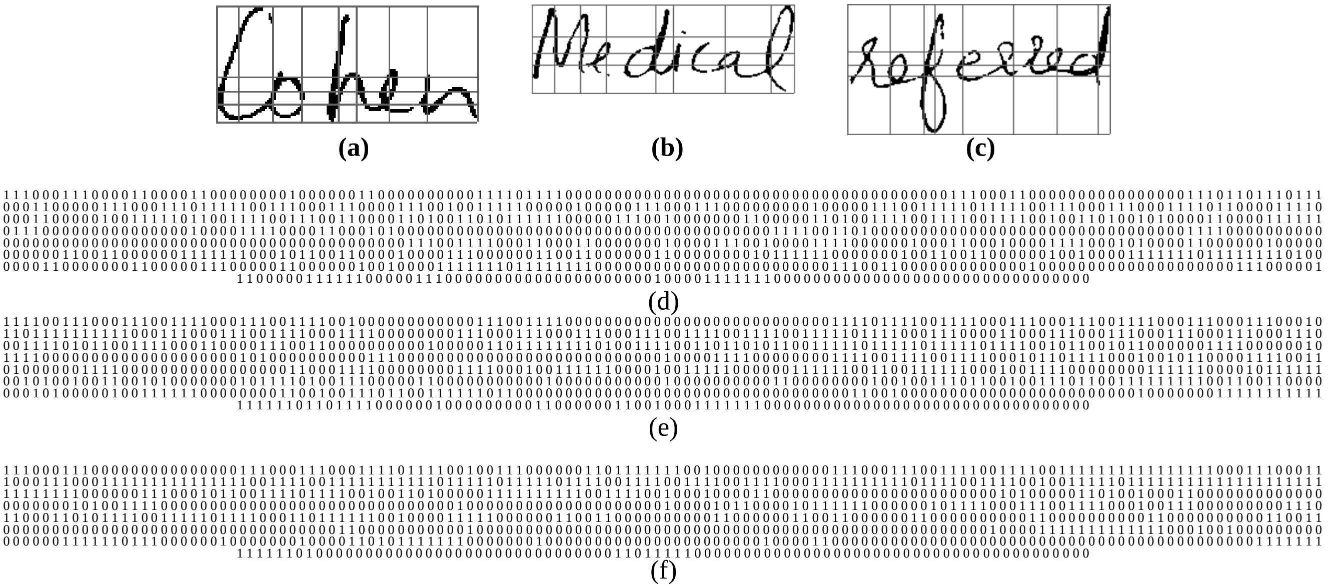


Figure 3. Three exemplar word images with 4 x 8 division: (a) a handwritten word "Cohen", (b) a handwritten word "Medical", (c) a handwritten word "referred", (d) the 1024-dimensional feature vector for the image (a), (e) the 1024-dimensional feature vector for the image (b), (f) the 1024-dimensional feature vector for the image (c).

written document, whereas, in the verification model, a real-valued distance vector (each component represents the distance between two words of the same content) is used to describe the difference between a pair of documents.

4.1 Similarity Measure for Binary Feature 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 defined in [9].

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 (2):

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

4.2 Similarity Measure Functions for Heterogeneous Features

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

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

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

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

4.3 Classification Techniques

Simple k-nearest neighbor classification is used for identification and verification. Specifically, for the identification model we use nearest neighbor classification based on the similarity function (2) and for the verification model we employ 6-nearest neighbor classification.

5 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 5000 elements randomly chosen from Θ.

From each document, a number (from 1 to 4) of images of characteristic words are segmented, then GSC features are extracted from each of them.

In the next section, we will examine the identification and verification performance of each characteristic word under five different divisions and combination of four characteristic words.

6 Handwriting Identification and Verification

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

6.1 Identification Results

Figure 4 shows the identification performance of four characteristic words and their simple combination under five different divisions. Obviously, 4 x 4 division presents the worst performance for all four characteristic words, and under 4 x 6 division the word "been" gets to its highest accuracy 45.4%. At meantime, "Cohen", "Medical" and "referred" all reach their own best performance under 4 x 8 division. While the experiment in [8] shows that the macro word-features alone can only identify less than 10% of 900 writers, in this study the combination of all four words under 4 x 6 division can identify correctly more than 80% of the writers.

Figure 5 compares four words under 4 x 8 division and sixty-two characters. As expected, all words bear higher discriminative power of handwriting individuality than characters. Moreover, the combination of all four words under 4 x 8 division can identify correctly 83% of the writers.

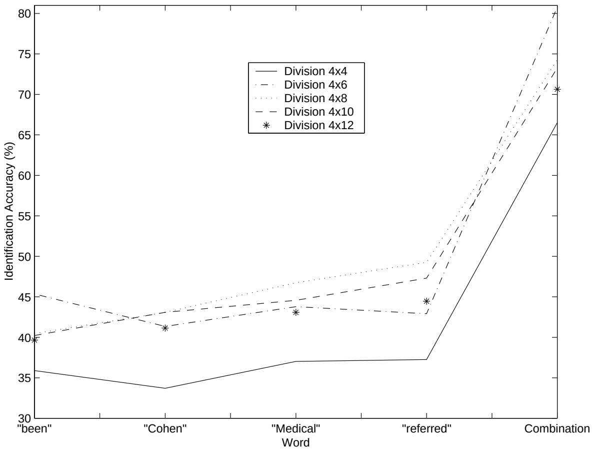


Figure 4. Handwriting identification performance of four characteristic words under five different divisions and their combination.

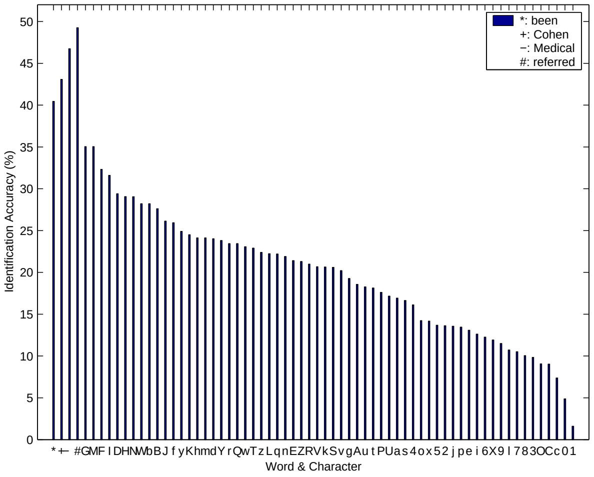


Figure 5. Handwriting identification performance of individual characters and 4 characteristic words under 4 x 8 division.

6.2 Verification Results

Figure 6 compares the verification performance of four words under 4 x 8 division and sixty-two characters. Each of the words still shows higher verification accuracy than any character. The simple combination of four words under 4 x 8 division gives 90.94% verification accuracy.

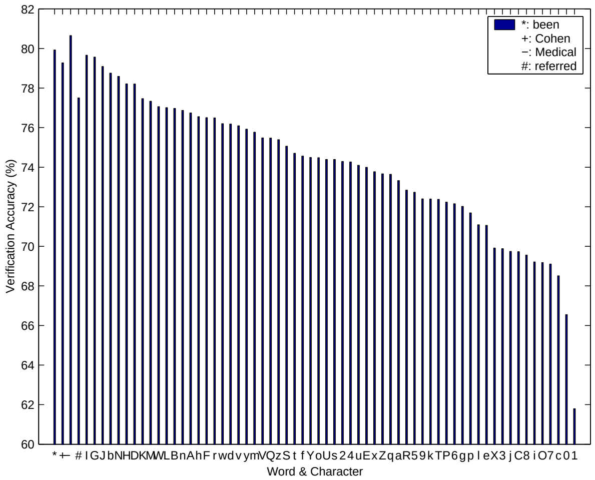


Figure 6. Handwriting verification performance of 62 characters and 4 characteristic words under 4 x 8 division.

From the observation and discussions above, some general remarks can be made as follows: (i) the features resulted from the modified GSC algorithm are effective in writer identification and verification, (ii) division of a word image in the GSC algorithm should be done dynamically, (iii) handwritten words usually bear more individuality than characters, and (iv) GSC word features are more effective than the macro word features defined in [8].

7 Conclusions

An algorithm for extracting features from handwritten words was developed for the purpose of writer identification and verification. The extensive experiments were conducted to evaluate the effectiveness of the features. In the future, more characteristic words will be added into the existing settings in order to establish an objective measurement of individuality of handwritten words.

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. Lorrette. 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] X. Wang, V. Govindaraju, and S. N. Srihari. Holistic digit pair recognition. *Journal of Pattern Recognition*, 33(12):1967–1974, December 2000.

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

[12] L. Zuo, Y. Wang, and T. Tan. Personal identification based on pca. In *http://nlpr-web.ia.ac.cn/english/irds/papers/zuolong/PR025.pdf*.