# Group Discriminatory Power of Handwritten Characters

Catalin I. Tomai, Devika M. Kshirsagar, Sargur N. Srihari

CEDAR, Computer Science and Engineering Department State University of New York at Buffalo, Buffalo, NY 14228, U.S.A.

## ABSTRACT

Using handwritten characters we address two questions (i) what is the group identification performance of different alphabets (upper and lower case) and (ii) what are the best characters for the verification task (same writer/different writer discrimination) knowing demographic information about the writer such as ethnicity, age or sex. The Bhattacharya distance is used to rank different characters by their group discriminatory power and the k-nn classifier to measure the individual performance of characters for group identification.

Keywords: Handwriting Identification, Character Discriminability

# 1. INTRODUCTION

The Handwriting Identification problem  $(^{12})$  is composed of two sub-problems: verification (given two documents, are they written by the same writer or not?) and identification (what is the writership identity of the query document?).

The identification of writer group attributes like gender, age and handedness from handwriting is an important goal in the forensic studies field. It is not yet clear whether particular handwriting features can be attributed to these group characteristics.<sup>1</sup> For example, the direction of strokes and loops may be affected by handedness. Also, aging may influence the quality of handwriting. In<sup>3</sup> Briggs concludes that, given the check-writing specimens from about 100 persons, males and females, the gender identification is not possible. While character (micro-features) have been found to be powerful for handwriting discrimination,<sup>2</sup> their capability in identifying the gender/handedness/age has not been evaluated yet.

In most criminal cases where handwriting is used as an evidence, few handwritten characters extracted from checks, tax-forms, etc. are at the disposal of the forensic document examiners. In these cases, the knowledge of the individual discriminatory power of characters can be used to estimate the validity of the writer verification/identification tasks or to improve their accuracy. Since in some cases partial information about the writer of the query document is available from other sources (his/her gender,approximate age, etc) we also evaluate how useful this information is to perform a faster and more accurate writer verification.

The objectives of this paper are: (i) estimate the performance of the current character features for gender/age/education/handedness identification (ii) find out for a given group (gender/handedness/age,etc.) what are the characters that present the best same-writer/different-writer discrimination and estimate their writer verification performance.

The rest of the paper is organized as follows. In Section 2 we describe the experimental framework used in this work. In Section 3 we measure the same-writer/different-writer discriminatory power of different characters for different categories of writers and estimate their accumulated writer verification performance. In Section 4 we estimate the performance of the character features for gender/age/handedness identification. We draw conclusions in Section 5.

Further Author information: (Send correspondence to Catalin I. Tomai) Catalin I. Tomai: E-mail: catalin@cedar.buffalo.edu Devika Kshirsagar: E-mail: dmk29@cedar.buffalo.edu Sargur N. Srihari: E-mail: srihari@cedar.buffalo.edu

#### 2. EXPERIMENTAL SETTINGS

**CEDAR database** The CEDAR letter database consists of more than 3,000 handwritten document images written by more than 1,000 writers representative of the US population. Each individual provided three handwriting samples of the same text (CEDAR letter<sup>2</sup>). The letter contains all the 26 alphabets in upper case and lower case and the 10 digits. For each writer, 3 samples of each digit and letter have been manually extracted. Figure 1 presents some of the characters extracted from the documents:

4 a La hk m3

Figure 1. Examples of characters extracted from the documents

The database is stratified along six categories: gender (G) male, female, handedness (H) right, left, age (A) under 15, 15-24, 25-44, 45-64, 65-84, over 85, ethnicity (E) white, black, hispanic, asian and pacific islander, American Indian, Eskimo, Aleut, highest level of education (D) below highschool graduate, above, and place of schooling (S) USA, Foreign.

The writers population is unequally distributed in these categories: for example, there are 443 males and 1119 females, 599 are high school graduates, 276 have a bachelor degree, 1406 are right-handed, 156 are left-handed, 337 are under the age of 24 and 347 are above the age of 45.

**Features and Classification** The character (micro) features that were used to experimentally determine the individuality of handwriting  $(^2)$  have been first used for character recognition.<sup>4</sup> For a given character, the micro-features set is of length 512 bits corresponding to gradient (192 bits), structural (192 bits), and concavity (128 bits) features. Each of these three sets of features derive from 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 topological and geometrical features: direction of bays, presence of holes, and large vertical and horizontal strokes.

The distance between two characters is given by a real valued distance vector computed using a similarity distance  $(^5)$  between the two binary vectors.

Since we need a classification method that would adapt to missing and variable number of features, we used K-nn (with K = 6) classification for both group identification and writer verification.

## **3. CHARACTERS PERFORMANCE FOR GROUP IDENTIFICATION**

The goal here is to estimate the group identification performance of different characters. For each group we have split the writers set into two subsets corresponding to the considered pairs of classes: (males/females), (left-handed/right/handed), etc. Each set contains the same number of document samples for each class, which varies with the representativeness of that class in the writers population. The training (exemplars) set contains double the number of documents in the test set. For example, for the gender group we have considered 300 writers in each class (male/female). Therefore, one sample of each character of the alphabet is extracted from 600 documents in the test set and 1200 exemplar documents which are then used by the K-nn classifier. For the other groups the number of writers considered is as follows: age: 313 writers for each class, handedness: 100 writers for each class, ethnicity: 106 writers in each class.

The individual performance results are obtained by considering each character sample from the test set and compare it with all the character samples in the training set (excluding its own). The test character is classified as part of the group class to which the majority of its neighbors (weighted by their proximity) belong. In the classification process the identity of the writers of the samples is considered unknown (that is, character samples are tagged only by their group class).

Table 1 displays the rankings of characters by their corresponding identification performance. Some characters present a high capability in identifying the classes of the age group, while the gender or handedness detection is still difficult.

**Table 1.** Characters ranked (top ones) according to their group identification performance over four binary demographic classification tasks

Male/Female		LH/RH		Under	24/Above 45	Foreign Born/US Born		
Char	$\operatorname{Perf}(\%)$	Char	$\operatorname{Perf}(\%)$	Char	$\operatorname{Perf}(\%)$	Char	$\operatorname{Perf}(\%)$	
b	70	Ν	61	h	82	а	65	
3	67	С	60	d	82	х	62	
m	66	R	60	х	82	Y	62	
у	66	n	59	b	82	b	61	
Y	65	9	58	v	81	q	61	
а	65	t	58	1	81	W	60	

Figure 2 displays the individual performance of different characters for each group considered. Some characters (e.g. b) appear in the top only for most of these groups.

Figure 3 presents the accumulated performance of the considered characters for group identification. As observed, the more characters are available the better the identification. The accumulated performance is significantly better than the individual performances. The differences in performance for the different groups becomes more evident. While more experiments are needed to draw a strong conclusion, we may observe that determining the age group is a significantly simpler task than determining the gender or ethnicity of a handwriting sample.

Because of the actual content of the documents considered and the occasional bad quality of the handwriting input, features may be extracted only from some characters of certain writers. Therefore, the distance computation between two sets of character features as well as the classification procedure has to be flexible in terms of the number of character features and samples being compared.

Let's consider two different writers  $w_1$  and  $w_2$  and two documents (with same or different content):  $D_1$  written by writer  $w_1$  and  $D_2$  written by writer  $w_2$ . Each document is described by a vector of character featuresone sample for each character,  $\bar{v}^1 = (c_1^1, c_2^1, ..., c_M^1)$  and  $\bar{v}^2 = (c_1^2, c_2^2, ..., c_M^2)$ , where M(=62) is the alphabet size. The distance  $D(D_1, D_2)$  between the two documents is given by :  $D(\bar{v}^1, \bar{v}^2) = \sum_i d(c_i^1, c_i^2)/N$ .  $d(c_i^1, c_i^2) = 0$ if one character feature is missing or both character feature vectors are missing. N is the number of character pairs  $(c_i^1, c_i^2)$  for which the distance between the feature vectors of the two characters  $d(c_i^1, c_i^2)$  exists.

## 4. DISCRIMINABILITY OF LETTERS KNOWING GROUP IDENTITY

#### 4.1. Discriminability ranking of Letters

The discriminatory power of letters can be directly derived from the writer verification and identification performances, method which, however, depends on the different classification schemes used. Here we use two different classifier-independent methods: the Bhattacharya distance between the same-writer/different-writer distance distributions and the receiver operating characteristic (ROC) curve.

Let's consider N letters  $c^i$  with i = 1, N from an alphabet A of M letters, with  $N \leq M$ . We also consider W writers and from each writer' set of handwriting samples we extract K instances of a certain letter  $c^i$ . The  $k^{th}$  letter image sample from the set of images extracted for writer w is  $c^i_{w,k}$ .

Once we convert each letter image into a feature vector we compute sets of similarity distances of two types:  $D_{w,w}^i = d(c_{w,l}^i, c_{w,m}^i)$ , for  $l \neq m$  - Distances between letter images belonging to the same writer

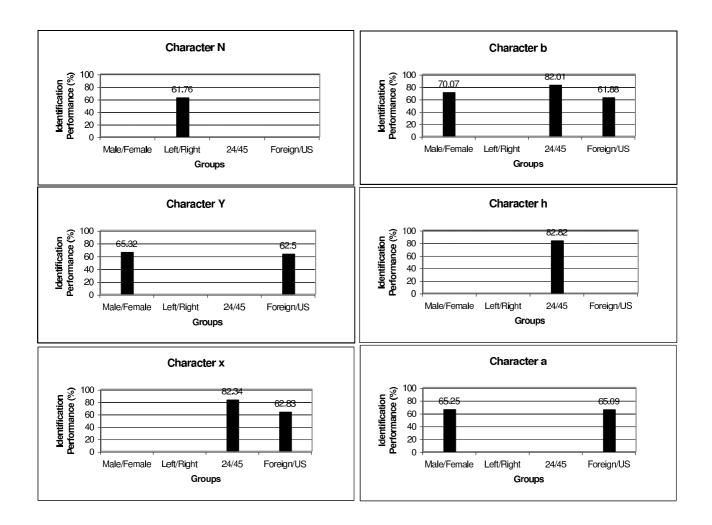


Figure 2. Individual performance of selected characters for group identification

 $D_{w,x}^i = d(c_{w,l}^i, c_{x,m}^i)$ , for  $w \neq x$  - Distances between letter images belonging to different writers

For example, for the (male/female) pair of classes we have randomly picked 144 male writers for the training set. The within-writer distances have been computed by using 3 documents of each of the 144 writers, thus having 432 pairs of documents written by (male-male) writers. The between-writer distances have been derived from 432 pairs of documents written by (male-male) and (male-female) pairs.

We approximate the histograms of  $D_{w,w}^i$  and  $D_{w,x}^i$  distances by normal distributions and the confusion is measured by the amount of overlap between the two distributions. If we have two arbitrary distributions  $p_1$  and  $p_2$ , at any position x, the probability of classification error is given by:  $P(error) = min(p_1(x), p_2(x))/(p_1(x) + p_2(x))$ . An upper bound on P(error) is given by the Bhattacharya distance which measures the separation between two pdf's:

$$d_B = \int_{-\inf}^{\inf} p_1(x)^{1/2} p_2(x)^{1/2} dx \tag{1}$$

Assuming that the distributions are Gaussian the formula becomes:

$$d_B = \frac{1}{2} \sum_{i=1}^{N} \frac{\sigma_{i,1}^2 + \sigma_{i,2}^2}{2\sqrt{\sigma_{i,1}^2 \sigma_{i,2}^2}} + \frac{1}{2} \sum_{i=1}^{N} \frac{\mu_{i,1} - \mu_{i,2}^2}{\sigma_{i,1}^2 + \sigma_{i,2}^2}$$
(2)

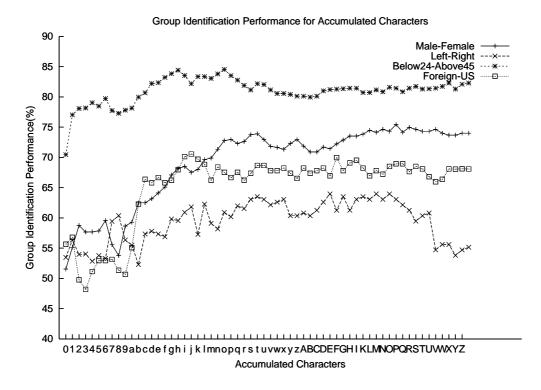


Figure 3. Accumulated performance of characters for group identification

The larger the overlap between the distributions for a certain letter, the higher the uncertainty regarding the writership of two images of that letter. Therefore, letters that are more "discriminatory" have a lower Bhattacharya distance.

Table 2 presents the ranking of the top 6 letters by their Bhattacharya distance between the same-writer and different-writer distance distributions.

Males Fema		emales	Bachelor		High School		LH		RH		Under 24		Above 45		
С	Dist	С	Dist	С	Dist	С	Dist	С	Dist	С	Dist	С	Dist	С	Dist
G	0.5853	G	0.5555	D	0.5564	Ν	0.5998	J	0.5597	G	0.4874	G	0.6055	Ν	0.5391
Ι	0.6045	W	0.5740	Ι	0.5612	G	0.6024	D	0.5632	h	0.5059	d	0.6195	D	0.5688
h	0.6061	Ι	0.5789	Ν	0.6067	Ι	0.6159	h	0.5661	Ι	0.5350	J	0.6265	G	0.5697
b	0.6079	D	0.5860	b	0.6067	F	0.6325	S	0.5790	Α	0.5355	Ι	0.6309	Μ	0.5723
J	0.6102	w	0.5963	G	0.6106	W	0.6393	b	0.6026	b	0.5598	W	0.6363	W	0.5745
у	0.6206	Ν	0.5966	W	0.6145	D	0.6407	f	0.6062	Η	0.5614	b	0.6398	F	0.5783

Table 2. Top 6 letters ranked by their same-writer/different writer discriminatory capability for different groups

While there are no significant differences among the best letter identities between the males/female and bachelor/high-school pairs, there is significantly less discriminability for the letters written by left-handed writers than right-handed writers. Also, there is an evident difference in discriminability performance between the under-24 writers and above-45 writers. This observation can be accounted to handwriting becoming more individualistic with age. With the exception of the "under-24" group of writers, for all the other categories, upper-case letters present more discriminatory power than the lower-case ones. This may be due to their positioning at the beginning of a word.

The ROC curves are computed by measuring the true positive and false negative rates for different chosen values of the distance in the SW/DW distances interval range. Figure 4 presents the ROC curves for the top letters of some of the groups considered. The ROC curves mostly validate the findings obtained using the Bhattacharya distance, the top characters (e.g. G,I) for males presenting the best performance in both cases.

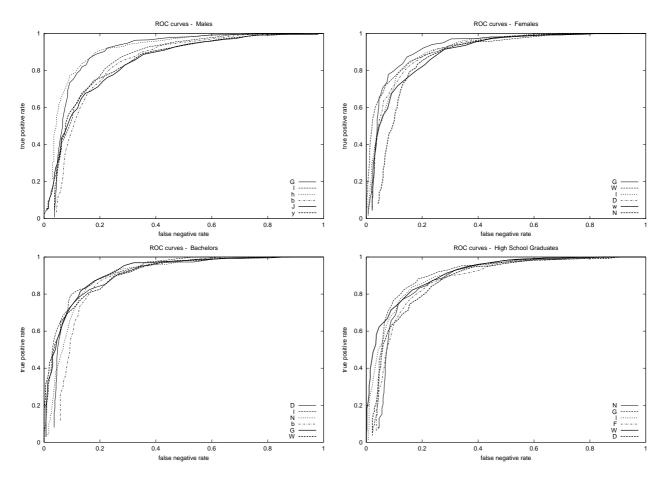


Figure 4. ROC curves of writer verification performance for the top ranked letters

#### 4.2. Writer Verification Performance Evaluation

Next, we use the letter rankings obtained in the previous section and measure the accumulated writer verification performance of these letters. The testing set consists of 1500 within-writer distances obtained from a set of 1500 documents written by 500 randomly chosen writers (3 documents per writer). Another 1500 between-writer distances were obtained from 1500 pairs of documents written by different writers.

Since typically in HWI analysis, given a pair of documents, one of them belongs to the "query" writer and the other one belongs to the "exemplars" set, we have looked at the group information of the first writer of the pair. For example, for the gender group we check the gender of the first writer. If the writer is male, we consider 1,2,3,...52 letters, starting with the "best" one as given by the ranking obtained for the male group, to do the writer verification analysis.

Figure 5 displays the accumulated writer verification performance of these letter features for all groups considered. The "Base Case" plot is the performance obtained when the letters are considered in sequential (alphabetical) order. The plots show that knowing the group information always help us choosing better characters to do the writer verification analysis. However, some group information is more important than other. For example, knowing the gender or age is more important that knowing the handedness information.

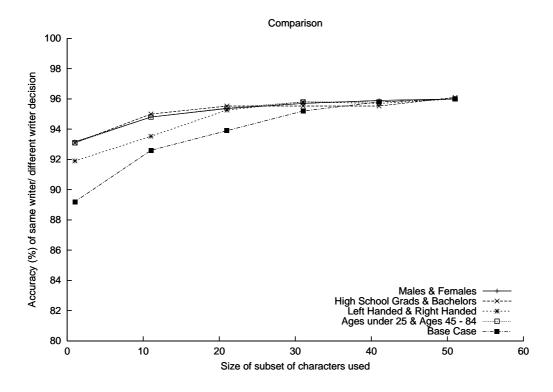


Figure 5. Writer Verification performance of accumulated characters for all groups considered

## 5. CONCLUSION

We have estimated the group identification performance of different characters and measured their accumulated performance. The results are encouraging for some groups, however, more features are needed to have a greater confidence in the decision. Our work shows that, for different categories of writers, characters present different writer verification performance. Knowing the demographic information about writers helps in identifying the best characters and achieve better writer verification performance using fewer characters. We plan to further improve on these results by using new document and word features to help the decision.

## 6. ACKNOWLEDGMENTS

Work supported by the Department of Justice, National Institute of Justice, grant 2002-LT-BX-K007

## REFERENCES

- R. A. Huber and A. M. Headrick, Handwriting Identification: Facts and Fundamentals, CRC Press LLC, 1999.
- S. N. Srihari, S.-H. Cha, H. Arora, and S. Lee, "Individuality of handwriting," *Journal of Forensic Sciences* 47, July 2002.
- M. E. Briggs, "Empirical study iv, writer identification: Determination of gender from check writing style," Journal of Questioned Document Examination 10(1), pp. 3–25, 2002.
- J. T. Favata and G. Srikantan, "A multiple feature/resolution approach to handprinted character/digit recognition," International Journal of Imaging Systems and Technology 7, pp. 304–311, 1996.
- 5. B. Zhang and S. N. Srihari, "Binary vector dissimilarities for handwriting identification," in *SPIE Document Recognition and Retrieval X*, (Santa Clara, California, USA), January 2003.