Writer Identification: Statistical Analysis and Dichotomizer

Sung-Hyuk Cha and Sargur N. Srihari

Center of Excellence for Document Analysis and Recognition
State University of New York at Buffalo, NY, 14260
{scha,srihari}@cedar.buffalo.edu

Abstract. This paper is to determine the statistical validity of individuality in handwriting based on measurement of features, quantification and statistical analysis. In classification problems such as *writer, face, finger print* or *speaker identification*, the number of classes is very large or unspecified. To establish the inherent distinctness of the classes, i.e., validate individuality, we transform the many class problem into a dichotomy by using a "distance" between two samples of the same class and those of two different classes. A measure of confidence is associated with individuality. Using ten feature distance values, we trained an *artificial neural network* and obtained 97% overall correctness. In this experiment, 1,000 people provided three sample handwritings.

Key Words: Dichotomizer, Hypothesis Testing, Individuality, Writer Identification

1 Introduction

The *Writer Identification problem* is a process to compare questioned handwriting with samples of handwriting obtained from known sources for the purposes of determining authorship or non-authorship. In other words, it is the examination of the design, shape and structure of handwriting to determine authorship of given handwriting samples. Document examiners or handwriting analysis practitioners find important features to characterize individual handwriting as features are consistent with writers in normal undisguised handwriting [1]. Authorship may be determined due to the following hypothesis that people's handwritings are as distinctly different from one another as their individual natures, as their own finger prints. It is believed that no two people write the exact same thing the exact same way.

Since the writer identification plays an important investigative and forensic role in many types of crime, various automatic writer identification by computer techniques, feature extraction, comparison and performance evaluation methods have been studied (see [8] for the extensive survey). Osborn suggested a statistical basis to handwriting examination by the application of the *Newcomb rule of probability* and Bertillon was the first who used the Bayesian theorem to handwriting examination [5]. Hilton calculated the *odds* by taking the likelihood

F.J. Ferri et al. (Eds.): SSPR&SPR 2000, LNCS 1876, pp. 123–132, 2000.

ratio statistic that is the ratio of the probability calculated on the basis of the similarities, under the assumption of identity, to the probability calculated on the basis of dissimilarities, under the assumption of non-identity [5,4]. However, relatively little study has been carried out to demonstrate its scientific and statistical validity and reliability as forensic evidence. To identify writers, it is necessary to determine the statistical validity of individuality in handwriting based on measurement of features, quantification, and statistical analysis.

Consider the multiple class problem where the number of classes is small and one can observe enough instances of each class. To show the individuality of class statistically, one can cluster samples into classes and infer it to the population. It is an easy and valid setup to establish the individuality. Now consider the *many class problem* where the number of classes is too large to be observed (n is very large). Most pattern identification problems such as *writer, face, fingerprint* or *speaker identification* fall under the aegis of the many class problem. Most parametric or non-parametric multiple classification techniques are of no use and the problem is seemingly insurmountable because the number of classes is too large or unspecified.

To establish the inherent distinctness of the classes, i.e., validate individuality, we transform the many class problem into a dichotomy by using a "distance" between two samples of the same class and those of two different classes. We tackle the problem by defining a distance metric between two writings and finding all writings which are within the threshold for every feature. In this model, one need not observe all classes, yet it allows the classification of patterns. It is a method for measuring the reliability of classification about the entire classes based on information obtained from a small sample of classes drawn from the class population. In this model, two patterns are categorized into one of only two classes; they are either from the same class or from the two different classes. Given two handwriting samples, the distance between two documents is first computed. This distance value is used as data to be classified as positive (authorship, inner-variation, within author or identity) or negative (non-authorship, intra-variation, between different authors or non-identity). We use within author distance and between authors distance throughout the rest of this paper. Also, we use subscriptions of the positive ($\oplus$) and negative ($\ominus$) symbols as the nomenclature for all variables of within author distance and between authors distance, respectively.

The subsequent sections are organized as follows. The section 2 discusses the dichotomy transformation. The section 3 shows the experimental database of writer, exemplar and features. In section 4, the full statistical analysis of the collected database and gives the experimental results. Finally, the section 5 concludes the paper.

2 Transformation of Polychotomizer to Dichotomizer

The writer identification can be viewed as a U.S. population category classification problem, so called *Polychotomizer*. As the number of classes is enormously

large and almost infinite, this problem is seemingly insurmountable. In this section, we show how to transform a large polychotomizer to a simple *dichotomizer*, a classifier that places a pattern in one of only two categories.

To illustrate, suppose there are three writers, $\{W_1, W_2, W_3\}$. Each writer provides three documents and two scalar value features extracted per document. Fig. 1 (a) shows the plot of documents for every writer. To transform into di-

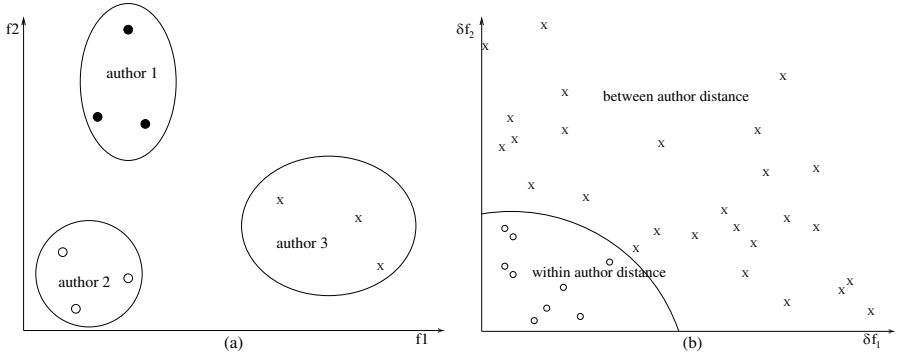

Fig. 1. *Transformation from (a) Feature domain to (b) Feature distance domain*

stance space, we take the vector of distances of every feature between writings by the same writer and categorize it as a *within author distance* denoted by $x_{\oplus}$. The sample of *between author distance* is, on the other hand, obtained by measuring the distance between two different person's handwritings and is denoted by $x_{\ominus}$. Let d_{ij} denote i'th writer's j'th document.

$$\boldsymbol{x}_{\oplus} = \boldsymbol{\delta}(d_{ij} - d_{ik}) \text{ where } i = 1 \text{ to } n, j, k = 1 \text{ to } m \text{ and } j \neq k \quad (1)$$

$$\boldsymbol{x}_{\ominus} = \boldsymbol{\delta}(d_{ij} - d_{kl}) \text{ where } i, k = 1 \text{ to } n, i \neq k \text{ and } j, l = 1 \text{ to } m \quad (2)$$

where n is the number of writers, m is the number of documents per person, δ is the distance between two documents. Fig. 1 (b) represents the transformed plot. The feature space domain is transformed to the feature distance space domain. There are only two categories: *within author distance* and *between author distance*.

It would be desirable if all distances between the same class (writer) in feature domain belong to the within class distance class in feature distance domain. Similarly, we would like all distances between two different classes in feature domain belong to the between class distance class in feature distance domain. Unfortunately, this is not always the case; perfectly clustered class in feature domain may not be perfectly dichotomized in feature distance domain. Thus, we have a trade-off between tractability and accuracy. Since sampling a sufficiently large sample from each individual person is intractable, we may wish to transform feature domain to the feature distance domain where we can get large samples for

both classes. By the transformation, the problem becomes a tractable inferential statistic problem but we might get the lesser accuracy.

3 Experimental Database

There are three steps to validate the individuality of handwriting: i) data collection, ii) feature extraction and iii) statistical analysis. In this section, we discuss the first two issues and the following section 4 covers the statistical analysis. The first one is data collection of writers, exemplars, and features. We collected seven attributes of writers through the questionnaire data-sheet. They are *gender, age, handedness, highest level of education, country or states of primary education, ethnicity* and *country of birth*. We built a database that is "representative" of the US Population. This has been achieved by basing our sample distribution on the US census data (1996 Projections) [7]. There are 510 female and 490 male population distributions and 36% of white ethnicity group, etc. The database contains handwriting samples of 1000 distinct writers.

3.1 Exemplar: CEDAR Letter

Each subject provides three exemplars of the CEDAR Letter shown in Figure 2. The *CEDAR Letter* is concise (it has just 156 words), easy to understand and

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

Fig. 2. CEDAR Letter

complete. It's *complete* in that, each alphabet occurs in the beginning of a word as a capital and a small letter, and as a small letter in the middle and end of a word. In addition, it also contains punctuation, numerals, interesting letter and numeral combinations (ff, tt, oo, 00) and a general document structure that would allow us to extract document level features such as word and line spacing, line skew etc. Forensic literature refers to many such documents - the *"London Letter"*, the *"Dear Sam Letter"* to name a few. But none of them are *complete* in the sense of the CEDAR Letter as follows. All capitals must appear in the letter and it is desirable to have all small letters in the beginning, middle and terminal positions of the word. We score the letter according to these constraints:

$$\text{score(letter x)} = \frac{104 - \text{Number of 0's}}{104} \tag{3}$$

The CEDAR letter scores 99% whereas the London letter scores 76%. the cedar letter has only 1 zero entry that is a word that ends with a letter *"j"*. Since there is no common English word that ends with the letter *"j"*, the cedar letter excludes this entry.

3.2 Feature Extraction

Encouraged by the recent success in off-line handwriting recognition and handwritten address interpretation [9], we utilize the similar features for the individuality validation. Albeit there are numerous features in line, word, character and spacing features, we give some document level computational features.

The darkness value is the threshold value that separates the character parts and background parts of the document image. A digital image is a rectangular array of picture elements called *pixels* and each pixel has a darkness value between 0 and 255. A histogram is built and it has two peaks. One is due to dark handwritten characters and the other is due to the bright background. The valley between two peaks is the grey level threshold. We use the darkness value, grey level threshold value as an indicator of pen pressure. Another document level feature is the number of blobs that is the number of connected components in the document image. A blob is also known as an exterior contour. This feature is related to intra-word and inter-word connections. Those writers who connects characters or words have fewer number of blobs while those who do not connect have lots of blobs. A similar feature is the number of holes that is the number of closed loops. A hole is often called an interior contour or a lake. This feature gives the tendency of making loops while writing. The average stroke width feature is computed by measuring the highest frequency of width per line. We compute the slant, skew and average height of character features.

4 Analysis

4.1 Size of Sample

Let $n_{\oplus} = |x_{\oplus}|$ and $n_{\ominus} = |x_{\ominus}|$.

Fact 1 *If n people provide m writings, there are $n_{\oplus} = \binom{m}{2} \times n$ positive data, $n_{\ominus} = m \times m \times \frac{n(n-1)}{2}$ negative data and $\binom{mn}{2}$ data in total.*

Proof. $n_{\oplus} = \binom{m}{2} \times n$ is straight-forward. To count the negative data, we can enumerate them as $m \times (m \times (n-1)) + m \times (m \times (n-2)) + \cdots + m \times (m \times 1)$. For the first author, there are $m \times (n-1)$ number of other writer's writing data and he has three writing data. For the second author, there are $m \times (n-2)$ number of other writer's writing data that are not counted yet. Therefore, $n_{\ominus} = m \times m \times \sum_{i=1}^{n-1} i$. Now, $n_{\oplus} + n_{\ominus}$ must be $\binom{mn}{2}$.

$$\binom{mn}{2} = \frac{(mn)!}{(mn-2)!2} = \frac{(mn)(mn-1)}{2} = \frac{m(m-1)}{2} n + m^2 \frac{n(n-1)}{2} = n_{\oplus} + n_{\ominus}$$

□

In our data collection, 1000 people (statistically representative U.S. population) provide exactly three samples. Hence, there are $n_{\oplus} = 3000, n_{\ominus} = 4,495,500$ and $4,498,500$ data in total.

Most statistical testing requires the assumption that observed data be statistically independent. The distance data is not statistically independent: one obvious reason being the triangle inequality of three distance data of the same person. This caveat should not be ignored. One immediate solution is to choose randomly a smaller sample from a large sample obviating the triangle inequality. One can partition $n_{\oplus} = 3000$ data into disjoint subsets of 500 guaranteeing no triangle inequality.

4.2 Feature Evaluation

A good descriptive way to represent the relationship between two populations is calculating overlaps between two distributions. Fig. 3 illustrates the two distributions assuming that they are normal. Although this assumption is invalid, we use it to describe the behavior of two population figuratively. The *type I error*, α occurs when the same author's documents are identified as different authors and the *type II error*, β occurs when the two document written by two different writers are identified as the same writer as shown in Figure 3.

$$\alpha = Pr(\text{dichotomizer}(d_{ij}, d_{kl}) \geq T | i = k) \tag{4}$$

$$\beta = Pr(\text{dichotomizer}(d_{ij}, d_{kl}) < T | i \neq k) \tag{5}$$

Let $\hat{X}$ denote the distance x position where two distributions intersect. As shown in Fig. 3, *type 1 error* is the right side area of positive distributions where the decision bound $T = \hat{X}$. Suppose one must make a crisp decision and choose the intersection as a classification bound. Then the type one error means that the probability of error that one classifies two writings as different authors even though they are written by a same person. The *type 2 error* is the left side area

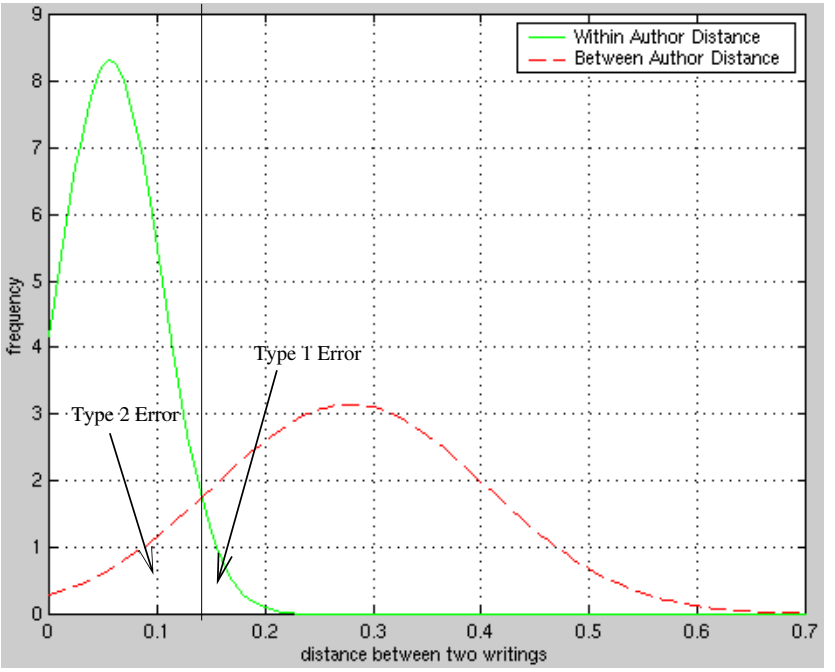

Fig. 3. *Type I and II errors*

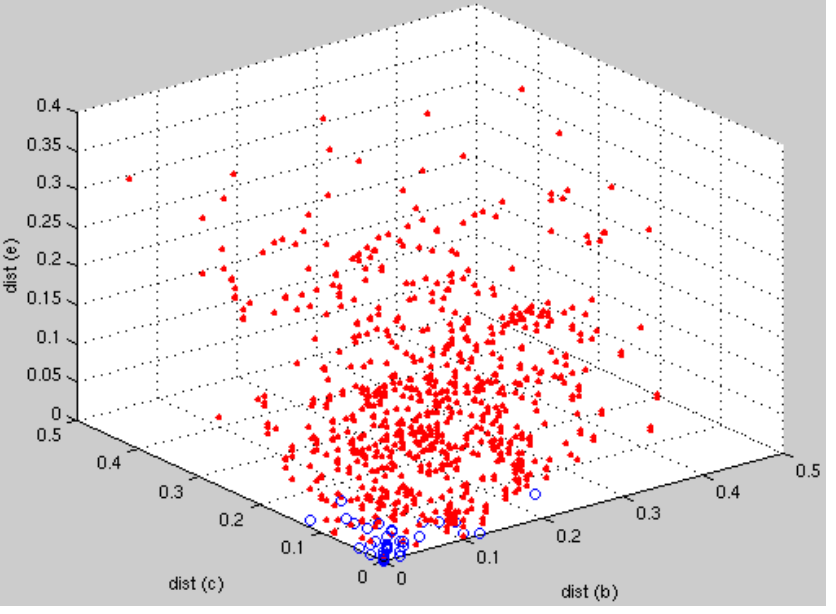

Fig. 4. *Three Feature Space Distributions*

Table 1. Evaluating Features by overlaps

	δ_A	δ_B	δ_C	δ_D	δ_E	δ_F	δ_{ABCE}
$\hat{X}$	0.0172	0.1029	0.0825	0.0317	0.0576	1.7300	0.1407
Type 1 error	9.0%	6.94%	5.0%	24.54%	0.81%	3.0%	3.84%
Type 2 error	38.6%	27.3%	26.0%	51.4%	15.7%	27.0%	14.0%
Rem.	Good	Good	Good	Bad	Best	Good	

of negative distributions meaning the probability of error that one classifies two writings as a same author even though they are written by two different writers. Table 1 shows the intersection positions, $\hat{X}$'s and proportion of each error for each feature. Note that feature (E) is an excellent feature whereas feature (D), the average stroke width is a bad one. Note that the last column in Table 1 is not the multivariate results but the univariate overlaps of the Euclidean distance of multiple features.

Another novel way to handle multiple features is to get the distance value for each feature and produce the multi-dimensional vector distances. Fig. 4 illustrates three dimensional distance values, $\{\delta_b, \delta_c, \delta_e\}$. Similar to the one-dimensional case, the within author distances tend to cluster toward the origin while the between authors distances tend to be apart from the origin. Various multivariate analysis [6] such as *Hotelling* T^2 *statistics* to test hypotheses on two multivariate means but we use the *artificial neural network*.

4.3 Dichotomizer: Artificial Neural Network

Samples of both class are divided into 6 groups of 500 in size. One pair set is used as a training set and the other set is used as a validation set. The rest of them are used as testing sets. Using ten feature distance values, we trained an *artificial neural network* as shown in Fig. 5. It is observed that the higher

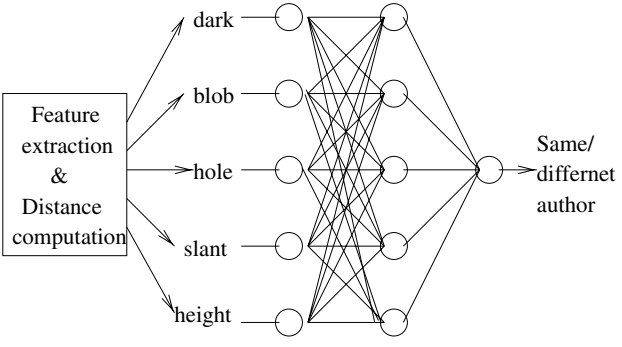

Fig. 5. ANN Dichotomizer Design

number of features the better the dichotomizer is as shown in Table 2.

Table 2. Experimental results vs. the number of features

no. of features	5	9	10
Type I error	5.3%	4.6%	3.5%
Type II error	5.2%	3.5%	2.1%
Accuracy	95%	96%	97%

5 Conclusion

In this paper, we showed that the multiple category classification problem can be viewed as a two-categories problem by defining the distance and taking those values as positive and negative data. This paradigm shift from the polychotomizer to the dichotomizer makes the writer identification that is a hard U.S. population multiple class problem very simple.

We designed an experiment to show the individuality of handwriting by collecting samples from people that is representative of the US population. Given two randomly selected handwritten documents, we can determine whether the two documents were written by the same person or not. Our performance is 97%.

One advantage of the dichotomy model working on distribution of distances is that many standard geometrical and statistical techniques can be used as the distance data is nothing but scalar values in feature distance domain whereas the feature data type varies in feature domain. Thus, it helps to overcome the non-homogeneity of features. Techniques in *pattern recognition* typically require that features be homogeneous. While it is hard to design a polychotomizer due to non-homogeneity of features, the dichotomizer simplifies the design by mapping the features to homogeneous scalar values in the distance domain. Types of features can be nominal, linear, angular, strings, histograms [2], etc. Full discussion on the multiple feature integration for writer identification and various distance measures can be found in [3].

5.1 Work on Progress

Features used in the analysis are document level features. As the segmentation tools are developed, features in the line, word, and character level features will be applied. The higher performance is expected.

We are currently dealing with the following five issues: **i) comparison between polychotomy and dichotomy:** comparing polychotomy in feature domain and dichotomy in distance domain from the view point of tractability vs. accuracy, **ii) distance measures:** use and evaluate several distance measures, e.g., element, histogram, probabilistic density function, string, and convex hull distances, **iii) efficient search:** nearest-neighbor algorithms for distance measures **iv) applications:** designing and analyzing an algorithm for *writer identification* for a known number of writers and a method for *handwritten document image indexing and retrieval*, and **v) discovery:** mining a database consisting

of writer data and features obtained from a handwriting sample, statistically representative of the US population, for feature evaluation and to determine similarity of a specific group of people.

Acknowledgments

This research has been possible funded by National Institute of Justice (NIJ) in response to the solicitation entitled *Forensic Document Examination Validation Studies*: Award Number 1999-IJ-CX-K010 [10].

References

1. Russell R. Bradford and Ralph B. Bradford. *Introduction to Handwriting Examination and Identification.* Nelson-Hall Publishers: Chicago, 1992.
2. Sung-Hyuk Cha and Sargur N. Srihari. Distance between histograms of angular measurements and its application to handwritten character similarity. In *Proceedings of 15th ICPR*, pages –. IEEE CS Press, 2000.
3. Sung-Hyuk Cha and Sargur N. Srihari. Multiple feature integration for writer identification. In *Proceedings of 7th IWFHR 2000*, pages –. Springer-Verlag, September 2000.
4. Ordway Hilton. The relationship of mathematical probability to the handwriting identification problem. In *Proceedings of Seminar No. 5*, pages 121–130, 1958.
5. Roy A. Huber and A. M. Headrick. *Handwriting Identification: Facts and Fundamentals.* CRC Press LLC, 1999.
6. Donald F. Morrison. *Multivariate statistical methods.* New York : McGraw-Hill, 1990.
7. U.S. Department of Commerce. Population profile of the united states. Current Population Reports Special Studies P23-194, Semtember 1998.
8. Rejean Plamondon and Guy Lorette. Automatic signature verification and writer identification - the state of the art. *Pattern Recognition*, 22(2):107–131, 1989.
9. Sargur N. Srihari and E.J. Keubert. Integration of hand-written address interpretation technology into the united states postal service remote computer reader system. In *Proceedings of 4th International Conference on Document Analysis and Recognition (ICDAR'97)*, pages –, Ulm, Germany, 1997.
10. Jeremy Travis. Forensic document examination validation studies. Solicitation: http://ncjrs.org/pdffiles/sl297.pdf, October 1998.