J Forensic Sci, March 2008, Vol. 53, No. 2
doi: 10.1111/j.1556-4029.2008.00682.x
Available online at: www.blackwell-synergy.com

Sargur Srihari,[1] *Ph.D.; Chen Huang,*[2] *M.S.; and Harish Srinivasan,*[2] *M.S.*

On the Discriminability of the Handwriting of Twins

ABSTRACT: As handwriting is influenced by physiology, training, and other behavioral factors, a study of the handwriting of twins can shed light on the individuality of handwriting. This paper describes the methodology and results of such a study where handwriting samples of twins were compared by an automatic handwriting verification system. The results complement that of a previous study where a diverse population was used. The present study involves samples of 206 pairs of twins, where each sample consisted of a page of handwriting. The verification task was to determine whether two half-page documents (where the original samples were divided into upper and lower halves) were written by the same individual. For twins there were 1236 verification cases—including 824 tests where the textual content of writing was different, and 412 tests where it was the same. An additional set of 1648 test cases were obtained from handwriting samples of nontwins (general population). To make the handwriting comparison, the system computed macro features (overall pictorial attributes), micro features (characteristics of individual letters), and style features (characteristics of whole-word shapes and letter pairs). Four testing scenarios were evaluated: twins and nontwins writing the same text and writing different texts. Results of the verification tests show that the handwriting of twins is less discriminable than that of nontwins: an overall error rate of 12.91% for twins and 3.7% for nontwins. Error rates with identical twins were higher than with fraternal twins. Error rates in all cases can be arbitrarily reduced by rejecting (not making a decision on) borderline cases. A level of confidence in the results obtained is given by the fact that system error rates are comparable to that of humans (lower than that of lay persons and higher than that of questioned document examiners [QDEs]).

KEYWORDS: forensic science, questioned document examination, handwriting processing, document analysis, writer verification, twins study

The distinctiveness of each person's handwriting has long been intuitively observed. Methods have been developed for a human expertise of handwriting matching over many decades (1–5). Yet there is a need for studies in the quantitative assessment of the discriminative power of handwriting particularly for the acceptance by the courts of evidence provided by questioned document examiners (QDE). In a previous study of handwriting individuality (6), we reported on the discriminability of handwriting of a diverse population from across the United States. The present paper reports on a complementary study of the discriminatory power of handwriting when the population consists of a cohort group consisting of twins. Both the previous study and the current study are based on automated methods for handwriting comparison. The current study uses algorithms that are updated with respect to the types of handwriting features that are computed.

The necessity of studying cohort groups such as twins has been considered to be important in various medical (7), social (8), biometric, and forensic fields. The similarities of genetic and environmental influences of twins allow the importance of the characteristic to be studied in its limiting conditions where extraneous factors are minimized. Any methodology needs to be tested for boundary conditions where the possibility of error is maximum. Satisfactory performance with twins strengthens the reliability of the method. Research has been done on twins for biometrics such as fingerprints (9) and DNA (10), which are physiological in nature, i.e., they do not change after birth. On the other hand, handwriting is more of a behavioral characteristic with a significant psychological component associated with it, which makes the study of the handwriting of twins to be meaningful (11).

Computational methods for handwriting analysis have been more recently developed (6,12–14). When designed as a system they allow conducting large scale and statistically meaningful tests. They provide accuracy rates for verification (which is the task of determining whether two handwriting samples were written by the same person) and for identification (which is the task of determining the writer of a questioned document from a set of known writers). They provide a base-line result for human and automated questioned handwriting examination. Subjecting automatic methods to the handwriting of twins will throw some light on the effect of genetic and environmental factors.

Specific goals of the present study are to extend a previous study (6) on automated handwriting analysis by (1) comparing performance on handwriting of twins with those of the general population, (2) determining performance when the textual content of the questioned and known writing is different, (3) comparing performance on fraternal and identical twins, and (4) comparing system performance with that of humans. The evaluation was performed using an automatic method of writer verification, which provides a quantitative measure of the degree of match between a pair of handwriting samples, known and questioned, based on the shapes of characters, bi-grams and words, and the global structure of the composition, e.g., slant, word spacing, etc.

The rest of the paper is organized as follows. We first describe the verification system including the features, similarity computation, and decision algorithm. Then, we present the twins test-bed, i.e., the way in which the test cases were obtained and grouped. After that we give the results of the experiments performed, followed by comparison between human performance and system performance on the same testing scenarios, and comparison of the current results with

[1]Department of Computer Science and Engineering and Director, Center of Excellence for Document Analysis and Recognition, University at Buffalo, State University of New York, Buffalo, NY 14228.
[2]Department of Computer Science and Engineering, University at Buffalo, State University of New York, Buffalo, NY 14228.

Received 13 July 2006; and in revised form 6 Oct. 2007; accepted 13 Oct. 2007.

those that were reported previously with samples not specialized to twins. The last section contains concluding remarks.

Automatic Writer Verification Method

To begin, it is useful to define the terms verification, identification, and recognition. Verification is the task of determining whether a pair of handwriting samples was written by the same individual. Identification is to find the writer having the closest match with the questioned document out of a pool of writers. Recognition is to convert images to text. The CEDAR-FOX system was used to perform these functions (15). The system has interfaces to scan handwritten document images, obtain line and word segments, and automatically extract features for handwriting matching after performing character recognition and/or word recognition (with or without interactive human assistance in the recognition process, e.g., by providing word truth).

Statistical parameters for writer verification are built into CEDAR-FOX, which were obtained using several pairs of documents, which were either written by the same writer or by different writers. Writer verification consists of four steps: (1) writing element extraction, (2) similarity computation, (3) estimating conditional probability density estimates for the difference being from the same writer or from different writers (as Gaussian or Gamma).

Given a new pair of documents, verification is performed as follows: (1) writing element extraction, (2) similarity computation, (3) determining the log-likelihood ratio (LLR) from the estimated conditional probability density estimates.

Features and Similarity

The system computes three types of features—macro features at the document level, micro features at the character level, and style features from bi-grams and words. Each of these features contributes to the final result to provide a confidence measure of whether two documents under consideration are from the same or different writers.

Macro features capture the global characteristics of the writer's individual writing habit and style. They are extracted from the entire document. Totally, there are 13 macro features including the initial eleven features reported in the previous study (6) and two new ones—stroke width and average word gap. The initial 11 features are entropy of gray values, binarization threshold, number of black pixels, number of interior contours, number of exterior contours, contour slope components consisting of horizontal (0°or flat stroke), positive (45° or 225°), vertical (90° or 270°), and negative (135° or 315°), average height, and average slant per line. In our current system, 11 of 13 macro features (except entropy and number of black pixels) are set to be the default features and were used in the experiments.

Micro features are designed to capture the finer details at the character level. Each micro feature is a gradient-based binary feature vector, which consists of 512 bits corresponding to gradient (192 bits), structural (192 bits), and concavity (128 bits) features, known as GSC features (6). Figure 1 shows three examples of the letter "e", the first two of which were written by the same person and the third was written by a different person. The pairwise score is positive for the first pair (indicating same writer) and negative for the other pairs (indicating different writers).

The use of micro features depends on the availability of recognized characters, i.e., character images associated with truth. Four possible methods are available in CEDAR-FOX to get recognized characters: (1) *manually crop characters* and label each with its truth; this labor-intensive method has the highest accuracy, (2) *automatically recognize characters* by using a built-in character recognizer; the method is error prone for cursive handwriting where there are few isolated characters, (3) *automatically recognize words* using a word-lexicon from which segmented and recognized characters are obtained; error rates depend on the size of lexicon which can be as high as 40% for a page, and (4) use *transcript mapping* to associate typed words in a transcript of the entire page with word images (16); it involves typing the document content once which can then be reused with each scanned document. As the full-page documents have the same content (CEDAR letter), the transcript mapping approach was used as shown in Fig. 2. This method has about 85% accuracy among words recognized. As most words are recognized, they are also useful for extracting features of letter pairs and whole words, as discussed next.

Style features are features based on the shapes of whole *words* and shapes of letter pairs, known as *bi-grams*. They are similar to micro features of characters as described below. Figure 3 shows three examples of word images, where (*a*) and (*b*) were written by the same writer and (*c*) was from a different writer. A bi-gram is a partial word that contains two connected characters, such as *th*, *ed*, and so on, as shown in Figs. 4*a–c*, where (*a*) and (*b*) were written by the same writer while (*c*) was from a different writer.

Style features are similar to micro features (i.e., GSC features) where a binary feature vector is computed from the bi-gram or word image. While the bi-gram feature has the same size as the micro feature (with 512 binary values), the word feature has 1024 bits corresponding to gradient (384 bits), structural (384 bits), and concavity (256 bits). As with micro features, to use these two style features, word recognition needs to be done first. As mentioned above when describing micro features, a transcript-mapping algorithm was used to do the word recognition automatically. The words being recognized are saved and used to compute the word-GSC features. Then characters are segmented from these words and the two consecutive segmented characters both being confirmed by a character recognizer are combined as one bi-gram component.

Similarity Computation—Given two handwritten documents, questioned and known, the system first extracts all the features mentioned above; therefore three types of feature vectors are generated for each document: macro, micro, and style features. To determine whether they are written by the same or different writers, the system computes the distance between all the feature vectors. For macro features, as they are all real-valued features, the distance is just the absolute value of their difference. For micro and style features, several methods have been recently evaluated (17), which has led to the choice of a so-called "Correlation" measure as being used to compute the dissimilarity between two binary feature vectors. It is defined as follows. Let $S_{ij}(i,j \in \{0,1\})$ be the number of occurrences of matches with i in the first vector and j in the second vector at the corresponding positions, the dissimilarity, or distance, D between the two feature vectors X and Y is given by the formula:

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

For example, to compute the dissimilarity between vector $X = [111000]$ and vector $Y = [101010]$, we have $S_{11} = 2$, $S_{10} = 1$, $S_{01} = 1$, $S_{00} = 2$, therefore $D(X, Y) = 1/3$. It can be observed that the range of $D(X, Y)$ has been normalized to [0,1]. That is, when $X = Y$, $D(X, Y) = 0$, and when they are completely different, $D(X, Y) = 1$.

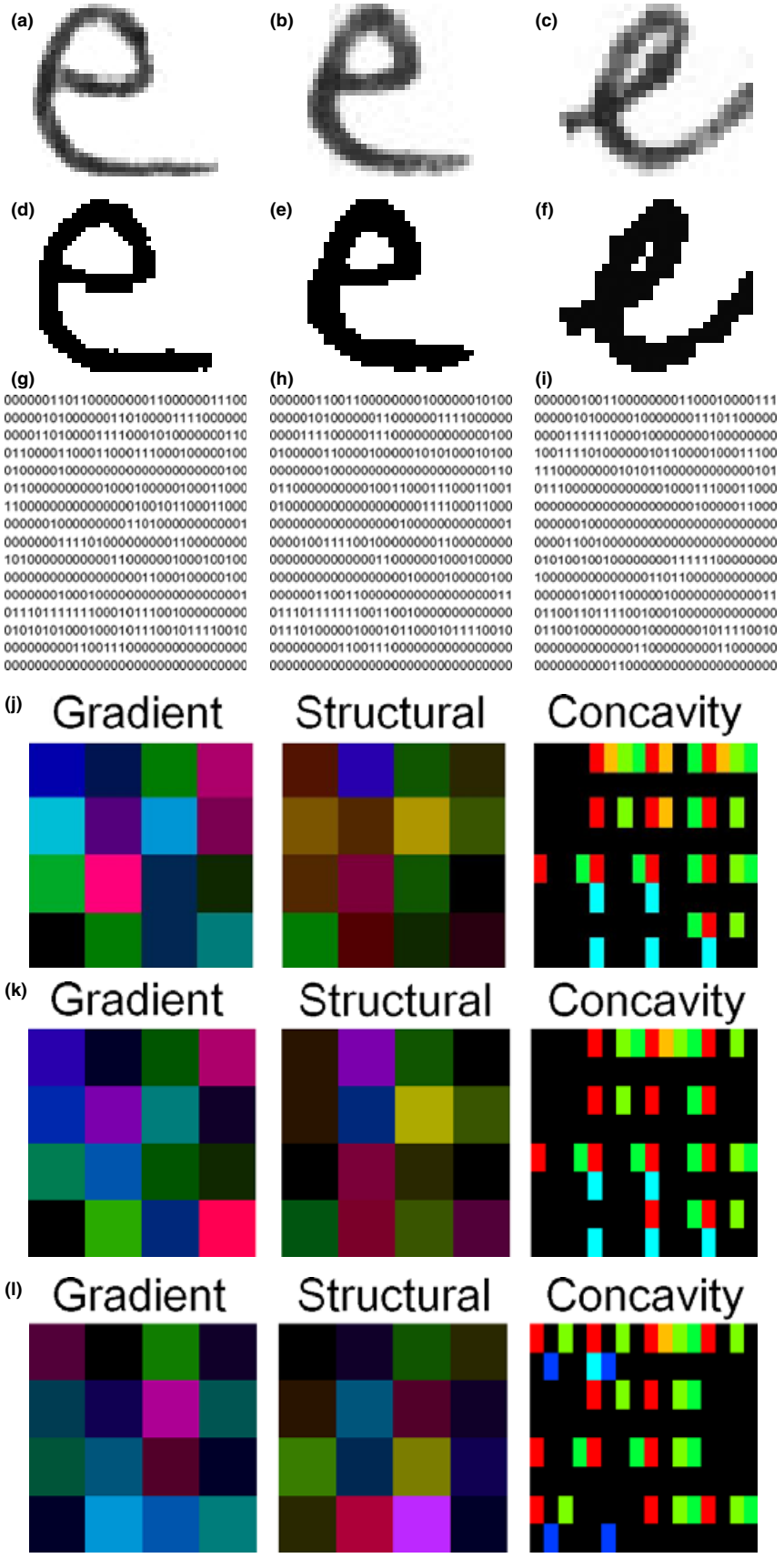

FIG. 1—*Micro features for three handwritten letters. (a)–(c) are grayscale images, (d)–(f) are corresponding thresholded binary images, (g)–(i) are corresponding GSC 512-bit vectors, (j)–(l) are color-coded versions of the bit vectors where the 192 bits of G and S are each sub-divided into 4×4 groups of 12 bits each (resulting in a 4096-color palette) and the 128 bits of C into an 8×16 pattern. The correlation distance between (a) and (b) is 0.16 while that between (a) and (c) is 0.43 and between (b) and (c) is 0.35. The LLR value between (a) and (b) is 1.49 while that between (a) and (c) is −0.97 and between (b) and (c) is −0.26.*

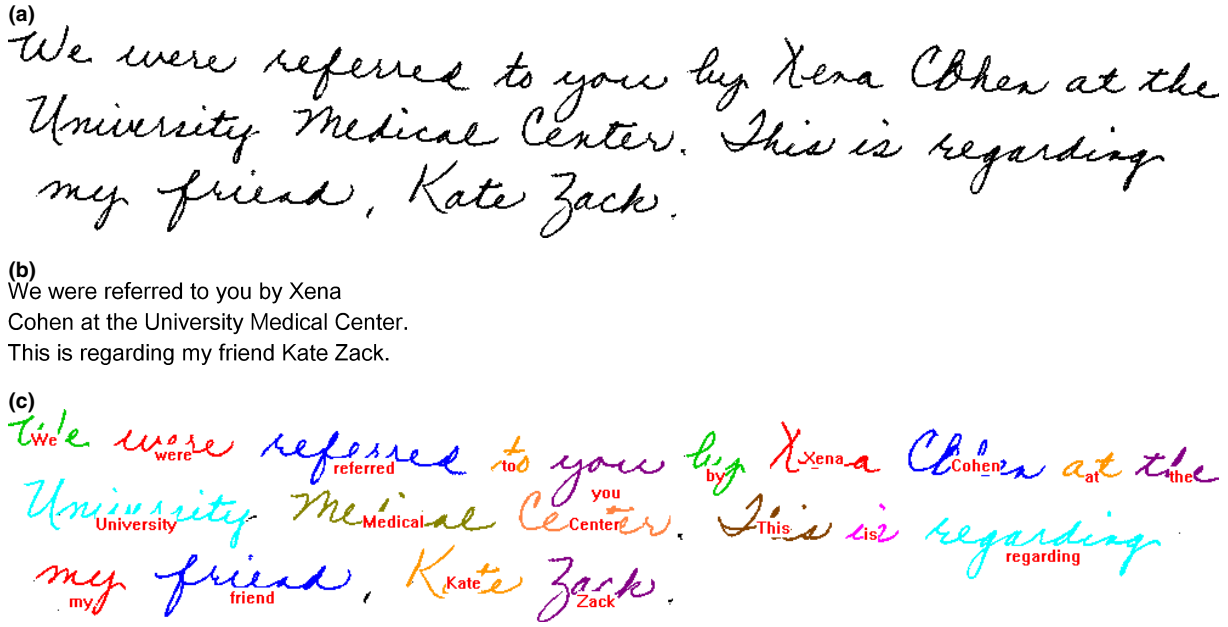

FIG. 2—*Transcript mapping: (a) handwritten text, (b) typed transcript, and (c) transcript-mapped handwritten text.*

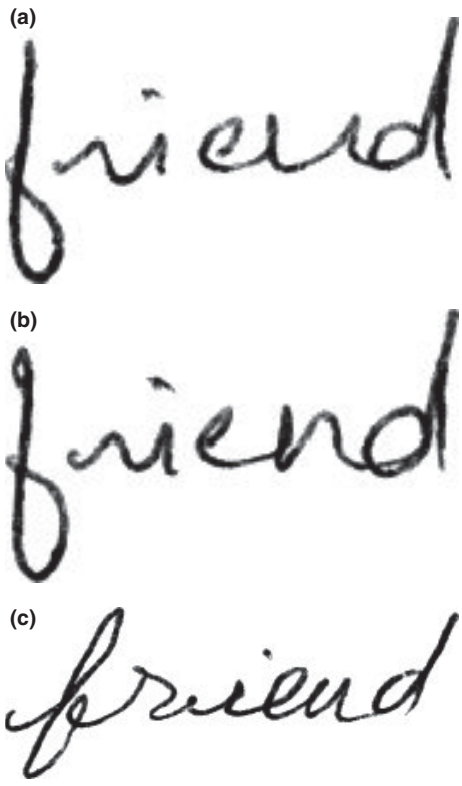

FIG. 3—*Word shapes: (a) and (b) are word samples from the same writer while (c) is a word sample from a different writer. The correlation distance between (a) and (b) is 0.2022 and between (a) and (c) is 0.3702. The LLR value between (a) and (b) is 4.44 and between (a) and (c) is −0.35.*

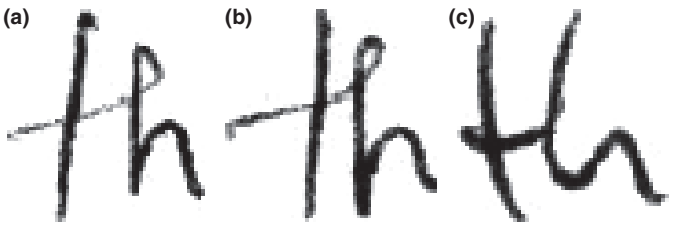

FIG. 4—*Bi-gram shape: (a) and (b) are two bi-gram samples from the same writer while (c) is a bi-gram sample from a different writer. The correlation distance between (a) and (b) is 0.1996 and between (a) and (c) is 0.3735. The LLR value between (a) and (b) is 4.35 and between (a) and (c) is −0.47.*

Statistical Model of Similarity

The distributions of dissimilarities, or distances, conditioned on being from the same or different writer are used to compute likelihood functions for a given pair of samples. The distributions can be learned from a training dataset and represented either as probability tables or as probability density functions. Probability distributions for discrete valued distances are represented as probability tables. In the case of continuous-valued distances, parametric methods are used to represent probability density functions. Parametric methods are more compact representations than nonparametric methods. For example, with the *k*-nearest-neighbor algorithm, a well-known nonparametric method, we need to store some or all training samples. This compactness results in a faster run-time algorithm. Also parametric methods are more robust in a sense that nonparametric methods are more likely to over-fit and therefore generalize less accurately. Of course the challenge is to find the right parametric model for the problem at hand. While the Gaussian density is appropriate for mean distance values that are high relative to the standard deviation, the Gamma density is more appropriate than the Gaussian when the distances are low relative to the standard deviation as it will not overlap with the negative part of the real line.

Assuming that the dissimilarity data can be acceptably represented by Gaussian or Gamma distributions, the probability density functions of distances conditioned upon the same-writer and different-writer categories for a single feature x have the parametric forms $p_s(x) \sim p(a_{\text{same}}, b_{\text{same}})$ and $p_d(x) \sim p(a_{\text{diff}}, b_{\text{diff}})$. For macro features, we model both categories by Gamma distribution as $p_s(x) \sim \text{Gamma}(\alpha_s, \beta_s)$ and $p_d(x) \sim \text{Gamma}(\alpha_d, \beta_d)$. For micro and

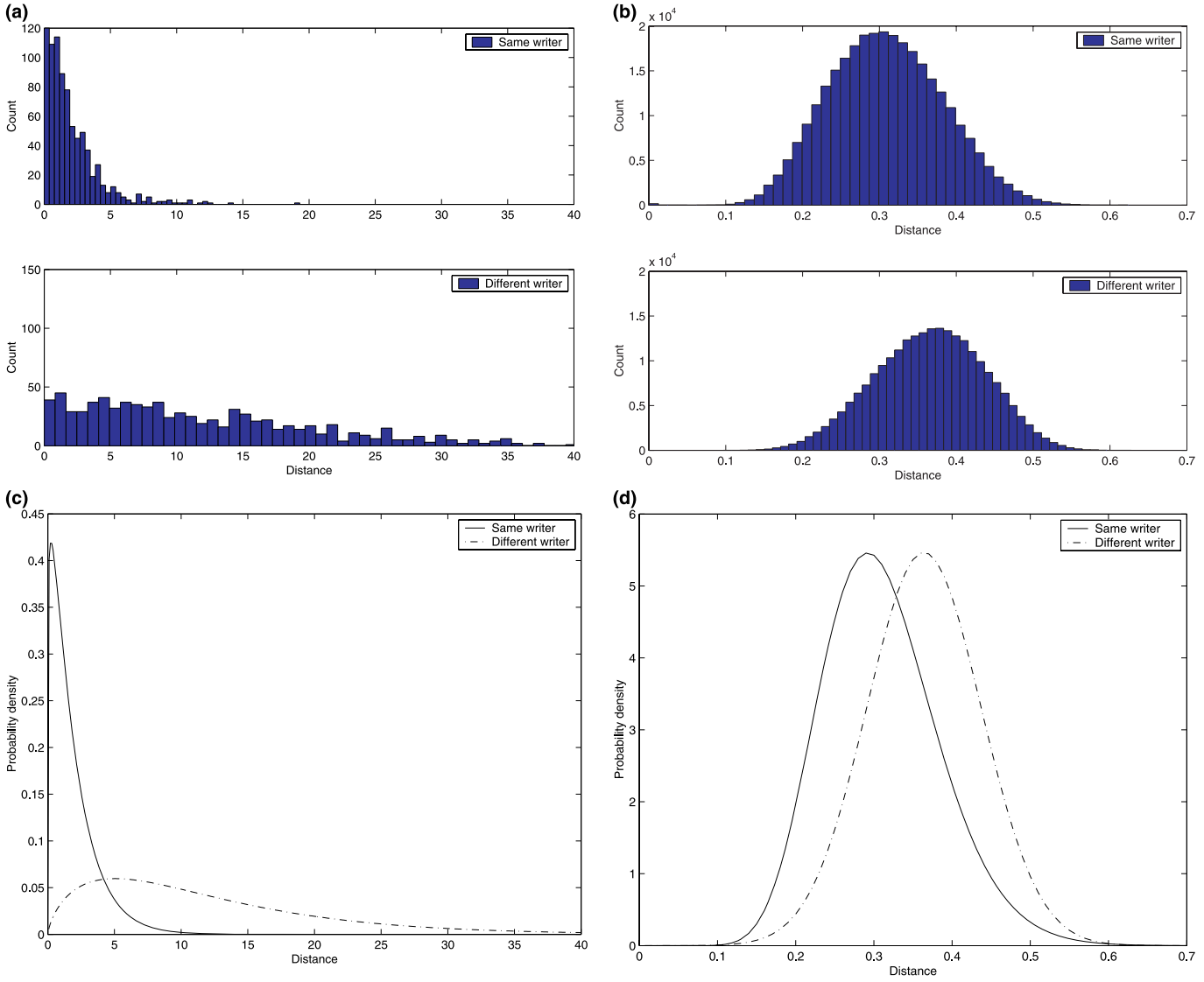

FIG. 5—*Histograms and parametric probability density functions (p.d.f.s): (a) same and different histograms for the "slant" (macro feature), (b) same and different histograms for the letter "e"(micro feature), (c) same and different p.d.f.s for "slant," and (d) same and different p.d.f.s for letter "e."*

style features, while the "same-writer" category is modeled as $p_s(x)$~Gamma(α_s, β_s) for Gamma distribution, the "different-writer" is modeled as $p_d(x) \sim N(\mu_d, \sigma_d^2)$ for Gaussian distribution. Dissimilarity histograms corresponding to the same writer and different writer distributions for "slant" (macro feature) and for the letter "e" (micro feature) are shown in Figs. 5*a* and 5*b*, respectively. Conditional parametric parametric probability density functions for "slant" and for "e" are shown in Figs. 5*c* and 5*d*.

The Gaussian and Gamma probability density functions are as follows:

$$\text{Gaussian}(x) = \frac{1}{(2\pi)^{1/2}\sigma} \exp^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ for } x \in R$$

$$\text{Gamma}(x) = \frac{x^{\alpha-1}\exp^{(-x/\beta)}}{(\Gamma(\alpha))\beta^{\alpha}} \text{ for } x > 0$$

Here μ and $\sigma > 0$ are the mean and the standard deviation of the Gaussian distribution. The parameters of the gamma distribution, $\alpha > 0$ and $\beta > 0$ can be evaluated from the sample mean $\hat{\mu}$ and sample variance $\hat{\sigma}^2$ as follows $\alpha = \hat{\mu}^2/\hat{\sigma}^2$ and $\beta = \hat{\sigma}^2/\mu$. "$\alpha$" is called the shape parameter and "β" is the scale parameter.

The current system has both micro and style features modeled as Gamma (for same-writer) and Gaussian (for different-writer) distribution. Macro features that have discrete values are modeled with a probability table and those that have continuous values are modeled as Gamma distributions. The features and the distributions by which they are modeled are summarized in Table 1.

Parameter Estimation

All statistical parameters for each of the features used by CEDAR-FOX were estimated by using maximum likelihood estimation. The training data are a set of learning, or design, samples provided by 1000 nontwin writers who provided three samples each. Each document is a handwritten copy of a source document in English, known as the CEDAR letter (6). The source document is concise and complete in that it captures all characters (all numerals, small case and upper case English letters), punctuations and distinctive letter and numeral combinations (ff, tt, oo, 00). The learning set provided by 1000 writers is a subset of samples

TABLE 1—*Modeling distributions for all the features.*

Features	Type of Distribution	Modeled Using
Entropy	Continuous	Gamma
Threshold	Discrete	Probability table
No. of black pixels	Continuous	Gamma
No. of exterior contours	Discrete	Probability table
No. of interior contours	Discrete	Probability table
Horizontal slope	Continuous	Gamma
Positive slope	Continuous	Gamma
Vertical slope	Continuous	Gamma
Negative slope	Continuous	Gamma
Stroke width	Discrete	Probability table
Average slant	Continuous	Gamma
Average height	Discrete	Probability table
Average word gap	Continuous	Gamma
Micro features	Continuous	Gamma-Gaussian
Bi-gram	Continuous	Gamma-Gaussian
Word	Continuous	Gamma-Gaussian

provided by 1568 nontwin writers. The remainder of the samples was kept aside for testing as nontwins data.

The list of macro features modeled as Gamma distribution and their estimated parameters are given in Table 2. The parameters were estimated by using 3000 pairs of half-page documents including 1500 from the same writer and 1500 from different writers.

Likelihood Ratio Computation

When two handwritten document images, which are labeled as known and questioned, are presented to the CEDAR-FOX verification subsystem, the system segments each document into lines and words. A set of macro, micro, and style features are extracted that capture both global characteristics from the entire document and local characteristics from individual characters, bi-grams, and words. The system has available to it a set of statistical parameters for each of the macro, micro, and style features. These parameters are used to determine the probabilities of the differences observed in the value of the feature for each of the two documents, if they are from the same writer distribution or different writer distributions.

In proceeding further, it is noted that the probability of a given pair of documents belonging to the same or different writer is modeled as the joint probability of the feature distances between the two documents. When more than one feature is used, we make the assumption that feature distances are statistically independent, although there do exist correlations between some pairs of feature distances as can be seen from the scatter plots for the first 12 macro features in Fig. 6. A more complex model to account for correlations, i.e., a Bayesian neural network, got an accuracy improvement of 1–2% on a particular dataset, which was not significant. There are several other reasons for making the independence assumption for an interactive system for the comparison of handwriting: first, the features used can be modified interactively, e.g., because LLR are additive, it allows us to observe the effects of adding and removing features on system performance; on the other hand a neural network would have to be retrained for each feature combination which is infeasible when there are a large number of features. Second, as has been observed in other machine learning tasks, more complex models tend to overfit to the data, which can lead to poorer performance on large amounts of unseen data. Third, its simplicity goes back to not only Occam's razor but also to traditional QDE literature suggesting the multiplying of probabilities of handwriting elements (1).

TABLE 2—*Gamma parameters for continuous macro features.*

	Shape Parameter, α		Scale Parameter, β	
Feature	Same Writer	Different Writer	Same Writer	Different Writer
Entropy	0.8036	1.0462	0.0302	0.0263
No. of black pixels	1.7143	1.4838	1736.6	3520.0
Horizontal slope	1.2237	1.6041	0.0199	0.0444
Positive slope	0.9005	1.8346	0.0187	0.0569
Vertical slope	1.5775	1.8138	0.0150	0.0523
Negative slope	0.9506	1.3219	0.0100	0.0426
Average slant	1.1471	1.7135	1.6609	7.0830
Average word gap	3.4487	2.9363	0.0128	0.0372

Each of the two likelihoods that the given pair of documents was either written by the same individual or by different individuals can be expressed, assuming statistical independence of the distances of features, as follows. For each writing element e_i, $i = 1,\ldots, c$, where c is the number of writing elements considered, we compute the distance $d_i(j, k)$ between the jth occurrence of e_i in the first document and the kth occurrence of e_i in the second document for that writing element. The likelihood is extracted as

$$L_{\mathrm{s}} = \prod_{i=1}^{c} \prod_{j} \prod_{k} p_{\mathrm{s}}(d_j(j,k))$$

$$L_{\mathrm{d}} = \prod_{i=1}^{c} \prod_{j} \prod_{k} p_{\mathrm{d}}(d_j(j,k))$$

The LLR in this case has the form

$$\mathrm{LLR} = \sum_{i=1}^{c} \sum_{j} \sum_{k} [\ln p_{\mathrm{s}}(d_i(j,k)) - \ln p_{\mathrm{d}}(d_i(j,k))]$$

That is, the final LLR value is computed using all the features, considering each occurrence of each element in each document. The CEDAR-FOX verification system outputs the LLR of two documents being tested. When the likelihood ratio (LR) is above 1, the LLR value is positive and when the LR is below 1, the LLR value is negative. Hence, if the final score obtained is positive, the system concludes that the two documents were written by the same writer. Similarly, if the final LLR score is negative, the system concludes that the two documents were written by different writers.

Correction to Tails

Finally, it is necessary to introduce a correction to the computed LLR at the extremely unlikely tail regions of the distributions. For a given feature, when the distance between two elements being matched decreases we expect the score to be a positive value that is increasing. Similarly when the distance increases, we expect the LLR to be a negative value that decreases. However, for very small distance values and very large distance values, which lie at the tails of the Gamma and Gaussian distributions, the asymptotic behavior can be anomalous. Figure 7*a* shows an example of the original relationship between LLR score and the distance for letter "e" based on the model as shown in Fig. 5*d*. From the figure, it can be observed that for small values of distance (<0.17 in this case) when decreasing the distance the LLR value also drops, while for large value of distance (>0.49 in this case) when increasing the distance the LLR value starts to increase monotonically. This is because we use a parametric method (Gaussian or Gamma) to model the data.

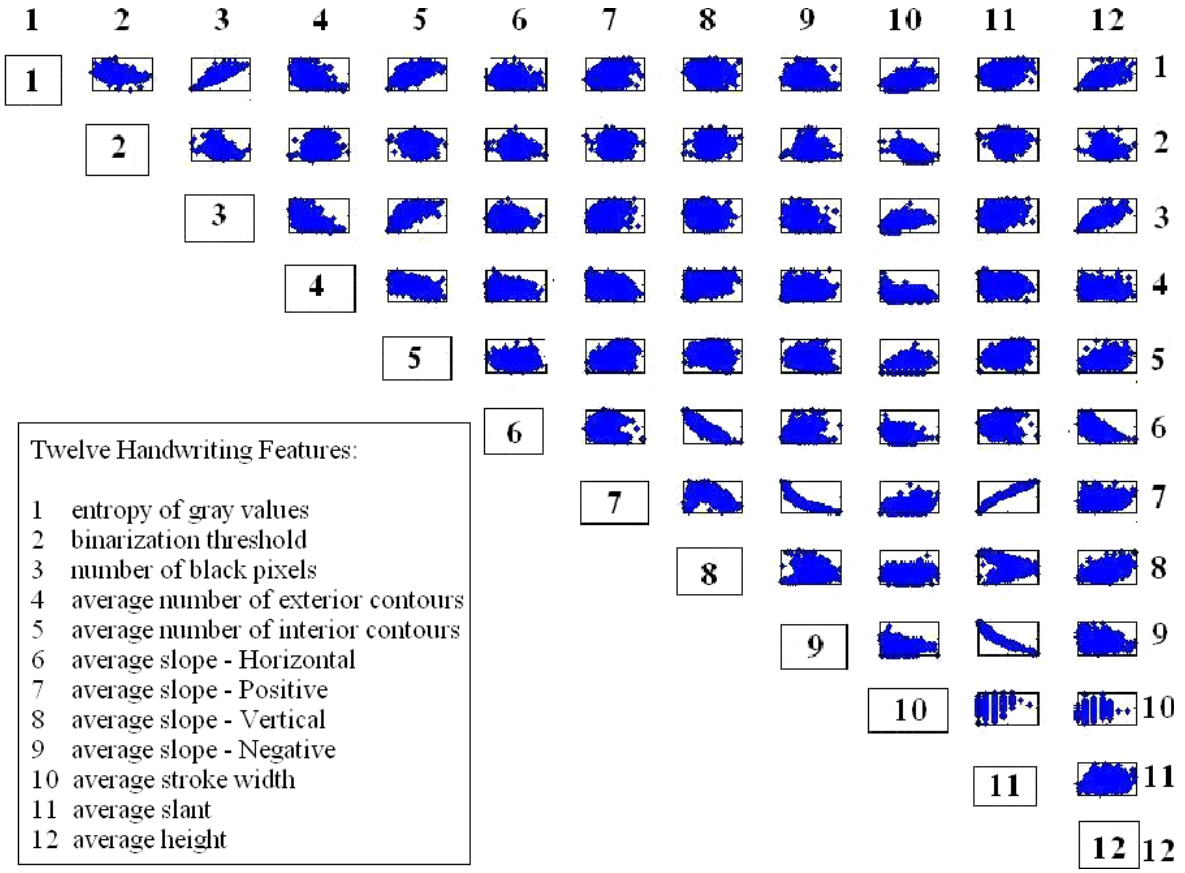

FIG. 6—*Scatter plot for 12 macro features.*

The probability density functions learned from the training set are approximations whose interactions in the tail region are not modeled accurately. This phenomenon is a consequence of distributions and parametric models that assign nonzero values to the distribution asymptotically. Therefore, it is necessary to correct the LLR value in the boundary cases so that we have only one decision boundary. The adjusted model has a distance-LLR relationship as shown in Fig. 7*b*. For small values of distance, we assign a fixed LLR value that is the maximum value in the region. Similarly for large value of distance we assign the minimum value in that region. Lacking a closed form solution for the distance value at which the maximum and minimum values of the LLR are reached, we obtain them from the LLR-distance curve. If the maximum value of LLR, i.e., LLR_{max}, occurs at d_{max} and the minimum value LLR_{min} at d_{min}, then for any $d < d_{max}$ $LLR = LLR_{max}$ and for any $d > d_{min}$ $LLR = LLR_{min}$.

Twins Test-bed

The purpose of this study was to determine the performance of CEDAR-FOX on handwriting samples provided by twins. Handwritten documents from 206 pairs of twins and nontwins were collected. None of the twins' handwriting was used in the design of the CEDAR-FOX system and hence the results on twins provide an objective evaluation of the system.

Demographic Distribution of Twins

The demographic data for the individuals who contributed to the twins' dataset are as follows:

1. Location: Samples were obtained from people coming from at least 150 different cities and seven different countries.
2. Age: The age distribution of the sample pairs is as shown in Fig. 8. Here age is divided into intervals with the first interval representing age from 0–10, second from 10–20 and so on. As seen from the graph, age from 20–30 is the interval with the most twin pairs.
3. Schooling: All the twins attended the same school as their twin pair. Sixteen percent had private schooling while the rest (84%) had public schooling.
4. Handedness: Of the 412 test cases, 15 (3.6%) were ambidextrous, 45 (10.9%) were left-handed, and 352 (85.5%) were right-handed. Among twins, 36 (17.5%) pairs had different handedness.
5. Type of twins: Out of the 206 pairs of twins, 31 (15.05%) were fraternal, 169 (82.04%) were identical, and the remaining six (2.91%) pairs were unsure.
6. Sex: Sixty-nine (16.7%) test cases were male and 343 (83.3%) female.
7. Injuries: Twenty-one (5.1%) had injuries that could affect their handwriting.

Writing Styles

The handwriting samples were divided into three different categories based on the style of writing: both twins used printing (including upper case printing), both used cursive, one used printing and the other used cursive (mixed). Table 3 shows the number of twin pairs for each category.

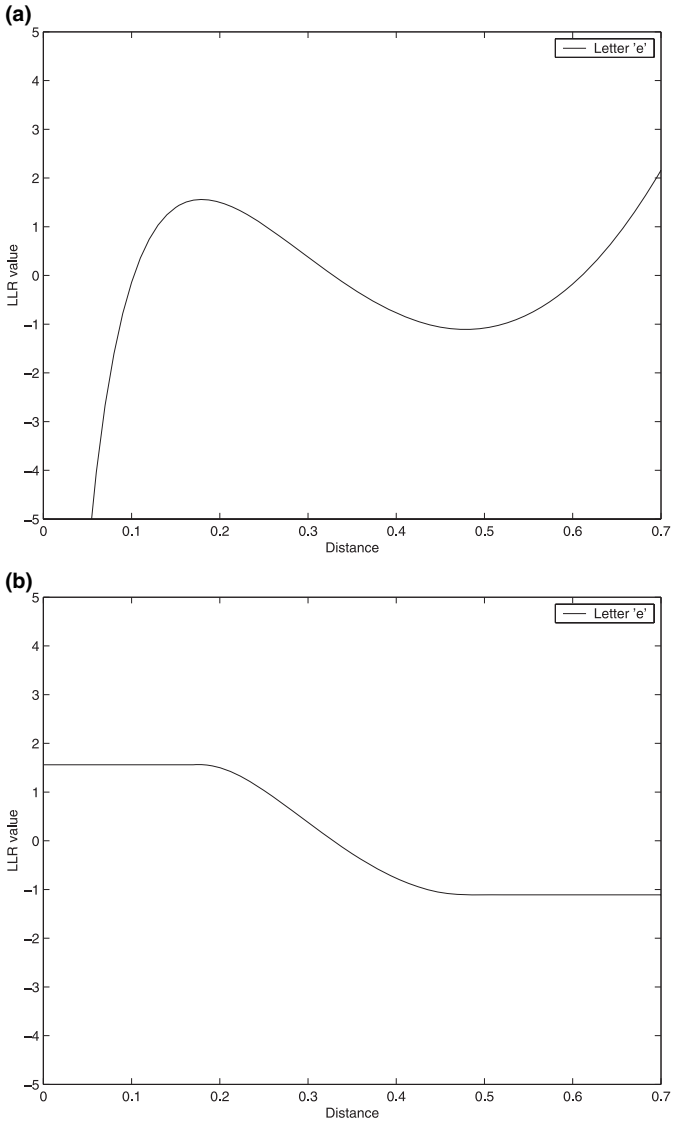

FIG. 7—*Relationship between LLR value and the distance for letter "e." (a) The original one and (b) the adjusted one.*

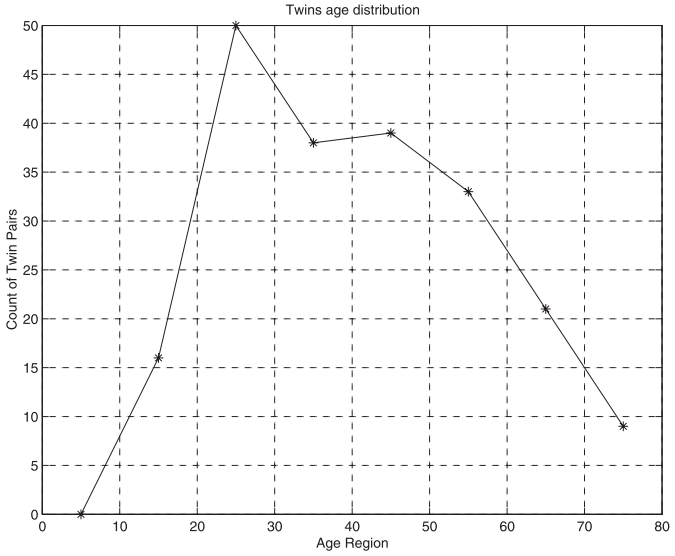

FIG. 8—*Age distribution among twins in database.*

TABLE 3—*Distribution of writing styles among twins.*

Both Printing				
Normal	All Upper Case	Both Cursive	Used Two Different Style (Mixed)	Total
36	2	128	40	206

Document Content

As in the case of the design samples, each twin participant was required to copy the CEDAR letter once in his/her most natural handwriting using plain unlined sheets, and a medium black ballpoint pen. Portions of the CEDAR letter for two pairs of twins having similar and dissimilar handwriting are shown in Figs. 9 and 10. They are also examples of Different-writer Same-content (DS) test (see details discussed in next section). The similarity scores of these two pairs of documents are shown in Tables 4 and 5.

Verification Test Cases

Two hundred and six pairs of twins provided one sample each. Thus a total of 412 (206×2) documents were obtained. Each of the document images was cut into two roughly, logically equal size images—the upper and lower parts—and these half page images were used for testing same and different document content. Two documents images, corresponding to the upper and lower parts from the same writer, which is one test case of "Same-writer Different-content" (SD) are shown in Fig. 11. The similarity scores of these two documents are shown in Table 6.

Figure 12 shows the schematics of the method by which the test cases were generated. Verification test scenarios are therefore divided into the following four classes:

1. Same-writer, Same-content (SS): The two samples are from the same writer having the same textual content.
2. SD: The two samples are from the same writer having different textual content.
3. DS: The two samples are from different writers having the same textual content.
4. Different-writer, Different-content (DD): The two samples are from different writers having different textual content.

Note that SD and DD are complementary in that both involve different content data. The average of the SD and DD cases gives the overall accuracy on different content data. The DS does not have a complementary SS case as there was only one sample provided by each twin writer. For nontwins, as we have three samples from each writer, we generated another set of 412 half-page documents from the original dataset to form the SS test cases, as shown in Fig. 12*b*. The numbers of test cases for each of the four scenarios are given in Table 7.

Thus, from 824 half-page documents of the twins, we generated a total of 1236 (412 + 412 + 412) test cases for the experiments. A total of 1648 test cases were obtained from a different set of 412 writers taken from the general population samples.

Testing Process

CEDAR-FOX version 0.53 was used for all testing. The verification process was run with all the test cases specified. The results of these test cases were in the LLR format as explained above. Scatter plot graphs were made of the final LLR score to show the

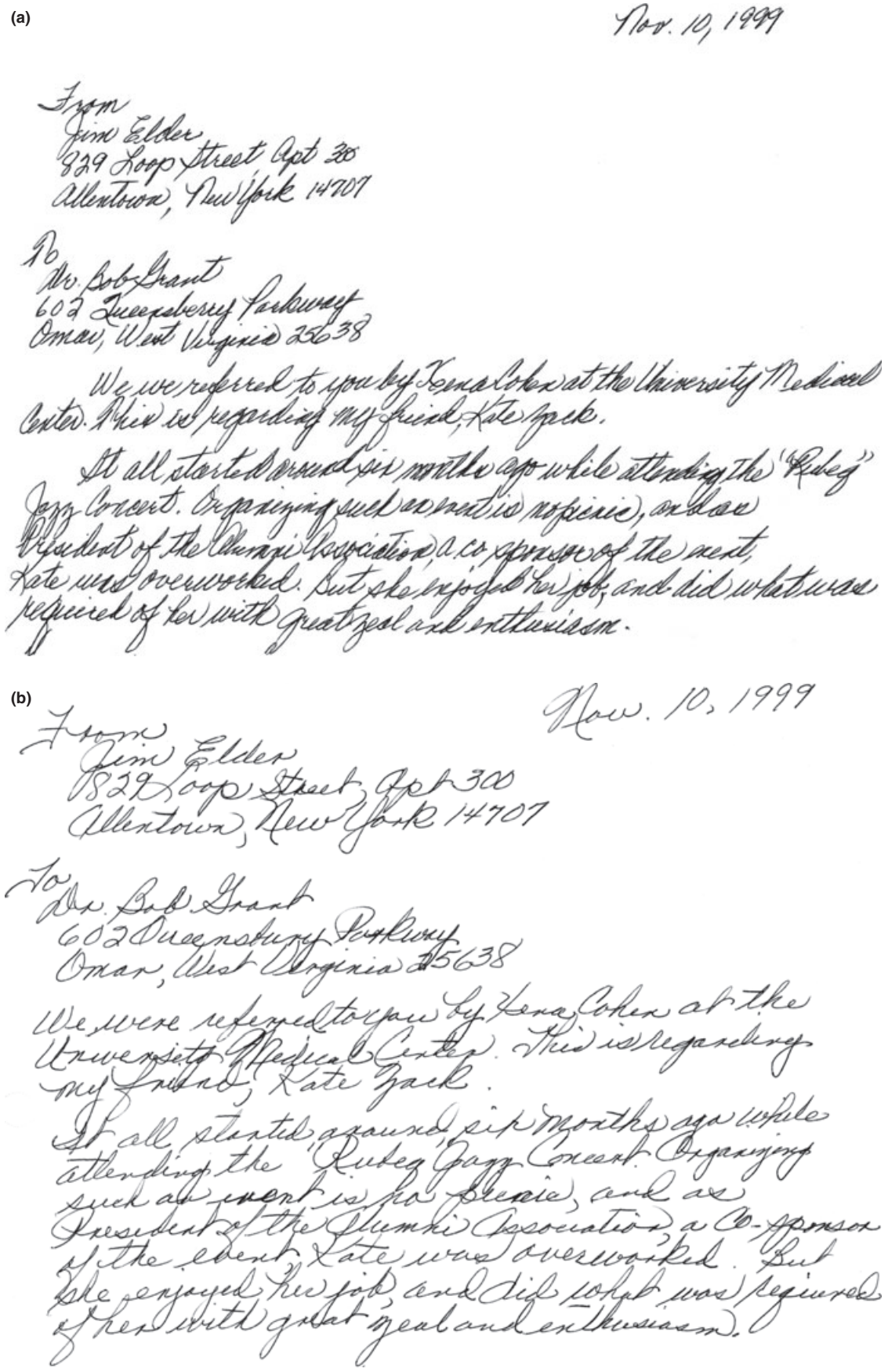

FIG. 9—*Similar handwriting samples of a pair of twins: (a) Twin 003a and (b) Twin 003b. The LLR value between these two documents is 7.15.*

(a)

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

Nov. 10, 1999

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

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

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,

(b) FROM

Jim Elder
829 Loop Street Apt 300
Allentown, New York 14707

NOV 10, 1999

TO
Dr. Bob Grant
602 Cranberry Parkway
Omar, West Virginia 25638

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

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 enjoy her job, and did what was required of her with great zeal and enthusiasm.

FIG. 10—*Dissimilar handwriting samples of a pair of twins: (a) Twin 006a and (b) Twin 006b. The LLR value between these two documents is −98.36.*

TABLE 4—*The similarity table for two document samples: Twin003a and Twin003b.*

Feature	No. of Comparisons	Distance	LLR Value
Threshold	N/A	1	0.49
No. of exterior contours	N/A	2	0.38
No. of interior contours	N/A	1	0.83
Horizontal slope	N/A	0.07	−1.43
Positive slope	N/A	0.06	−1.83
Vertical slope	N/A	0.02	1.24
Negative slope	N/A	0.01	0.76
Stroke width	N/A	3	−5.11
Average slant	N/A	2.45	1.11
Average height	N/A	3	0.63
Average word gap	N/A	0.03	1.22
Micro features	10	N/A	2.68
Bi-gram	0	N/A	0
Word	9	N/A	6.17
Total	N/A	N/A	7.15

LLR, log-likelihood ratio.

TABLE 5—*The similarity table for two document samples: Twin006a and Twin006b.*

Feature	No. of Comparisons	Distance	LLR Value
Threshold	N/A	0	1.44
No. of exterior contours	N/A	3	−0.09
No. of interior contours	N/A	8	−0.63
Horizontal slope	N/A	0.14	−3.63
Positive slope	N/A	0.1	−3.61
Vertical slope	N/A	0.1	−3.14
Negative slope	N/A	0.15	−10.96
Stroke width	N/A	1	−0.41
Average slant	N/A	22.41	−9.35
Average height	N/A	10	−2.24
Average word gap	N/A	0.07	−0.14
Micro features	256	N/A	−22.07
Bi-gram	0	N/A	0
Word	38	N/A	−43.54
Total	N/A	N/A	−98.36

LLR, log-likelihood ratio.

distribution. The accuracy of each individual feature in discriminating among writers as well as the accuracy of the system, when all the features are considered simultaneously, is calculated.

Verification Results

The outcome of a writer verification experiment can be one of four possibilities: (1) *true positive*: documents written by the same writers and it is confirmed by the results; (2) *true negative*: documents written by different writers and it is confirmed by the results; (3) *false positive*: documents written by different writers but the results concluding them to be written by the same writer; and (4) *false negative*: documents written by the same writers but the results concluding that they are written by different writers. True positive and true negative are correct results while false positive and false negative are erroneous results. In CEDAR-FOX same/different decision is based on computing the LLR value.

Scatter Plots

The test results for the data were plotted as a scatter plot graph where the *x*-axis represents the number of test cases and the *y*-axis represents the LLR values obtained. LLR values indicate the extent of similarity between the writers. The higher the LLR, the greater is the similarity between the writers, while low LLR indicates dissimilarity within the writers.

Different Content (SD + DD)—For different content tests, when using 11 macro features plus micro and style features (which gives the best performance, as we will see in the next section), the LLR values range from −39.25 to 197.95 for SD data with the majority of them lying in the 0–50 range (as shown in Fig. 13). Test cases having LLR values less than 0 are false negatives. For DD data, LLR values range from −112.62 to 54.36 with most of them lying in the −70 to 10 range.

Different Writer (DS + DD)—For DS, the LLR values range from −243.24 to 138.4. The majority of the cases lie in the −75 to 20 range. As seen from Fig. 14, DS data have more false positives than DD. This can be attributed to the pictorial effects where different content appears as different pictorial content.

Verification Error Rates with Twins and Nontwins

The error rates of verification using different content data are given in Table 8. This is the case when there is no rejection, i.e., in each case a hard decision is made whether the writers are the same or different. As shown in the table, using the combination of all the features together shows an overall improvement over any of the individual features. The overall error rate for twins is 12.6%. In comparison the overall error rate for nontwins is 3.15%.

Error rates using same content data are shown in Table 9. Because only a single document per twin was available, the SS testing was performed only on nontwins as shown in Fig. 12*b*. However, as the same writer testing is independent of twins, we used the same values (column 4) for both twins and nontwins to compute overall error rates (columns 5 and 6).

The overall error rates for twins versus nontwins can be observed by combining the results of same content (Table 8) and different content (Table 9). When using all features the average overall error rate for twins is 12.91% (average of 12.6% and 13.22%) and the overall error rate for nontwins is 3.7% (average of 3.15% and 4.24%). This attests to the fact that twins' handwriting is indeed more similar than that of nontwins.

The error rate of the system can be decreased by rejecting test cases that lie in a region of LLR ambiguity. In this experiment, the rejection intervals are chosen symmetrically around LLR = 0. A bigger interval will lead to a higher rejection rate and a lower error rate. As the final LLR value is the summation of all the features considered, the upper and lower thresholds of LLR values for rejection depend on how many features are used. In the current system, the default features setting is the combination of all three types of features, i.e., macro, micro, and style features. There are another two factors that could affect the selection of rejection intervals. The first one is whether the compared documents have same (or different) content. Because same content documents have more common elements to be compared, generally they have a larger range of the final LLR values. Therefore, to having a same rejection rate (for example 20%), one needs to choose a bigger interval for same content cases (−20 to 20 in this case) and a smaller interval for different content cases (−10 to 10 in this case). The second factor is how difficult the testing cases are. For example, in our experiments, it can be observed that, with a same rejection interval twin cases have a much higher rejection rate, which also means that handwriting of twins are more similar than that of nontwins.

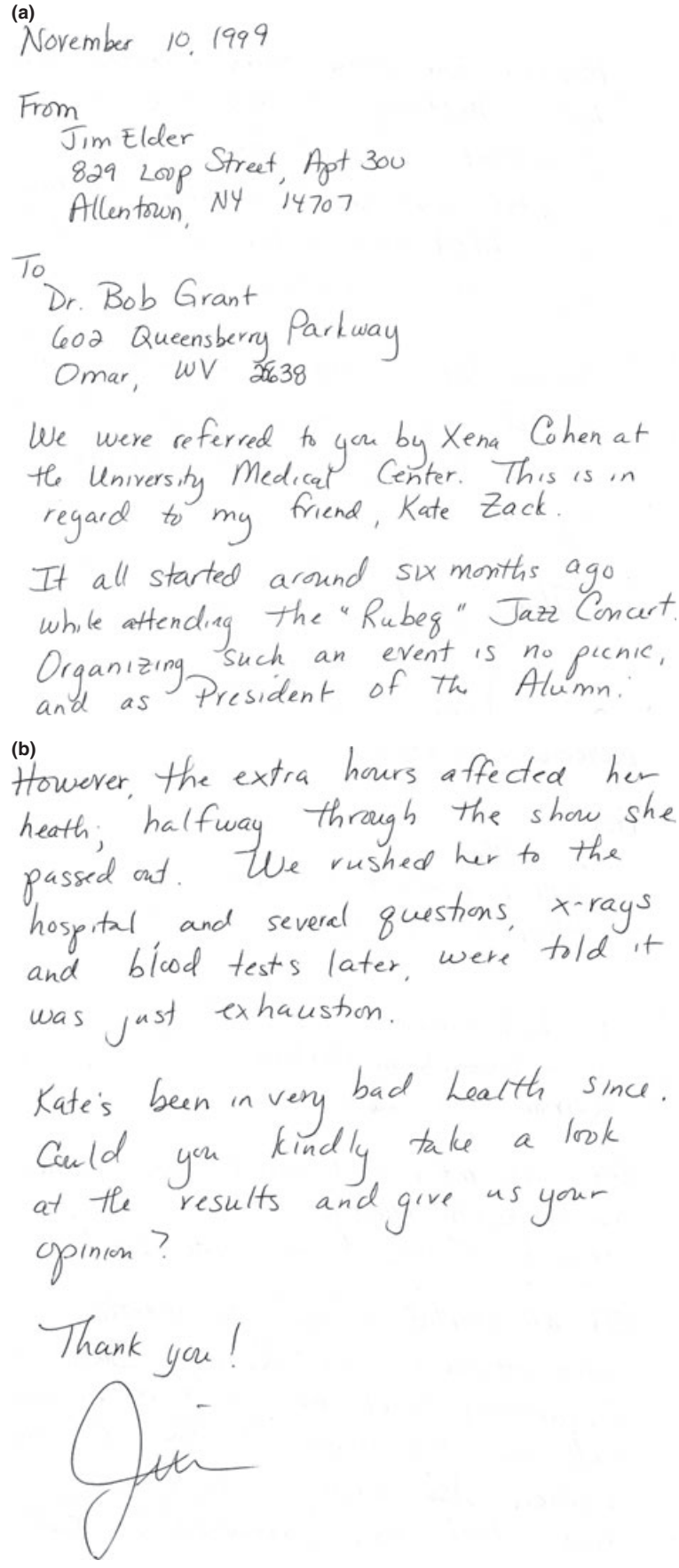

FIG. 11—*Example of Same-writer Different-content (SD) data obtained by dividing the full page of the CEDAR letter (Twin 178a) into (a) upper half and (b) lower half. The LLR value between these two documents is 34.19.*

TABLE 6—*The similarity table for two document samples: Twin178a-U and Twin178b-U.*

Feature	No. of Comparisons	Distance	LLR Value
Threshold	N/A	2	−1.53
No. of exterior contours	N/A	1	0.82
No. of interior contours	N/A	0	1.1
Horizontal slope	N/A	0.01	0.88
Positive slope	N/A	0.01	1.3
Vertical slope	N/A	0	2.15
Negative slope	N/A	0	1.14
Stroke width	N/A	0	0.65
Average slant	N/A	1.91	1.5
Average height	N/A	1	1.03
Average word gap	N/A	0.05	0.65
Micro features	1117	N/A	13.81
Bi-gram	6	N/A	3.23
Word	18	N/A	−1.67
Total	N/A	N/A	25.05

LLR, log-likelihood ratio.

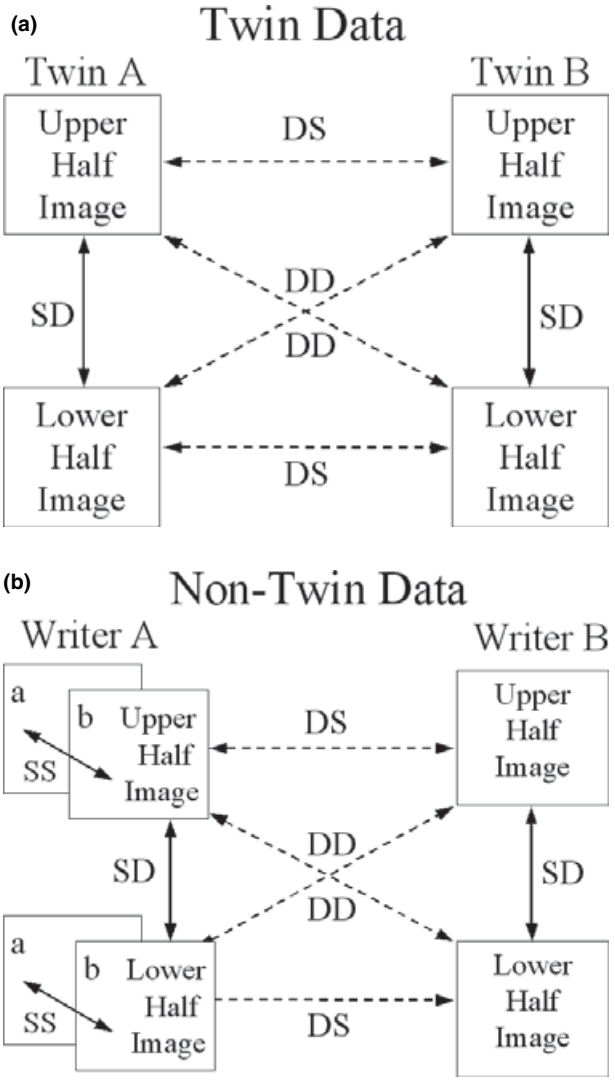

FIG. 12—*Generating test cases for (a) twin samples and (b) nontwin samples: each bidirectional arrow represents 206 test cases.*

TABLE 7—*Distribution of test cases in four verification scenarios.*

Verification Scenarios	Number of Half-page Document Pairs: Twins	Nontwins
Same-writer Different-content	412	412
Different-writer Different-content	412	412
Different-writer Same-content	412	412
Same-writer Same-content	0	412

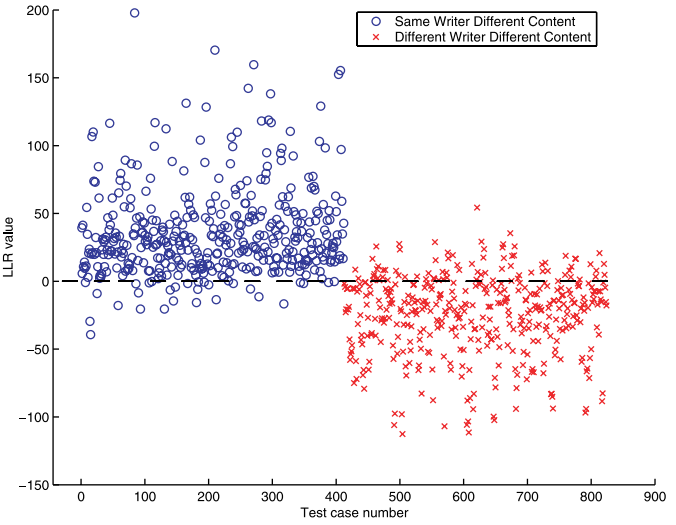

FIG. 13—*Scatter plot for twins with different content (SD and DD). Same writer (SD) has a lower error rate than different writer (DD).*

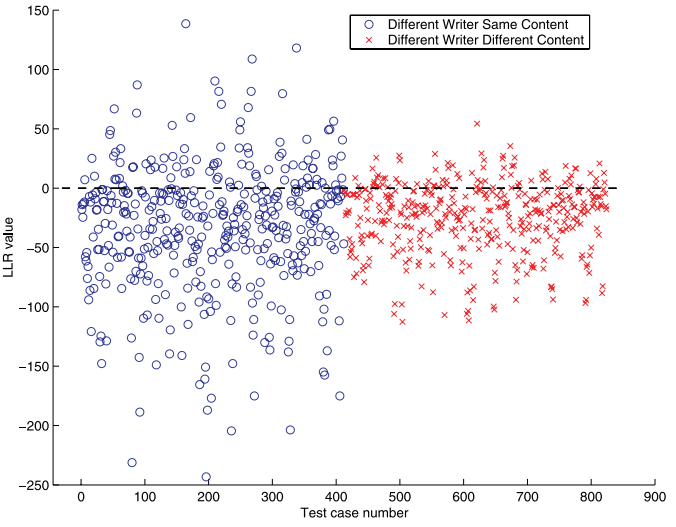

FIG. 14—*Scatter plot for different writers with same and different content (DS and DD). Same content data (DS) has a slightly higher error rate than different content data (DD).*

In this experiment, for different content testing, the rejection range was set to be from −20 to 20. For each rejection interval, the error and reject rates were calculated as shown in Table 10.

Figure 15 shows the error rates with the increasing reject rates for twins as well as nontwins. Error rate for twins can be reduced from 12.6% to 3.73% by rejecting 40% of the cases. Similarly, the

TABLE 8—*Verification error rates (%) for twins and nontwins using different content data with no rejection.*

	Different-writer Different-content (DD)		Same-writer Different-content (SD)		Overall (DD + SD)	
Feature under consideration	Twin	Nontwin	Twin	Nontwin	Twin	Nontwin
11 Macros	22.57	7.03	4.85	6.07	13.71	6.55
Micro	33.92	11.88	14.56	12.14	24.24	12.01
Bi-gram (BG)	26.16	10.31	20.00	16.15	23.08	13.23
Word (WD)	11.73	1.46	32.11	31.30	21.92	16.38
Macro + Micro + BG + WD	18.89	2.17	6.31	4.13	12.60	3.15

TABLE 9—*Verification error rates (%) for twins and nontwins using same content data with no rejection.*

	Different-writer Same-content (DS)		Same-writer Same-content (SS)	Overall (DS + SS)	
Feature under consideration	Twin	Nontwin	Nontwin	Twin	Nontwin
11 Macros	26.21	6.31	4.37	15.29	5.34
Micro	36.64	9.22	7.52	22.08	8.37
Bi-gram (BG)	33.85	9.19	8.57	21.21	8.88
Word (WD)	22.56	3.16	10.44	16.50	6.80
Macro + Micro + BG + WD	21.83	3.87	4.61	13.22	4.24

error rate for nontwins can be reduced from 3.15% to 0.14% by rejecting about 25% of the test cases.

For same content testing, the rejection range was set to be from −45 to 45. For each rejection interval, the error and reject rates were shown in Table 11. Error rate for twins can be reduced from 13.22% to 6.23% by rejecting 38% of the cases. Similarly, the error rate for nontwins can be reduced from 4.24% to 1.04% by rejecting about 18% of the test cases. The error rate can be further reduced but at the expense of rejecting a higher number of cases.

Verification Accuracy for Different Writing Styles

Each handwriting sample was assigned one of three styles: cursive, handprint, or mixed. The following three categories were evaluated separately: cursive (both twins used cursive handwriting), hand-print (both used hand-print), and mixed (twins used two different styles). The distribution of writing styles among the twin data was shown in Table 3.

Table 12 gives verification results based on writing style. The overall error rates of verifying cursive, hand-print, and mixed handwriting using different content are 15.82%, 15.13%, and 2.5%, respectively. As seen in the table, the performance of hand-print is roughly the same (marginally better) than that of cursive. In DD testing, the error rate is zero for mixed style because the system always prefers a "different writer" answer when the styles of two documents are different. Such cases are usually rejected (offer no opinion) by QDE in practice. An automatic reject option can be put into CEDAR-FOX to reject cases where there are mixed types of writing styles, particularly as it is already able to make an assignment on the print-cursive scale.

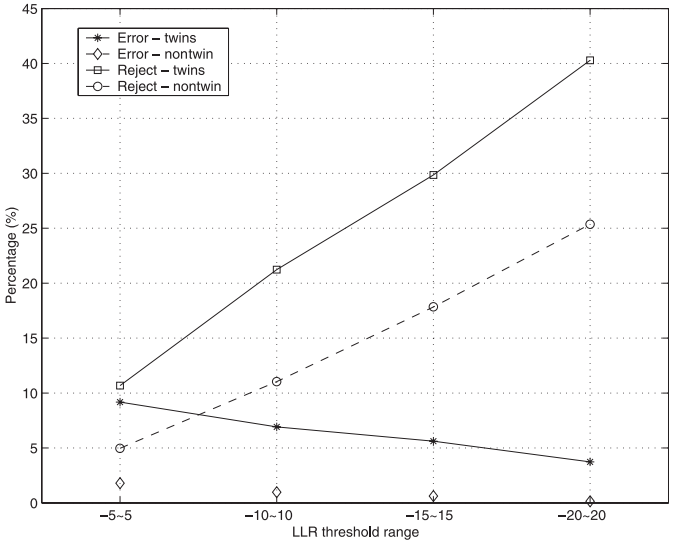

FIG. 15—*Verification error-reject rates for twins and nontwins using different content data.*

Verification Accuracy for Identical and Fraternal Twins

Identical and fraternal twins were tested separately to determine verification accuracy. The distribution of different types of twins was given in the previous section (31 identical, 169 fraternal, and six unsure). Verification error rates for each test category are shown in Table 13.

The average error rate for identical twins is 20.43% and the average for fraternal twins is 11.29%. This is a statistically

TABLE 10—*Verification error rates for twins and nontwins using different content data with rejection.*

	Diff-writer Diff-content (DD)				Same-writer Diff-content (SD)				Overall (DD + SD)			
	Twin		Nontwin		Twin		Nontwin		Twin		Nontwin	
LLR rejection interval	Error	Reject	Error	Reject	Error	Reject	Error	Reject	Error	Reject	Error	Reject
−5 to 5	14.45	14.32	1.24	2.43	3.9	7.04	2.36	7.52	9.17	10.68	1.8	4.97
−10 to 10	10.96	26.94	0.78	6.07	2.87	15.53	1.16	16.02	6.91	21.23	0.97	11.04
−15 to 15	8.99	35.19	0.27	10.19	2.25	24.51	0.98	25.49	5.62	29.85	0.62	17.84
−20 to 20	6.02	47.57	0.29	15.29	1.45	33.01	0.0	35.44	3.73	40.29	0.14	25.36

TABLE 11—*Verification error rates for twins and nontwins using same content data with rejection.*

	Diff-writer Same-content (DS)				Same-writer Same-content (SS)		Overall (DS + SS)			
	Twin		Nontwin		Nontwin		Twin		Nontwin	
LLR rejection interval	Error	Reject	Error	Reject	Error	Reject	Error	Reject	Error	Reject
−5 to 5	22.04	9.71	2.72	1.7	4.44	1.7	13.14	5.7	3.58	1.7
−10 to 10	20.23	17.23	2.49	2.67	4.05	4.13	12.14	10.68	3.27	3.4
−15 to 15	20.78	25.24	2.27	3.88	3.87	5.83	12.32	15.53	3.07	4.85
−20 to 20	17.69	32.77	1.79	5.34	3.4	7.28	10.54	20.02	2.59	6.31
−35 to 35	11.76	50.49	1.39	12.38	1.69	13.59	6.72	32.04	1.54	12.98
−45 to 45	11.28	58.5	0.9	18.69	1.18	17.96	6.23	38.23	1.04	18.32

TABLE 12—*Verification error rates (%) for writing styles, without rejection.*

	Both Cursive	Both Print	Mixed
Different-writer Different-content (DD) (412)	23.83	23.68	0.0
Same-writer Different-content (SD) (412)	7.81	6.58	5.0
Overall (DD + SD) (824)	15.82	15.13	2.5

Number of comparisons shown in parentheses.

TABLE 13—*Verification error rates (%) for identical and fraternal twins.*

	Identical	Fraternal	Unsure
Different-writer Different-content (DD) (412)	17.16	11.29	25
Different-writer Same-content (DS) (412)	23.96	11.29	25
Overall (DD + DS) (824)	20.43	11.29	25

Number of comparisons shown in parentheses.

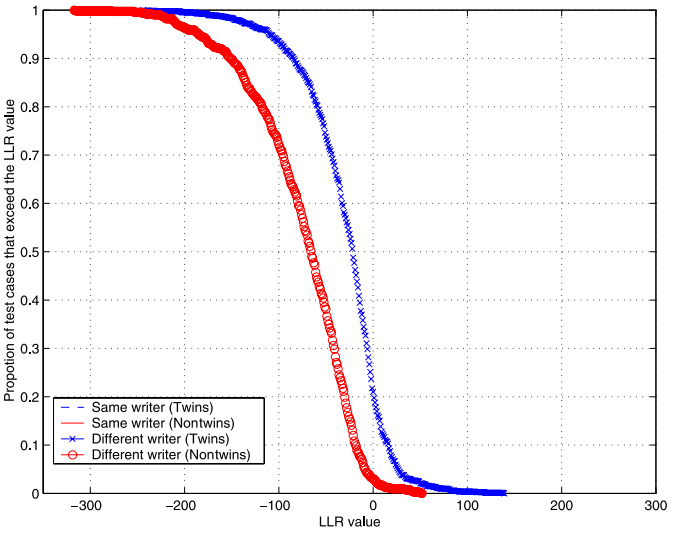

FIG. 16—*Tippet plots for twins and nontwins: the twin plots are less separated than the nontwin plots.*

significant difference as can be discerned from plots of confidence interval curves for error rates in comparing classifiers (18). Thus the handwriting of identical twins is twice as likely to be similar than that of fraternal twins. An earlier study on twin handwriting on a much smaller sample set and human examination concluded that there was no significant difference between identical and fraternal twins (19). Our study indicates that there does exist a significant difference, with the caveats that we used an automatic method and the sample of identical twins considered was small.

Strength of Evidence

It is useful to look at the distribution of LLR values to determine the strength of evidence when we are comparing the handwriting of twins and nontwins. The representation in Fig. 16, called the Tippett plot, was proposed by Evett and Buckleton in the field of interpretation of the forensic DNA analysis (20). The Tippett plot, first used in forensic analysis of paints (21), refers to "within-source comparison" and "between-sources comparison." The functions in the Tippett plot represent the inverse cumulative distribution functions, and are also known as reliability functions.

The Tippett plot of Fig. 16, whose *x*-axis consists of LLR values and *y*-axis the proportion of cases where the LLR values exceeds the x value, has two sets of two curves: one set for nontwins and another for twins. The two curves in each set correspond to same and different writer cases. In interpreting the plot, it has to be kept in mind that an LLR value greater than zero implies a decision of same-writer and a value less than zero implies different-writer. For any given LLR reading, the Tippett plot informs us as to the strength of that evidence, i.e., for a given positive (negative) value, the proportion of the positive (negative) cases that had a higher (lower) value than the current reading.

For the same-writer case, the proportion of cases where the LLR value exceeds zero is about 96% and 95% for nontwins and twins respectively, which follows from the observation that a vertical line at 0 on the *x*-axis intersects the twins and nontwins curves at 0.96 and 0.95. The error rates are the complementary values of 4% for nontwins and 5% for twins. For the different-writer case, the proportion of cases above zero is about 3% for nontwins and 18% for twins, which are the error rates. Thus, it becomes clear that the error rates for twins are higher than that for nontwins.

Another significant observation from the Tippet plot is that the separation between same and different writer curves is more for nontwin data than for twin data, i.e., the positive scores are more positive and the negative scores are more negative for nontwins than for twins. This is another indication that twin handwriting is more similar than nontwin handwriting.

Comparison with Human Performance

It is useful to compare performance of the system to that of the human examiners on the same set of data. As document examination involves cognitive skills, human performance can be considered to be a goal of automation. Such a comparison is also useful

TABLE 14—*Human performance versus System performance (at same rejection rates).*

Examiner	Number of Tests Given	SD Error Rate (%)	DD Error Rate (%)	Overall Error Rate (%)	Reject Rate (%)	Corresponding System Error Rate (%)
A	824	0.49	4.13	2.31	21.0	6.81
B	824	8.01	6.80	7.4	6.8	9.93
C	824	4.61	4.13	4.37	5.95	10.31
D	266	4.97	25.66	15.3	4.5	10.76
E	200	1.53	35.64	18.54	10.1	9.22
F	824	14.03	8.1	11.1	0.0	12.6
G	824	15.05	20.15	17.6	0.0	12.6
H	824	13.35	25.73	19.54	0.0	12.6
I	824	9.23	17.23	13.23	0.0	12.6
J	824	3.16	24.76	13.96	0.0	12.6
K	329	12.88	25.3	19.09	0.0	12.6
L	222	16.51	23.89	20.2	0.0	12.6

to calibrate system performance with respect to human performance.

A test consisting of 824 cases was set up with an online interface. This test is available at http://www.cedar.buffalo.edu/NIJ/test2/verifier2.php.

The system gives a score obtained at the time the user exits from the test. The tests were randomly generated where the two documents were the top and bottom half page images. The user has three options in each test case: same writer, different writer, and reject.

Table 14 gives the performance of twelve humans taking the test. Examiners A, B, and C were interns studying to become QDEs. Examiners D to L were lay persons (graduate students in computer science and engineering). As humans tend to reject borderline cases, the CEDAR-FOX system was also evaluated at the same rejection rates as those of human examiners and the results are shown in the rightmost column.

It is seen that examiners A–C had an average overall error rate of 4.69%, which is better than that of the system error rate of 9.02% at the same rejection rates (error rates of 2.31%, 7.4%, and 4.37% compared to 6.81%, 9.93%, and 10.31%, respectively). The average overall error rate of the nine laypersons (examiners D–L) was 16.51%, which was 4.49% higher than that of the system. Individually only layperson F had a 1.5% lower error rate than the system. Thus, system performance is better than that of lay persons but worse than that of QDEs.

As professional QDEs have been shown to perform better than lay persons (22), we can also conclude that their performance would be superior to that of the current system. One caveat in our results comparing the performance of humans and machines is that the present testing was on a small set of 12 individuals; a larger scale testing would be needed to compare the absolute performances of QDEs, lay persons, and the machine.

Although accuracy is not as high as that of QDEs, one advantage of using the system over human comparison is the speed in performing the tests. While the system takes a few minutes to compare the 812 sample pairs, human comparison can take several hours depending on the examiner. Another aspect of using an automated system is the objectivity of the result in that it will provide the same result with the same input data each time.

Comparison with Previous Study

Previous Study

In the earlier study on handwriting individuality, the size of the combined training and testing set consisted of 1500 writers providing three samples each (6). The handwriting sample consisted of a full page of text and matching was based on the same textual context in questioned and known documents. Writers in the previous study were not twins.

In the previous study, verification was performed by a neural network classifier. This matching was done on same content writing. The verification accuracy was about 96% using macro features based on whole documents and 10 manually cropped characters. When lesser content was used (a paragraph of text), verification performance, using macro features, reduces from 95% to 88%. As character cropping was manual, a perfect match between corresponding characters in the two documents was possible.

Current Study

In the current study, verification is done by the CEDAR-FOX system that uses a naive Bayes classifier, where each of the features was modeled either as Gamma-Gamma or Gamma-Gaussian distribution or using a probability table. The testing was done using both different content data and same content data, and the length of the content became shorter because we only used half page documents. The characters for micro features were obtained by using an automatic transcript-mapping based truthing, which can introduce some errors into the process. For different content testing, the verification correct rates were 87.4% for twins and 96.85% for nontwins. For same content testing, the resulting correct rates were 86.78% for twins and 95.76% for nontwins.

The current study uses a naive Bayes classifier, which is generally not as accurate as a neural network classifier that was used in the previous study. However, it was chosen as the method of comparison in the CEDAR-FOX system because the features can be dynamically selected and each of their contributions to classification can be individually listed, which is not possible for the neural network classifier. Yet, the results of both systems are comparable, as the current study shows an accuracy of about 96% for nontwins, which is about the same as that of the previous study.

Summary and Discussion

The discriminability of the handwriting of twins is a useful measure of the individuality of handwriting. Results of automated handwriting verification using handwriting specimens from twins and nontwins have been presented. When no rejection is allowed, the verification error rate using different content, half page writing is 12.6% for twins and 3.15% for nontwins. By allowing rejection, the verification error rate can be reduced to less than 4% and less than 0.5%, respectively. When comparing identical twins with

fraternal twins with different writer testing cases, the difference of error rates shows that handwriting of identical twins is more similar than that of fraternal twins.

The fact that error rate with twins is higher than with nontwins is probably consistent with biometrics that is based purely on physiological factors such as fingerprints and DNA. Distinguishing between the handwriting of twins is harder than that of nontwins because twins are more likely to have the same genetic and environmental influences than others. The results for nontwins are also consistent with the results of a previous study of the general population. The error rate for nontwins was about the same as that of the previous study, although the textual content of the pair of documents used in verification was different, the textual passages were smaller (half pages), and the characters used in the matching process were automatically determined (rather than manually, thus introducing some errors).

Comparison with human performance, on half-page of writing tests, shows that system performance is better than that of nonexpert humans. From a system design point of view, this is encouraging in that reaching human performance has been the goal of most work in artificial intelligence. With further system improvements, system performance can hope to reach the performance of QDEs. The current system is based on a set of simple features. The use of better features, e.g., those with a cognitive basis such as the ones used by QDEs, and higher accuracy classification algorithms should further decrease the error rates. As expert human performance has been shown to be significantly better than that of lay persons, many sophisticated improvements are likely to be needed to reach the higher goal.

Acknowledgments

We thank our research collaborators Kathleen Storer, Traci Moran, and Rich Dusak as well as the student interns who provided us with the twin data and helped refine the CEDAR-FOX system. We also wish to thank Vivek Shah who contributed to this work. This project was funded in part by Grant Number 2004-IJ-CX-K050 awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice.

References

1. Osborn AS. Questioned documents, 2nd edn. Albany, NY: Boyd Printing Company. Reprinted, Chicago: Nelson-Hall Co., 1929.
2. Robertson EW. Fundamentals of document examination. Chicago: Nelson-Hall, 1991.
3. Bradford RR, Bradford R. Introduction to handwriting examination and identification. Chicago: Nelson-Hall, 1992.
4. Huber RA, Headrick AM. Handwriting identification: facts and fundamentals. Boca Raton, FL: CRC Press, 1999.
5. Hilton O. Scientific examination of questioned documents. Boca Raton, FL: CRC Press, 1993.
6. Srihari SN, Cha SH, Arora H, Lee S. Individuality of handwriting. J Forensic Sci 2002;47(4):856–72.
7. Evans A, Van Baal GC, McCarron P, DeLange M, Soerensen TI, De Geus EJ, et al. The genetics of coronary heart disease: the contribution of twin studies. Twin Res 2003;6(5):432–41.
8. Goldberg S, Perrotta M, Minde K, Corter C. Maternal behavior and attachment in low-birth-weight twins and singletons. Child Dev 1986;57(1):34–46.
9. Jain AK, Prabhakar S, Pankanti S. On the similarity of identical twin fingerprints. Pattern Recog 2002;35(1):2653–63.
10. Rubocki RJ, McCue BJ, Duffy KJ, Shepard KL, Shepherd SJ, Wisecarver JL. Natural DNA mixtures generated in fraternal twins in utero. J Forensic Sci 2001;46(1):120–5.
11. Gamble DJ. The handwriting of identical twins. Can Soc Forensic Sci J 1980;13:11–30.
12. Plamondon R, Lorette G. Automatic signature verification and writer identification—the state of the art. Pattern Recognition 1989;22(2): 107–31.
13. Bulacu M, Schomaker L, Vuurpijl L Writer identification using edge-based directional features. Proceedings of the seventh international conference on document analysis and recognition; 2003 Aug 3–6, Edinburgh, Scotland. Los Alamitos (CA): IEEE Computer Society, 2003;937–41.
14. Van Erp M, Vuurpijl L, Franke K, Schomaker L. The WANDA measurement tool for forensic document examination. J Forensic Doc Exam 2005;16:103–18.
15. Srihari SN, Zhang B, Tomai C, Lee S, Shi Z, Shin YC. A system for handwriting matching and recognition. Proceedings of the symposium on document image understanding technology; 2003 Apr 9–11; Greenbelt (MD). College Park, MD: University of Maryland, Institute for Advanced Computer Studies, 2003;67–75.
16. Huang C, Srihari SN. Mapping transcripts to handwritten text. Proceedings of the tenth international workshop on frontiers in handwriting recognition; 2006 Oct 23–26, La Baule, France. Los Alamitos (CA): IEEE Computer Society, 2006.
17. Zhang B, Srihari SN. Binary vector dissimilarity measures for handwriting identification. Proceedings of SPIE, document recognition and retrieval X; 2003 Jan 20–24, Santa Clara, CA USA; Bellingham (WA): SPIE Press, 2003;5010:28–38.
18. Duda RO, Hart PE, Stork DG. Pattern classification. New York, NY: Wiley, 2001.
19. Boot D. An investigation into the degree of similarity in the handwriting of identical and fraternal twins in New Zealand. J Am Soc Quest Doc Exam 1998;1(2):70–81.
20. Evett IW, Buckleton JS. Statistical analysis of STR data. In: Carracedo A, Brinkmann B, Mayr W, editors. Advances in forensic haemogenetics 6. Heidelberg, Berlin: Springer, 1996;79–86.
21. Tippett CF, Emerson VJ, Lawton F, Lampert SM. The evidential value of the comparison of paint flakes from sources other than vehicles. J Forensic Sci Soc 1968;8:61–5.
22. Kam M, Fielding G, Conn R. Writer identification by professional document examiners. J Forensic Sci 1997;42(5):778–85.

Additional information and reprint requests:
Sargur Srihari, Ph.D.
SUNY Distinguished Professor
Department of Computer Science and Engineering and Director
Center of Excellence for Document Analysis and Recognition
University at Buffalo
State University of New York
Buffalo, NY 14228
E-mail: Srihari@cedar.buffalo.edu