# On the Discriminability of the Handwriting of Twins

Sargur Srihari, Chen Huang and Harish Srinivasan

TR-04-07 June 2007



Center of Excellence for Document Analysis and Recognition (CEDAR)
520 Lee Entrance, Suite 202
Amherst. New York 14228

To appear in the Journal of Forensic Sciences

# On the Discriminability of the Handwriting of Twins

Sargur Srihari, Chen Huang, Harish Srinivasan

This project was funded in part by Grant Number 2004-IJ-CX-K050 awarded by the National Institute of Justice, Office of Justice Programs, US 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 US Department of

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.

Doctoral Candidate, Department of Computer Science and Engineering, University at Buffalo, State University of New York, Buffalo, NY 14228.

#### Abstract

Since 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. The 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 1,648 test cases were obtained from handwriting samples of non-twins (general population). In order 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 non-twins 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 non-twins: an overall error rate of 12.91% for twins and 3.7% for non-twins. 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).

#### **Index Terms**

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

## I. Introduction

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], [2], [3], [4], [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 (QD) examiners. In a previous study of handwriting individuality [6] we reported on the discriminability of handwriting of a diverse population from across the Unites 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 [12], [13], [14], [6]. 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 (i) comparing performance on handwriting of twins with those of the general population; (ii) determining performance when the textual content of the questioned and known writing is different; (iii) comparing performance on fraternal and identical twins; and (iv) comparing system performance with that of humans. The evaluation was done 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 those that were reported previously with samples not specialized to twins. The last section contains concluding remarks.

#### II. 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: (i) writing element extraction, (ii) similarity computation, (iii) 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: (i) writing element extraction, (ii) similarity computation, (iii) determining the log-likelihood ratio (LLR) from the estimated conditional probability density estimates.

## A. 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 same or different writers.

*Macro-features* capture the global characteristics of writer's individual writing habit and style. They are extracted from the entire document. Totally there are thirteen 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 eleven 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 degree or flat stroke), positive (45 or 225 degree), vertical (90 or 270 degree) and negative (135 or 315 degree), average height and average slant per line. In our current system, 11 out 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]. Fig.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).

# [Figure 1 here]

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: (i) manually crop characters and label each with its truth; this labor-intensive method has the highest accuracy, (ii) automatically recognize characters by using a built-in character recognizer; the method is error prone for cursive handwriting where there are few isolated characters, (iii) 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 (iv) 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. Since the full-page documents have the same content (CEDAR-letter), the transcript mapping approach was used as shown in Fig.2 This method has an 85% accuracy among words recognized. Since most words are recognized, they are also useful for extracting features of letter pairs and whole words, as discussed next.

[Figure 2 here]

Style features are features based on the shapes of whole words and shapes of letter pairs, known as bigrams. They are similar to micro features of characters as described below. Fig.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 bigram is a partial word that contains two connected characters, such as th,ed and so on, as shown in Fig.4 (a) to (c), where (a) and (b) were written by the same writer while (c) was from a different writer.

[Figure 3 here]

[Figure 4 here]

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. In order 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, since 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 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.

## B. Statistical Model of Similarity

The distributions of dissimilarities 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 data set 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, the Gamma density is better for modeling distances since distance is never negative-valued.

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  $p_s(x) \sim p(a_{same}, b_{same})$  and  $p_d(x) \sim p(a_{diff}, b_{diff})$ . For macro features, we model both categories by Gamma distribution as  $p_s(x) \sim Gam(\alpha_s, \beta_s)$  and  $p_d(x) \sim Gam(\alpha_d, \beta_d)$ . For micro and style features, while the "same-writer" category is modeled as  $p_s(x) \sim Gam(\alpha_s, \beta_s)$  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 Fig.5 (a) and (b) respectively. Conditional parametric pdfs for "slant" and for "e" are shown in Fig.5 (c) and (d).

[Figure 5 here]

The Gaussian and Gamma probability density functions are as follows.

$$Gaussian(x) = \frac{1}{(2\pi)^{1/2}\sigma} \exp^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \qquad for \ x \in R$$
$$Gamma(x) = \frac{x^{\alpha-1} \exp^{(-x/\beta)}}{(\Gamma(\alpha))\beta^{\alpha}} \qquad for \ x > 0$$

Here  $\mu$  and  $\sigma$  are the mean and the standard deviation of the Gaussian distribution, while  $\alpha>0$  and  $\beta>0$  are gamma parameters which can be evaluated from the mean M and variance V as follows  $\alpha=M^2/V$  and  $\beta=V/M$ . ' $\alpha$ ' is called the shape parameter and ' $\beta$ ' 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.

[Table 1 here]

#### C. Parameter Estimation

All statistical parameters for each of the features used by CEDAR-FOX were estimated by using maximum likelihood estimation. The training data is a set of learning, or design, samples provided by 1,000 non-twin 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 1,000 writers is a subset of samples provided by 1,568 non-twin writers. The remainder of the samples was kept aside for testing as non-twins 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 3,000 pairs of half-page documents including 1,500 from same writer and 1,500 from different writer.

[Table 2 here]

## D. 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 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.

When each document is characterized by more than one feature CEDAR-FOX makes the assumption that the writing elements or features are statistically independent, although strictly speaking this is incorrect. Fig.6 shows a scatter plot for the first 12 macro features, which shows that there are some correlations between the features. However such a classification method, known as naive Bayes classification, has yielded good results in machine learning. We also investigated more complex models, such as a Bayesian neural network, and got an improvement of 1-2% on overall accuracy, which is not significant. There are several other justifications for choosing nave Bayes. First, its functioning is simple to explain and modify, e.g., since log-likelihood ratios are additive it allows us to observe the effects of adding and removing features easily. Its simplicity goes back to the earliest QDE literature [1] where there is reference to multiplying the probabilities of handwriting elements. Finally, a simpler model philosophically relates to Occam's razor and practically to the avoidance of over-fitting.

[Figure 6 here]

Each of the two likelihoods that the given pair of documents were either written by the same individual or by different individuals can be expressed, assuming statistical independence of the features as follows. For each writing element  $e_i$ ,  $i=1,\cdots,c$ , 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 likelihoods are extracted as

$$L_s = \prod_{i=1}^c \prod_j \prod_k p_s(d_i(j,k))$$
$$L_d = \prod_{i=1}^c \prod_j \prod_k p_d(d_i(j,k))$$

The log-likelihood ratio (LLR) in this case has the form

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

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 likelihood ratio 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.

#### E. 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. Fig.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. 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 non-zero 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.

[Figure 7 here]

#### III. 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 two hundred and six pairs of twins and non-twins 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.

## A. Demographic distribution of twins

The demographic data for the individuals who contributed to the twin's data set is as follows.

- Location Samples were obtained from people coming from at least 150 different cities and 7 different countries.
- Age The age distribution of the sample pairs is as shown in Fig.8. Here age is divided into regions with the first region 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.

[Figure 8 here]

- School All the twins attended the same school as their twin pair. 16% had private schooling while the rest 84% had public schooling.
- Handedness Of the 412 test cases, 15 (3.6%) were ambidextrous, 45 (10.9%) were left handed and the rest 352 (85.5%) were right handed. Among twin 36 (17.5%) pairs had different handedness.
- Type of twins Out of the 206 pairs of twins 31 (15.1%) were fraternal, 170 (82.5%) were identical and the rest of the 5 (2.4%) pairs were unsure.
- Sex 69 (16.7%) test cases were male and the rest 343 (83.3%) female.
- Injuries 21 (5.1%) had injuries that could affect their handwriting.

## B. 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.

[Table 3 here]

## C. 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 ball-point pen. Portions of the CEDAR letter for two pairs of twins having similar and dissimilar handwriting are shown in Fig.9 and Fig.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 Table 4 and Table 5.

```
[Figure 9 here]

[Table 4 here]

[Figure 10 here]

[Table 5 here]
```

## D. Verification test cases

Two hundred and six pairs of twins provided one sample each. Thus a total of  $412 (206 \times 2)$  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 11 here]

[Table 6 here]

Fig.12 shows the schematics of the method by which the test cases were generated.

[Figure 12 here]

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) Same writer, Different content (SD): The two samples are from the same writer having different textual content.
- 3) Different writer, Same content (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 since there was only one sample provided by each twin writer. For non-twins, since we have three samples from each writer, we generated another set of 412 half-page documents from the original data set to form the SS test cases, as shown in Fig.12(b). The number of test cases for each of the four scenarios are given in Table 7.

[Table 7 here]

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

## E. 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 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.

## IV. VERIFICATION RESULTS

The outcome of a writer verification experiment can be one four possibilities: (a) *true positive*: documents written by the same writers and it is confirmed by the results; (b) *true negative*: documents written by different writers and it is confirmed by the results; (c) *false positive*: documents written by different writers but the results concluding them to be written by the same writer and (d) *false negative*: documents written by the same writers but the results concluding that they are written by the 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.

#### A. 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 Same writer Different content data with majority of them lying in the  $0 \sim 50$  range (as shown in Fig.13). Test cases having LLR values less than 0 are false negatives. For Different writer Different content data LLR values range from -112.62 to 54.36 with most of them lying in  $-70 \sim 10$  range.

# [Figure 13 here]

Different Writer (DS + DD)-For Different writer Same content (DS), the LLR values range from -243.24 to 138.4. Majority of the cases lie in the  $-75 \sim 20$  range. As seen from Fig.14, different writer same content data has more false positives than different writer different content. This can be attributed to the pictorial effects where different content appears as different pictorial content.

[Figure 14 here]

## B. Verification error rates with twins and non-twins

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 non-twins is 3.15%.

## [Table 8 here]

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

# [Table 9 here]

The overall error rates for twins versus non-twins can be observed by combining the results of same content (Table 8) and different content (Table 9 IX). 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 non-twins 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 non-twins.

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. Since 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, in order to having a same rejection rate (for example 20%), one needs to choose a bigger interval for same content cases ( $-20 \sim 20$  in this case) and a smaller interval for different content cases ( $-10 \sim 10$  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 non-twins.

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.

[Table 10 here]

Fig.15 shows the error rates with the increasing reject rates for twins as well as non-twins. Error rate for twins can be reduced from 12.6% to 3.73% by rejecting 40% of the cases. Similarly the error rate for non-twins can be reduced from 3.15% to 0.14% by rejecting about 25% of the test cases.

[Figure 15 here]

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 non-twins 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.

[Table 11 here]

#### C. 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 different writer different content testing, the error rate is zero for mixed style since 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 QD examiners in practice. An automatic reject option can be put into CEDAR-FOX to reject cases where there are mixed types of writing styles, particularly since it is already able to make an assignment on the print-cursive scale.

[Table 12 here]

#### D. 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 6 unsure). Verification error rates for each test category are shown in Table 13.

# [Table 13 here]

The average error rate for identical twins is 20.43% and the average for fraternal twins is 11.29%. 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 [18]. Our study indicates that there does exist a difference, with the caveats that we used an automatic method and the sample of identical twins considered was small (31).

## E. 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 non-twins. 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 [19]. The Tippett plot, first used in forensic analysis of paints [20]. refers to "within-source comparison" and "between-sources comparison". The functions in the Tippett plot represent the inverse CDFs (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 non-twins 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 non-twins and twins respectively-which follows from the observation that a vertical line at 0 on the x-axis intersects the twins and non-twins curves at 0.96 and 0.95. The error rates are the complementary values of 4% for non-twins and 5% for twins. For the different-writer case, the proportion of cases above zero is about 3% for non-twins and 18% for twins-which are the error rates. Thus it becomes clear that the error rates for twins are higher than that for non-twins.

## [Figure 16 here]

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

## V. 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. Since document examination involves cognitive skills, human performance can be considered to be a goal of automation. Such a comparison is also useful to calibrate system performance with respect to human performance.

A test consisting of 824 cases was setup 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 QD examiners. Examiners D to L were lay persons (graduate students in computer

science and engineering). Since 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.

## [Table 14 here]

It is seen that examiners A-C had an average of 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 of 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.

Since professional QDEs have been shown to perform better than lay persons [21], 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.

#### VI. COMPARISON WITH PREVIOUS STUDY

#### A. 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%. Since character cropping was manual a perfect match between corresponding characters in the two documents was possible.

# B. Current study

In the current study verification is done by the CEDAR-FOX system which uses a naive Bayes classifier, where each of the features was modeled either as 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 since we only use 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 non-twins. For same content testing, the resulting correct rates were 86.78% for twins and 95.76% for non-twins.

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 non-twins, which is about the same as that of the previous study.

#### VII. 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 non-twins 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 non-twins. 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 are more similar than that of fraternal twins.

The fact that error rate with twins is higher than with non-twins 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 non-twins since twins are more likely to have the same genetic and environmental influences than others. The results for non-twins are also consistent with the results of a previous study of the general population. The error rate for non-twins 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), 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 non-expert 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 questioned document examiners, and higher accuracy classification algorithms should further decrease the error rates. Since expert human performance has been shown to be significantly better that that of lay persons many sophisticated improvements are likely to be needed to reach the higher goal.

## VIII. ACKNOWLEDGEMENT

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.

#### REFERENCES

- [1] A. S. Osborn, Questioned Documents. 2nd ED, Albany, NY: Boyd Printing Company. Reprinted, Chicogo: Nelson Hall Co., 1929.
- [2] E. W. Robertson, Fundamentals of Document Examination, Chicago: Nelson-Hall, 1991.
- [3] R. R. Bradford and R. Bradford, Introduction to Handwriting Examination and Identification, Chicago: Nelson-Hall, 1992.
- [4] R. Huber and A. Headrick, Handwriting Identification: Facts and Fundamentals, Boca Raton, FL: CRC Press, 1999.
- [5] O. Hilton, Scientific examination of questioned documents, Boca Raton, FL: CRC Press, 1993.
- [6] S. N. Srihari, S.-H. Cha, H. Arora, and S. Lee, "Individuality of handwriting," in Journal of Forensic Sciences, 47(4), pp. 856–872, July 2002.
- [7] A. Evans, G. C. V. Baal, and e. a. P. McCarron, "The genetics of coronary heart disease: the contribution of twin studies," in *Twin Research*, 6(5), pp. 432-441, Oct. 2003.
- [8] S. Goldberg, M. Perrotta, K. Minde, and C. Corter, "Maternal behavior and attachment in low-birth-weight twins and singletons," in *Child Development*, **57(1)**, pp. 34–46, Feb. 1986.
- [9] A. K. Jain, S. Prabhakar, and S. Pankanti, "On the similarity of identical twin fingerprints," in Pattern Recognition, 35(1), pp. 2653-2663, 2002.
- [10] R. J. Rubocki, B. J. McCue, K. J. Duffy, K. L. Shepard, S. J. Shepherd, and J. L. Wisecarver, "Natural dna mixtures generated in fraternal twins in utero," in *Journal of Forensic Science*, **46(1)**, pp. 120–125, Jan. 2001.
- [11] D. J. Gamble, "The handwriting of identical twins.," in Canadian Society of Forensic Science Journal, 13, pp. 11-30, 1980.
- [12] R. Plamondon and G. Lorette, "Automatic signature verification and writer identification state of the art.," in *Pattern Recognition*, **22(2)**, pp. 107–131, 1989.
- [13] M. Balacu, L. Schumaker, and L. Vuurpijl, "Writer identification using edge-based directional features," pp. 937–941, Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR), Aug. 2003.
- [14] M. Van-Erp, L. G. Vuurpijl, K. Franke, and L. R. Schumaker, "The wanda measurement tool for forensic document examination," in *Journal of Forensic Document Examination*, 16, pp. 103–118, 2005.
- [15] S. N. Srihari, B. Zhang, C. Tomai, S. Lee, Z. Shi, and Y. C. Shin, "A system for handwriting matching and recognition," in *Proceedings of the Symposium on Document Image Understanding Technology (SDIUT 03), Greenbelt, MD*, pp. 67–75, 2003.
- [16] C. Huang and S. Srihari, "Mapping transcripts to handwritten text," in *Proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition*, Oct. 2006.
- [17] B. Zhang and S. N. Srihari, "Binary vector dissimilarity measures for handwriting identification," in *Document Recognition and Retrieval X, SPIE, Bellingham, WA*, **5010**, pp. 28–38, 2003.
- [18] D. Boot, "An investigation into the degree of similarity in the handwriting of identical and fraternal twins in new zealand," in *Journal of American Society of Questioned Document Examiners (ASQDE)*, **1(2)**, pp. 70–81, 1998.
- [19] I. W. Evett and J. S. Buckleton, "Statistical analysis of str data," in *Advances in Forensic Haemogenetics* 6, pp. 79–86, Heidelberg, Berlin: Springer, 1996.
- [20] C. F. Tippett, V. J. Emerson, F. Lawton, and S. M. Lampert, "The evidential value of the comparison of paint flakes from sources other than vehicules," in *Journal of Forensic Science Society*, **8**, pp. 61–65, 1968.
- [21] M. Kam, G. Fielding, and R. Conn, "Writer identification by professional document examiners," in *Journal of Forensic Science*, **42(5)**, pp. 778–785, 1997.

#### IX. FIGURE LEGENDS

- FIG.1–Micro features for three handwritten letters. Each column gives the greyscale image, binary image, GSC binary feature vecter (512 bits), and color coded gradient, structural and concavity vectors. 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-Bigram shape: (a) and (b) are two bigram samples from the same writer while (c) is a bigram 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.
- FIG.5-Histograms and parametric probability density functions: (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 pdfs for "slant" and (d) same and different pdfs for letter 'e'.
- FIG.6-Scatter plot for 12 macro features.
- 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.
- 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.
- 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.
- 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.
- FIG.12–Generating test cases for (a) twin samples and (b) non-twin samples: each bidirectional arrow represents 206 test cases.
- 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).
- FIG.15-Verification Error-Reject rates for twins and non-twins using different content data.
- FIG.16-Tippet plots for twins and non-twins: the twin plots are less separated than the non-twin plots.

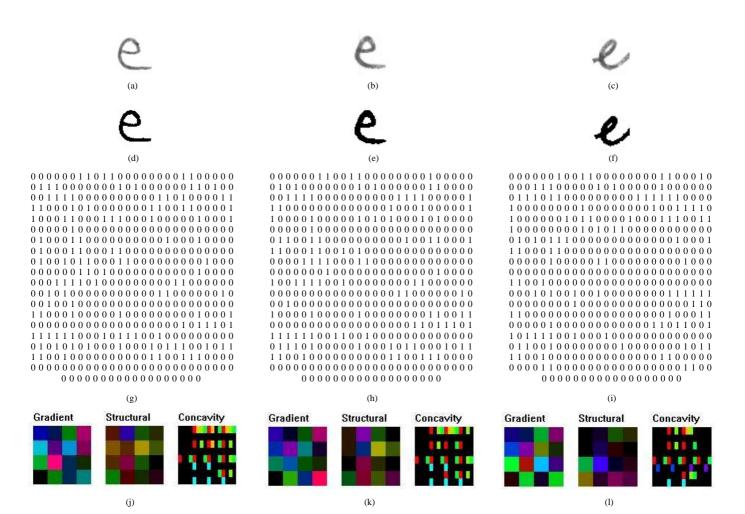


Fig. 1.

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

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

(b)

Twee were referred to you by Xxena Chonena at the University ity Medical Cecenterer. Ithis is the regarding may friend, Kate Zack.

Fig. 2.



Fig. 3.

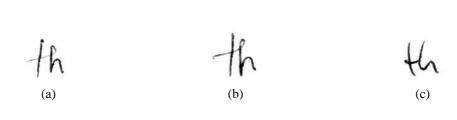


Fig. 4.

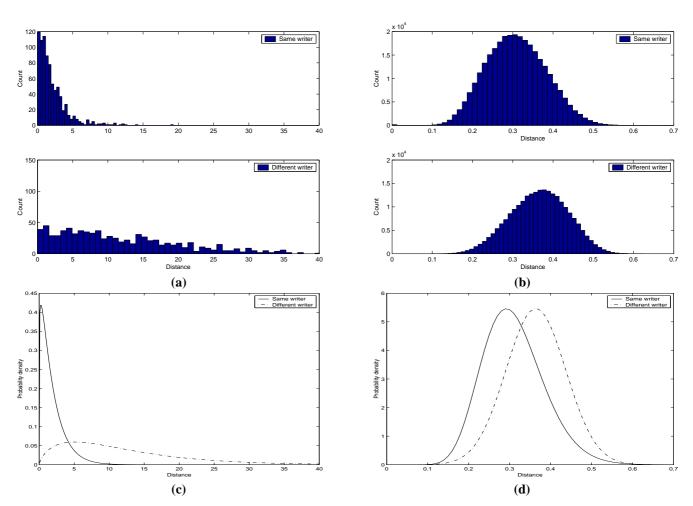


Fig. 5.

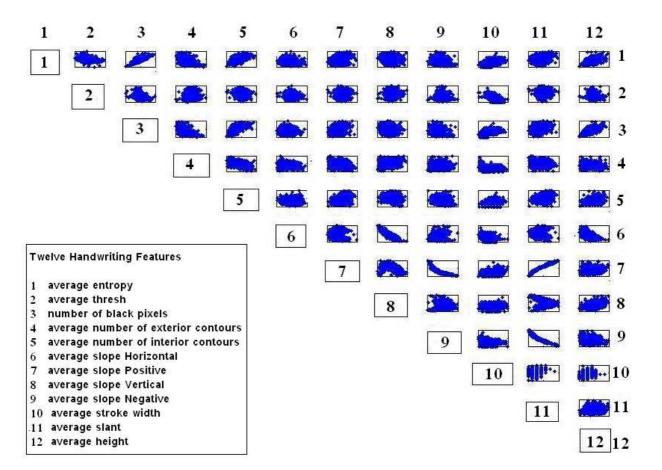
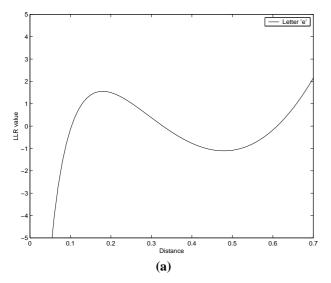


Fig. 6.



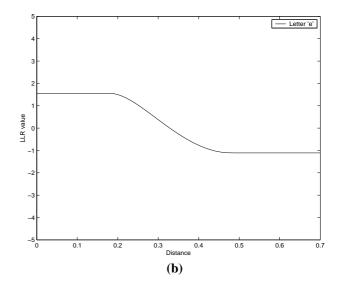


Fig. 7.

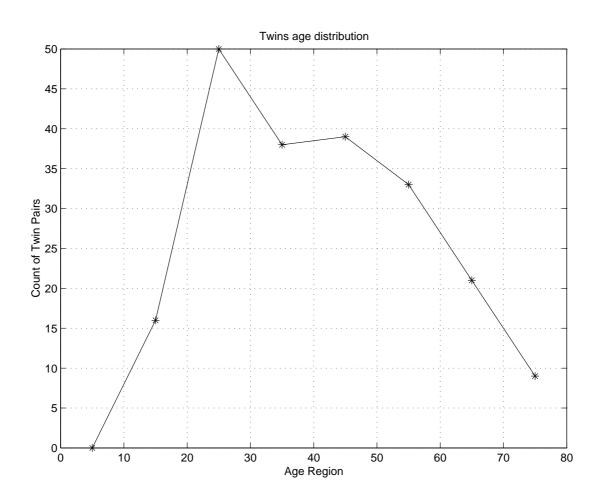


Fig. 8.

Jun Elder
Jun Juniphe 19101

Der Bothant
Juniphen Juniphe
Juniphen
Juniphe

Jaw. 10, 1999

Tim Elder

1829 oop Steek Joh 300

Allentown New York 14707

To be Dob Jonat

632 Jonat

632 Jonat

632 Jonat

182 Wese refreshing Softwar John at the

Unwersett fleshed Gentes This is reparsing

my John State Jack.

And stending the Ruley Jagy mean Injurying

such as inform to be sheare, and as

Justice of the State was averaged by

site when the way averaged but the shear was a server of the

steen with great geal and enthusiasm.

Fig. 9.

Jum Elder No. 10, 1999 San Loop Stood, Apd 300 Ollerton, New York 1400

Dr. Gob Gleant 603 Quandroy Salvery Omar, West Vergenia 25638

We were referred to you by Kenon Cohen at the University Medical Codes. This is respecting my Juend, Koto Marts.

enthalls about around six months ago white attending of the last grant of Land " and " and " and a continuation of the all months on the since on

From Sin Edder Street apt 300

Callectown, Down fork 11707

To Dr. Pob Thank 600 11707

To One were referred achieves Onea, west Virginia 25038

We were referred center. This is regarding my breing, Kate Jack.

all all beauted around hit months ago while attending on a Valley" Boy Covered Organizing hurl an event is no prince and as Previous of the Edunaic Association a construction of the event of the superior of the event of the input of the superior of the event of the input of the input of guest gal and estilessing.

(a)

November 10, 1999

From
Jim Elder
829 Loop Street, Apt 300
Allentown, NY 14707

To Dr. Bob Grant 602 Queensberry Parkway Omar, WV 2638

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

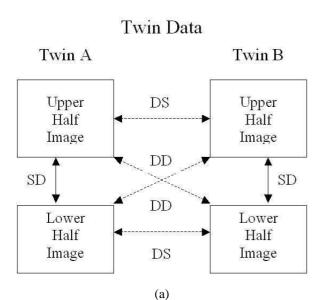
It all started around SIX months ago while attending the "Rubeg" Jazz Concert. Organizing Such an event is no picnic, and as President of the Alumn:

Fig. 11.

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

Kate's been in very bad Lealth Since. Could you kindly take a look at the results and give us your opinion?

Thank you!



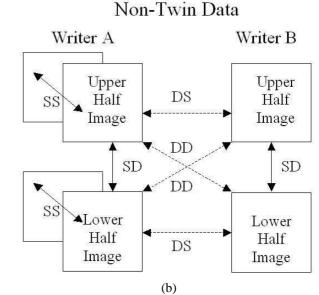


Fig. 12.

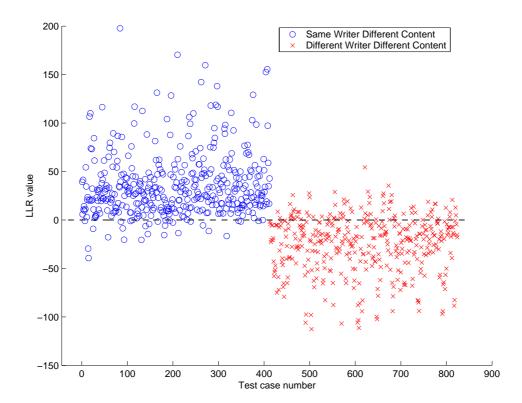


Fig. 13.

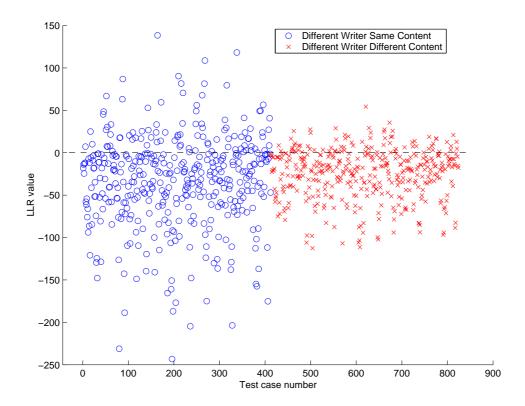


Fig. 14.

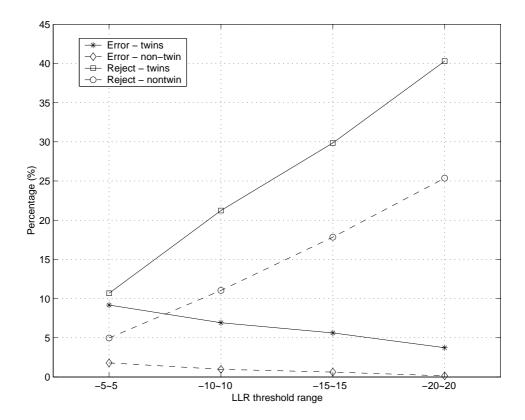


Fig. 15.

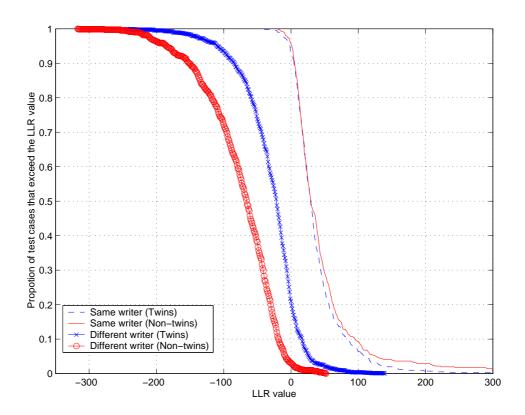


Fig. 16.

 $\label{eq:table in table in table in table in table in the features} \\ \text{Modeling distributions for all the features}$ 

Feature	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
Avg. slant	Continuous	Gamma
Avg. height	Discrete	Probability table
Avg. word gap	Continuous	Gamma
Micro features	Continuous	Gamma-Gaussian
Bi-gram	Continuous	Gamma-Gaussian
Word	Continuous	Gamma-Gaussian

 $\begin{tabular}{ll} TABLE \ II \\ GAMMA \ PARAMETERS \ FOR \ CONTINUOUS \ MACRO \ FEATURES \\ \end{tabular}$ 

Feature	Shape pa	rameter - $\alpha$	Scale pa	rameter - $eta$	
	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	
Avg. slant	1.1471	1.7135	1.6609	7.0830	
Avg. word gap	3.4487	2.9363	0.0128	0.0372	

 $\label{thm:table} \mbox{TABLE III}$  Distribution of writing styles among twins

Bo	Both printing		Used two different	Total
Normal	All upper case		styles (mixed)	
36	2	128	40	206

 $\label{thm:table_in_table_in_table} TABLE\ IV$  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
Avg. slant	N/A	2.45	1.11
Avg. height	N/A	3	0.63
Avg. word gap	N/A	0.03	1.22
Micro	10	N/A	2.68
Bi-gram	0	N/A	0
Word	9	N/A	6.17
Total	N/A	N/A	7.15

 $\label{thm:table v} TABLE\ V$  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
Avg. slant	N/A	22.41	-9.35
Avg. height	N/A	10	-2.24
Avg. word gap	N/A	0.07	-0.14
Micro	256	N/A	-22.07
Bi-gram	0	N/A	0
Word	38	N/A	-43.54
Total	N/A	N/A	-98.36

 $\label{thm:table VI} The \ similarity \ table \ for \ two \ document \ samples: Twin 178a-U \ and \ Twin 178b-U$ 

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
Avg. slant	N/A	1.91	1.5
Avg. height	N/A	1	1.03
Avg. word gap	N/A	0.05	0.65
Micro	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

 $\label{thm:constraint} TABLE\ VII$  Distribution of test cases in four verification scenarios

Verification scenarios	Number of half-page document pairs			
	Twins	Non-Twins		
Same Writer-Different Content (SD)	412	412		
Different Writer-Different Content (DD)	412	412		
Different Writer-Same Content (DS)	412	412		
Same Writer-Same Content (SS)	0	412		

 $TABLE\ VIII$   $Verification\ error\ rates\ for\ twins\ and\ non-twins\ using\ different\ content\ data\ with\ no\ rejection$ 

Features under	Differe	ent-writer	Sam	e-writer	Overall		
consideration	Different-content (DD)		Different-content (SD)		(DD+SD)		
	Twins Non-Twins		Twins	Non-Twins	Twins	Non-Twins	
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-Distance (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 IX

VERIFICATION ERROR RATES FOR TWINS AND NON-TWINS USING SAME CONTENT DATA WITH NO REJECTION

Features under	Different-writer		Same-writer	Overall	
consideration	Same-co	ntent (DS)	Same-content (SS)	(DS	S+SS)
	Twins Non-Twins		Non-Twins	Twins	Non-Twins
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-Distance (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 %

 $\label{table X} \mbox{Verification error rates for twins and non-twins using different content data with rejection}$ 

LLR	Diff-writer Diff-content (DD)			Same-writer Diff-content (SD)			Overall (DD+SD)							
Rejection	jection Twins		Non-twins		Twins Non-twins		Tv	vins	Non-	-twins	Tv	vins	Non-	-twins
Range	Error	Reject	Error	Reject	Error	Reject	Error	Reject	Error	Reject	Error	Reject		
-5 - 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 - 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 – 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 - 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		

LLR	Diff-writer Same-content (DS)			nt (DS)	Same-	writer Same-content (SS)	Overall (DS+SS)			
Rejection	Tw	ins	Non-	-twins		Non-twins	Tw	ins	Non-	-twins
Range	Error	Reject	Error	Reject	Error	Reject	Error	Reject	Error	Reject
-5 - 5	22.04	9.71	2.72	1.7	4.44	1.7	13.14	5.70	3.58	1.7
-10 - 10	20.23	17.23	2.49	2.67	4.05	4.13	12.14	10.68	3.27	3.4
-15 – 15	20.78	25.24	2.27	3.88	3.87	5.83	12.32	15.53	3.07	4.85
-20 – 20	17.69	32.77	1.79	5.34	3.4	7.28	10.54	20.02	2.59	6.313
-35 – 35	11.76	50.49	1.39	12.38	1.69	13.59	6.72	32.04	1.54	12.98
-45 – 45	11.28	58.5	0.9	18.69	1.18	17.96	6.23	38.23	1.04	18.32

 $\label{thm:table XII} \mbox{Verification error rates for writing styles, without rejection}$ 

	Both Cursive	Both Print	Mixed
Different Writer Different Content (DD)	23.83 %	23.68 %	0.0 %
Same Writer Different Content (SD)	7.81 %	6.58 %	5.0 %
Overall (DD+SD)	15.82 %	15.13 %	2.5 %

 $\label{thm:table XIII} \mbox{Verification error rates for identical and fraternal twins}$ 

	Identical	Fraternal	Unsure
Different writer Different content (DD)	17.16 %	11.29 %	25.0 %
Different writer Same content (DS)	23.96 %	11.29 %	25.0 %
Overall (DD+SD)	20.43 %	11.29 %	25.0 %

 $\label{thm:table XIV} \\ \mbox{Human performance vs System performance (at same rejection rates)}$ 

	Number of	Same writer	Different writer	Average	Reject	Corresponding
Examiner	Tests Given	Different content (SD)	Different content (DD)	Error Rate	Rate	System
		Error Rate (%)	Error Rate (%)	(%)	(%)	Error Rate (%)
A	824	0.49	4.13	2.31	21.0	6.81
В	824	8.01	6.80	7.40	6.80	9.93
С	824	4.61	4.13	4.37	5.95	10.31
D	266	4.97	25.66	15.30	4.50	10.76
Е	200	1.53	35.64	18.54	10.1	9.22
F	824	14.03	8.10	11.10	0.00	12.6
G	824	15.05	20.15	17.60	0.00	12.6
Н	824	13.35	25.73	19.54	0.00	12.6
I	824	9.23	17.23	13.23	0.00	12.6
J	824	3.16	24.76	13.96	0.00	12.6
K	329	12.88	25.30	19.09	0.00	12.6
L	222	16.51	23.89	20.20	0.00	12.6