

# Multi-Language Handwriting Derived Biometric Identification

**Donald T. Gantz, PhD John J. Miller, PhD**

George Mason University  
Fairfax, Virginia

**Mark A. Walch**

The Gannon Technologies Group  
Alexandria, Virginia

## Abstract

*This paper provides a discussion of key technologies underlying the development and implementation of a Handwriting Derived Biometric Identification system. Three applications for author identification through handwriting are in development based on this system:*

1. *Individual Identification*
2. *Document Clustering*
3. *Forensic Document Examination*

*Individual Identification involves relating a document of unknown authorship to a data base of reference samples of handwriting from known authors.*

*Document Clustering encompasses grouping documents based on handwriting characteristics without knowledge of specific author identity.*


*Forensic Document Examination involves building a statistical foundation that will support court room testimony by expert witnesses that will withstand Daubert Challenges raised against handwriting analysis as evidence.*

*This paper focuses on the technical foundations of Individual Identification.*

*Handwriting Derived Biometric Identification exploits the rich set of measurements available through Isomorphic Graph Matching which is a technique based on Graph-Theory that is used to identify the same written forms in different writing samples. By statistically comparing measurements on similar objects across different writing we are able to identify those writing characteristics that best distinguish or characterize individual authors. An author's biometric identity is defined through the measurements that are determined to characterize that author's writing in the sense that those measurements have the power to distinguish the author's writing from that of other authors. Handwriting Derived Biometric Identification is a computationally intense process that utilizes statistical discrimination algorithms.*

## 1 Overview

An automated process parses a handwriting sample to identify individual characters. This process uniquely associates each parsed character with a graph. For purposes of this paper, the software tool that makes this association is referenced as the "Graph Builder". Each graph is a collection of nodes connected by loops and curves. Nodes are located at the ends of curves or where curves cross. The following example demonstrates that there is a very large set of physical measurements defined for even the simplest graphs.

<b>Character:</b>	<b>a</b>	
<b>Graph Model:</b>	<b>4;112</b>	
<b>Number of Edges:</b>	<b>3</b>	
<b>Number of Vertices:</b>	<b>4</b>	

Measurements for Graph 4:112	Number of Measurements
Absolute Distance	
Vertex to Vertex	6
Vertex to Edge Centroid	12
Vertex to Edge Contours	12
Edge Centroid to Edge Centroid	3
Edge Contours to Edge Contours	3
	36
Centroid Distance	
Vertex to Graph Centroid	4
Edge Centroid to Graph Centroid	3
Edge Contours to Graph Centroid	3
	10
Graph Direction	
Vertex to Vertex	18
Vertex to Edge Centroid	24
Vertex to Edge Contours	24
Edge Centroid to Edge Centroid	6
Edge Contours to Edge Contours	6
	78
Centroid Direction	
Vertex to Graph Centroid	12
Edge Centroid to Graph Centroid	9
Edge Contours to Graph Centroid	9
	30



Each row of the table refers to a single letter; and each column of the table refers to a particular graph. The numbers in each row of the table present the percentage of occurrences of the letter that were

assigned to the column graph. Entries in the table are constrained in that percentages are only calculated for those letter/graph pairs which were observed at least ten times in the author's modeling London letters.

Table 2: The same information as in Table 1, but for Author=10.

	Author=10																		
C																			
h																			
a																			
r																			
a	2	1	-	-	3	1	-	-	-	4	4	4	1	1	-	1	4	-	-
c	-	1	6	2	-	2	0	6	9	-	-	-	-	1	6	2	-	0	6
t	1	2	4	-	2	0	-	4	6	0	0	1	1	2	6	0	0	-	4
e	9	-	-	6	2	-	1	-	-	-	-	2	2	-	-	-	-	3	-
r	2	0	0	4	4	0	6	0	0	0	0	8	8	0	0	0	2	0	8
a	19.2	--	48.1	--	--	--	--	--	--	--	--	--	--	32.7	--	--	--	--	--
c	--	100.0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
d	--	--	--	--	--	--	--	--	--	--	--	--	--	100.0	--	--	--	--	--
e	--	42.0	--	7.0	--	28.7	--	--	9.8	--	--	--	--	--	--	12.6	--	--	--
h	--	100.0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
i	--	33.8	--	--	--	66.2	--	--	--	--	--	--	--	--	--	--	--	--	--
l	--	100.0	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
n	--	75.8	--	--	--	24.2	--	--	--	--	--	--	--	--	--	--	--	--	--
o	42.9	--	57.1	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
r	--	29.7	--	56.0	--	--	--	--	--	--	14.3	--	--	--	--	--	--	--	--
s	--	56.7	43.3	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
t	--	--	--	--	--	--	--	25.4	--	55.2	--	--	--	--	--	--	19.4	--	--
u	--	--	--	--	--	--	--	100.0	--	--	--	--	--	--	--	--	--	--	--
w	--	--	--	--	--	--	--	--	--	100.0	--	--	--	--	--	--	--	--	--

Table 3: Observed frequencies of letter/isomorphism pairs occurring in Author #1's test paragraphs.

	Author=1																		
C																			
h																			
a																			
r																			
a	1	-	2	3	-	1	1	4	-	1	5	-	1	2	1	1	-	4	4
c	-	2	-	-	6	1	2	-	6	6	2	-	0	1	9	6	-	-	0
t	1	-	1	2	4	2	0	1	4	4	0	1	-	2	6	6	0	0	-
e	2	6	9	2	-	-	-	2	-	-	-	2	1	-	-	-	-	-	3
r	8	4	2	4	0	0	0	8	0	0	0	8	6	0	0	0	0	2	0
a	0	0	28	0	0	13	0	0	10	0	0	0	0	0	0	0	0	0	0
c	0	14	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0
d	0	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0	0
e	0	0	23	58	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0
i	0	0	0	0	0	0	0	0	0	0	0	0	50	0	0	0	0	0	0
l	0	0	0	13	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0
n	0	32	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0
o	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	50	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0
s	0	11	13	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0
t	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	25	0	5
u	0	0	0	0	0	8	0	0	0	0	0	0	0	16	0	0	0	0	0
w	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0



This strong discrimination was typical regardless of the author playing the role of the Unknown Author. Each of the 100 authors had the lowest sum of squared differences among all authors when that author played the role of the Unknown Author.

The above computations show that:

1. Authors are very consistent relative to the associations between letters of the alphabet and the graphs assigned to them.
2. When the writing samples from authors in a database are as rich as they are in our London letter database, the patterns of letter/graph associations can be a powerful identifier of authorship.

### 3 Biometric Identification

The paper will now turn to the problem of author identification for authors who are observed to be very similar in that the Graph Builder assigns the same graph to their writings of a particular letter. The graph in question, as noted above, will have hundreds of measurements defined for it. The quantity of writings available from the author will determine the number of occurrences of the letter/graph pair. Typically, there will be many more graph measurements than there will be observed occurrences of the letter/graph pair. This makes the data analysis subject to *the curse of dimensionality*.<sup>1</sup>

#### 3.1 Discriminant Analysis

One of the statistical methods we use to distinguish the data for similar authors is stepwise discriminant analysis.<sup>2</sup> We apply this technique by isolating corresponding letter/graph pairs from different handwriting specimens. That is, we inspect different writing samples to locate instances of the same letter written as the same graph, assigned by the Graph Builder. In practice, this task is accomplished by automation that uses recognition to identify the character and uses the Graph Builder to assign the appropriate graph. Once the pairing is done, discriminant analysis will find a small subset of the graph measurements that do a good job of

<sup>1</sup> In this article, the curse of dimensionality refers to the fact that the amount of data available (in terms of occurrences of a letter/graph pair) is insufficient to support an analysis with such high dimensional data (i.e., such a high number of measurements). Some technique must be employed to reduce the measurements to a small set (the Biometric Kernel).

<sup>2</sup> SAS/STAT Users Guide, Version 9.1.

discriminating the data on this letter/graph pair for two authors. Stepwise discriminant analysis will

1. Select a small number of measurements for discriminating the two authors' data; and
2. Provide a weight (canonical coefficient) for each selected measurement.

We combine the values of the selected measurements using the weights; the resulting new measurement (canonical variable) gives the best available discrimination of the data from the two authors. Figure 2 presents a histogram for such a canonical variable and shows the resulting separation of the data from two authors.

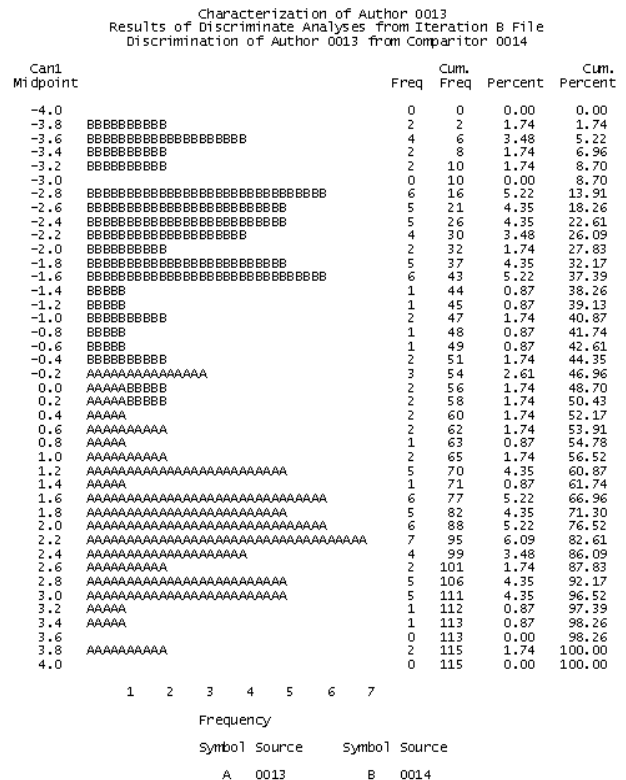


Figure 2: Discriminant Analysis for Two Authors

#### 3.2 The Biometric Kernel

Say that we have 100 authors in our data base who have been observed to be very similar in that the Graph Builder assigns the same graph to some of their writings of a particular letter. We can use stepwise discriminant analysis to compare Author 1 one-by-one to each of the other 99 authors. We combine the results of the 99 stepwise discriminant analyses to reduce the original list of hundreds of measurements defined for the letter/graph pair to a still large but much smaller list of measurements with demonstrated power in discriminating Author 1 from

each of the other 99 authors.<sup>3</sup> By repeating the entire process several times, we are able to isolate about a dozen of the measurements that are powerful for discriminating Author 1 from each of the other 99 authors. We refer to these final dozen (or so) measurements as Author 1's *Biometric Kernel* for the particular letter/graph pair.

Following the steps in the preceding paragraph for each author in the pool of similar<sup>4</sup> authors, we define a Biometric Kernel for each author. We form a database for identification by storing the following information for each cohort of similar<sup>5</sup> authors:

1. The names of the Biometric Kernel measurements for each author.
2. For each pair of authors (say, Author A and Author B), the weights (canonical coefficients), using Author A's Biometric Kernel, that form a canonical variable that discriminates Author A's data for the specified letter/graph pair from Author B's data for the same pair; and the corresponding weights associated with discrimination of Author A's data from Author B's data using Author B's Biometric Kernel.
3. The means and standard deviations of Author A's data and Author B's data for each canonical variable computed in #2.

### 3.3 The Competitive Matrix

We conceptualize the stored information for a cohort of similar authors as defining a *Competitive Matrix*; see Figure 3.

		Comparator Author				
		B1	B2	B3	B4	B5
Modeling Author	A1					
	A2					
	A3					
	A4					
	A5					

Figure 3: The Competitive matrix

<sup>3</sup> The results are a subset of the graph measurements with weights; each comparison to a new author provides a new subset of measurements and associated weights.

<sup>4</sup> Authors are similar in that in their three London Letters used for modeling our Graph Builder assigned the specified graph to some of their writings of the specified letter.

<sup>5</sup> See Footnote 4.

The structure of a Competitive Matrix is determined by associating the matrix's rows with the author whose Biometric Kernel is the basis of the discrimination; we say that rows are determined by the *Modeling Author*. Each column is then associated with a single *Comparison Author*, that is, the author whose Biometric Kernel is not used as the basis of the discrimination. Otherwise stated, the matrix uses each author twice, once as a model author and once as a comparator author. In this way, it is possible to distill those measurements that characterize each author-to-author comparison. Since the same authors exist both as rows and columns, the diagonal axis of the matrix would contain cells where authors are compared against themselves. Since discriminating an author's own data from itself is meaningless in the current context, cells along the diagonal of the matrix are excluded from consideration.

The physical database of three modeling London letters for each author is modeled by a collection of Competitive Matrices. For the examples of this paper, a database of 100 authors is used. Each Competitive Matrix corresponds to one specific letter/graph pair. There is a Competitive matrix for every letter/graph pair observed to occur via use of character recognition and the Graph Builder application.

## 4 Testing the Model

We now discuss testing our Competitive Matrix modeling of author characteristics by using the Competitive Matrices to assign probable authorship to the characters in the two held out London letters for all authors in the database.

### 4.1 Competitive Matrix Voting

Given a character of hypothetically *unknown authorship* from the held out London letters,

1. Quantification Scheme: Associate the character with a Letter and Graph pair by character recognition and the Graph Builder.
2. Evaluate potential authorship *among similar authors in the data base*, i.e., among all authors who were observed to use the referenced Letter/Graph pair in their three London letters which were used for Biometric Kernel modeling. The evaluation is accomplished via *Competitive Matrix Voting*.

In Competitive Matrix Voting, each author plays the roles of both row (i.e., *Modeling*) author and column (i.e., *Comparator*) author; hence the matrix structure.

Each *cell* (intersection of a row and column) is assigned a vote, either 1 or 0 for the row author and a

vote, either 1 or 0, for the column author. The voting logic is presented in Figure 4.

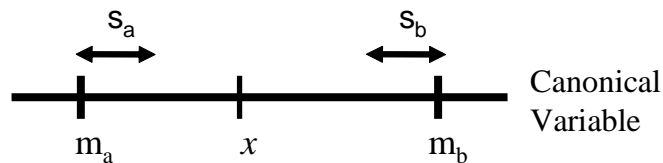
## Competitive Matrix Voting Testing a character of unknown authorship

- $H_0$ : A is the True Author
- Model: based on assumption  $H_0$  is true
  - So use the Biometric Kernel for Author A
- $H_{alt}$ : B is the True Author
- Statistic is Canonical Variable for (letter/graph) pair based on the Biometric Kernel for row Author A.
- Voting Algorithm: if  $|z_a| \leq |z_b|$  and  $-2 < z_a < 2$  then Vote for A  
if  $|z_b| \leq |z_a|$  and  $-2 < z_b < 2$  then Vote for B

It is possible that neither A nor B receives a vote. We call such a situation a 'no-vote' case.

$$z_a = (x - m_a) / s_a$$

$$z_b = (x - m_b) / s_b$$



$m_a$  and  $m_b$  are the means of the data for Author A and Author B.  
 $s_a$  and  $s_b$  are the standard deviations of the data for Author A and Author B.  
 $x$  is the canonical variable (statistic) value for the character.

**Comparator**

	B1	B2	B3	B4	B5
A1					
A2					
A3					
A4					
A5					

**Modeling  
Author**

Figure 4: Competitive Matrix Voting

The next step summarizes each author's row votes and column votes.

Row Voting Using the Competitive Matrix with  $m$  Authors

- For Author A, Add the votes for A across all columns (all comparator authors)
- This amounts to  $m-1$  tests of  $H_0$ : *A is the true Author*
- One test for each column Comparator Author B
- Each failure to reject the  $H_0$  with a plausible value for  $x$  is a vote for A
- An average vote close to 1 means that the unknown data is more consistent with the data used to model Author A than with the data from comparator authors.

Column Voting Using the Competitive Matrix with  $m$  Authors

- For Comparator Author B, sum the votes for B across all rows (all modeling authors)
- This amounts to 1 test of  $H_0$ : *A is the true Author* for each *modeling* Author A
- Each rejection of  $H_0$  in favor of  $H_{alt}$ : *B is the Author* with a plausible value for  $x$  is a vote for B
- An average vote close to 1 means that the unknown data is more consistent with the data for author B than with the data used to model most other Authors

As one's intuition would expect, an author's row and column votes are correlated; that is, both row and column votes are low or both are high. Typically, we have observed that the true author will have a fraction of row votes close to 1 and a fraction of column votes close to 1. However, as seen in Figure 5 some authors other than the true author might also have a fraction of row votes close to 1 and a fraction of column votes close to 1. In Figure 5, each author

from the Competitive Matrix is represented by a symbol (1 or 0) plotted according to that author's fraction of row votes (horizontally) and fraction of

column votes (vertically). The plotting symbol for the true author is '1' and the plotting symbol for all other authors is '0'.

### COMPARING COLUMN VOTING TO ROW VOTING

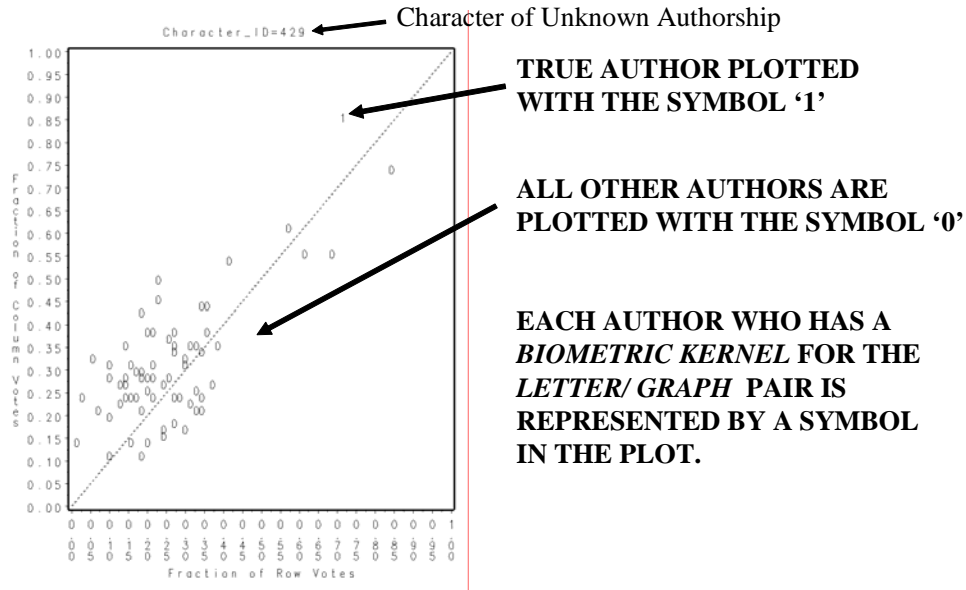


Figure 5: Comparing Row and Column Voting

We seek a single value (statistic) that reflects both how close the fraction of row votes is to 1 and how close the fraction of column votes is to 1. Note that

each symbol (author) in Figure 5 can be associated with a two-by-two frequency table. This table association is defined in Figure 6.

FOR EACH AUTHOR (A) IN THE COMPETITIVE MATRIX OF POTENTIAL AUTHORS, GET A 2x2 TABLE

Consider the LOG of the Odds Ratio  $\ln\left(\frac{ab}{cd}\right) = \ln\left(\frac{a}{d}\right) + \ln\left(\frac{b}{c}\right)$

	Row Tests	Column Tests
Accept $H_0$	a	d
Reject $H_0$ Accept $H_{alt}$	c	b

- a = Row Votes for Row Author ( $H_0$ : A)
- c = Row Votes for Column Author ( $H_{alt}$ : Other)
- d = Column Votes for Row Author ( $H_0$ : Other)
- b = Column Votes for Column Author ( $H_{alt}$ : A)

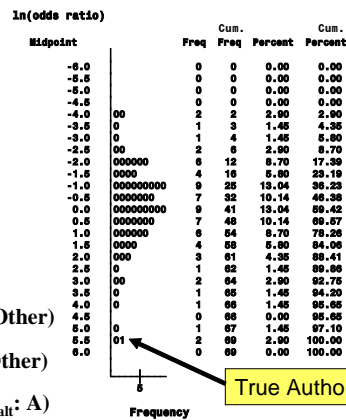


Figure 6: The Log of the Odds Ratio

## 4.2 The Log of the Odds Ratio

Considering this two-by-two table, a single value (statistic) with the desired properties is the log of the odds ratio, that is,  $\log(ab/cd)$ . A histogram for  $\log(ab/cd)$  for the data displayed in the plot of Figure 5 is presented in Figure 6. Note that the true author (symbol '1') and another author (symbol '0') appear tied for the highest value of  $\log(ab/cd)$ . These two authors are out in the far tail of the distribution of  $\log(ab/cd)$  values for all of the 69 authors who are in the Competitive Matrix.

Figure 6 parses the  $\log(ab/cd)$  value into the sum of two log odds values:

$$\log(ab/cd) = \log(a/d) + \log(b/c).$$

This breakdown of the log of the odds ratio helps us to appreciate the power for true author identification in this quantity. Figure 7 plots  $a$  versus  $d$  for the true authors associated with all characters in the held out two London letters for all authors in our database. Note in Figure 7 the strong pattern of plotted points with  $a > d$ ; such points will give large values of  $\log(a/d)$ .

Alternatively, Figure 8 plots  $a$  versus  $d$  for the authors who are not the true authors for the same

characters in the held out two London letters. Note in Figure 8, the symmetric pattern of plotted points around the 45 degree line where  $a = d$ ; for such points the values of  $\log(a/d)$  will be symmetric around 0.

Figure 9 plots  $c$  versus  $b$  for the true authors associated with all characters in the held out two London letters for all authors in our database. Note in Figure 9, the strong pattern of plotted points with  $b > c$ ; such points will give large values of  $\log(b/c)$ .

Alternatively, Figure 10 plots  $c$  versus  $b$  for the authors who are not the true authors for the same characters in the held out two London letters. Note in Figure 10, the symmetric pattern of plotted points around the 45 degree line where  $b = c$ ; for such points the values of  $\log(b/c)$  will be symmetric around 0.

Figure 11 presents a histogram of  $\log(ab/cd)$  for all authors (both true authors and others) from the Competitive Matrices for each character in all held out London letters. Note the smooth, symmetric structure of the histogram. The parts of the histogram associated with values for true authors are darkened in; note that the values for true authors are predominantly positive whereas the values for other authors are randomly positive or negative.

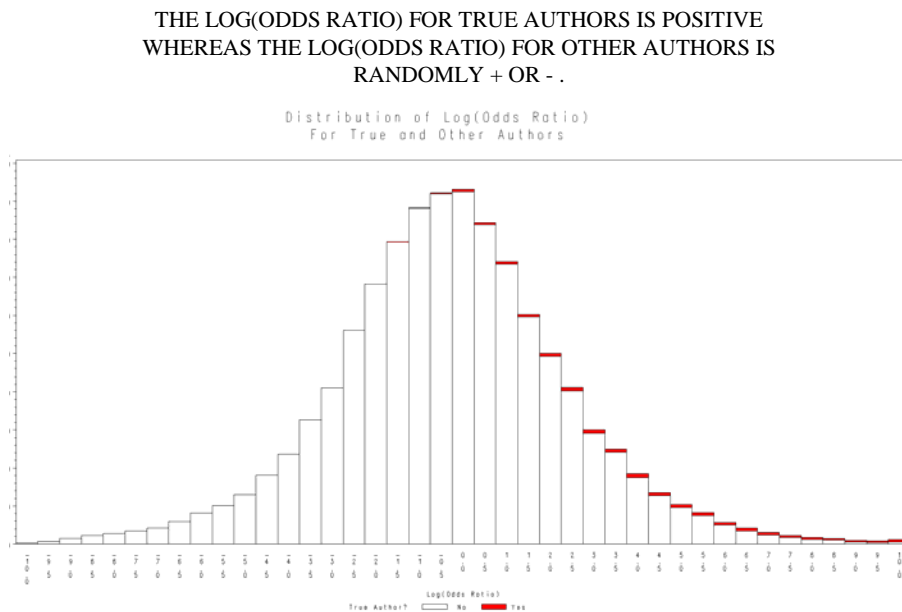
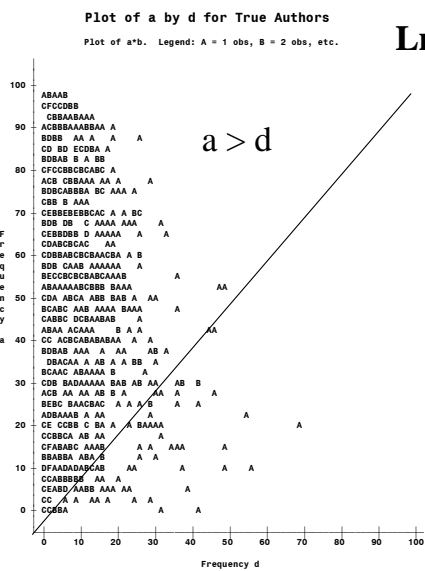


Figure 11: The distribution of the Log of the odds ratio

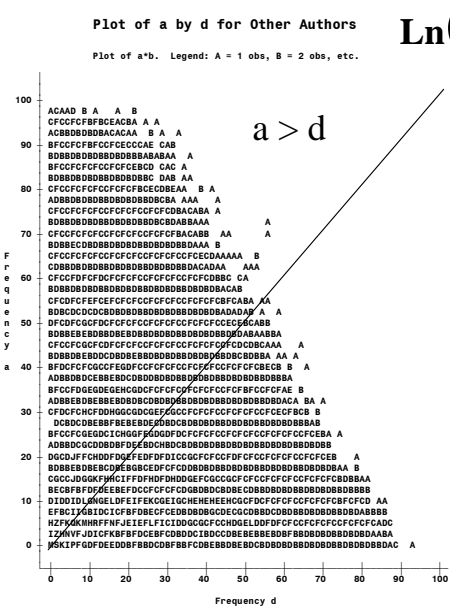


$$\ln\left(\frac{ab}{cd}\right) = \ln\left(\frac{a}{d}\right) + \ln\left(\frac{b}{c}\right)$$

	Row Tests	Column Tests
Accept $H_0$	a	d
Reject $H_0$ Accept $H_{alt}$	c	b

$a/d =$  Odds that an Accepted  $H_0$  is from a Row Test

Figure 7: Plot of a versus d for True Authors



$$\ln\left(\frac{ab}{cd}\right) = \ln\left(\frac{a}{d}\right) + \ln\left(\frac{b}{c}\right)$$

	Row Tests	Column Tests
Accept $H_0$	a	d
Reject $H_0$ Accept $H_{alt}$	c	b

$a/d =$  Odds that an Accepted  $H_0$  is from a Row Test

Figure 8: Plot of a versus d for Other Authors

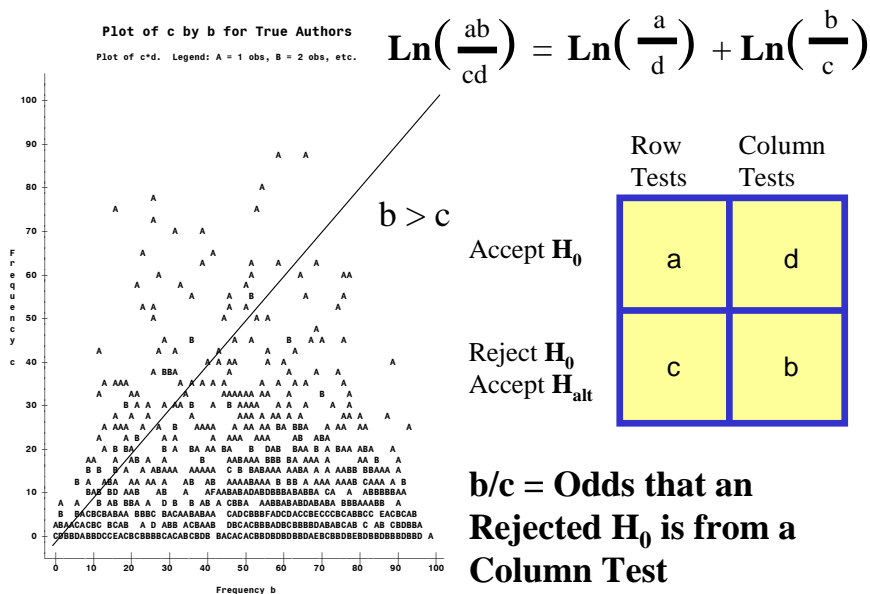


Figure 9: Plot of  $b$  versus  $c$  for True Authors

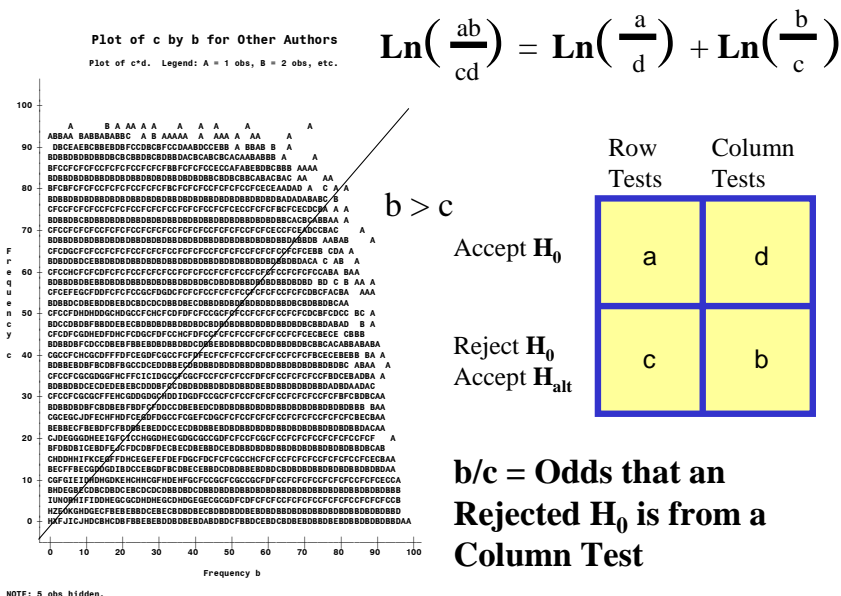


Figure 10: Plot of  $b$  versus  $c$  for Other Authors



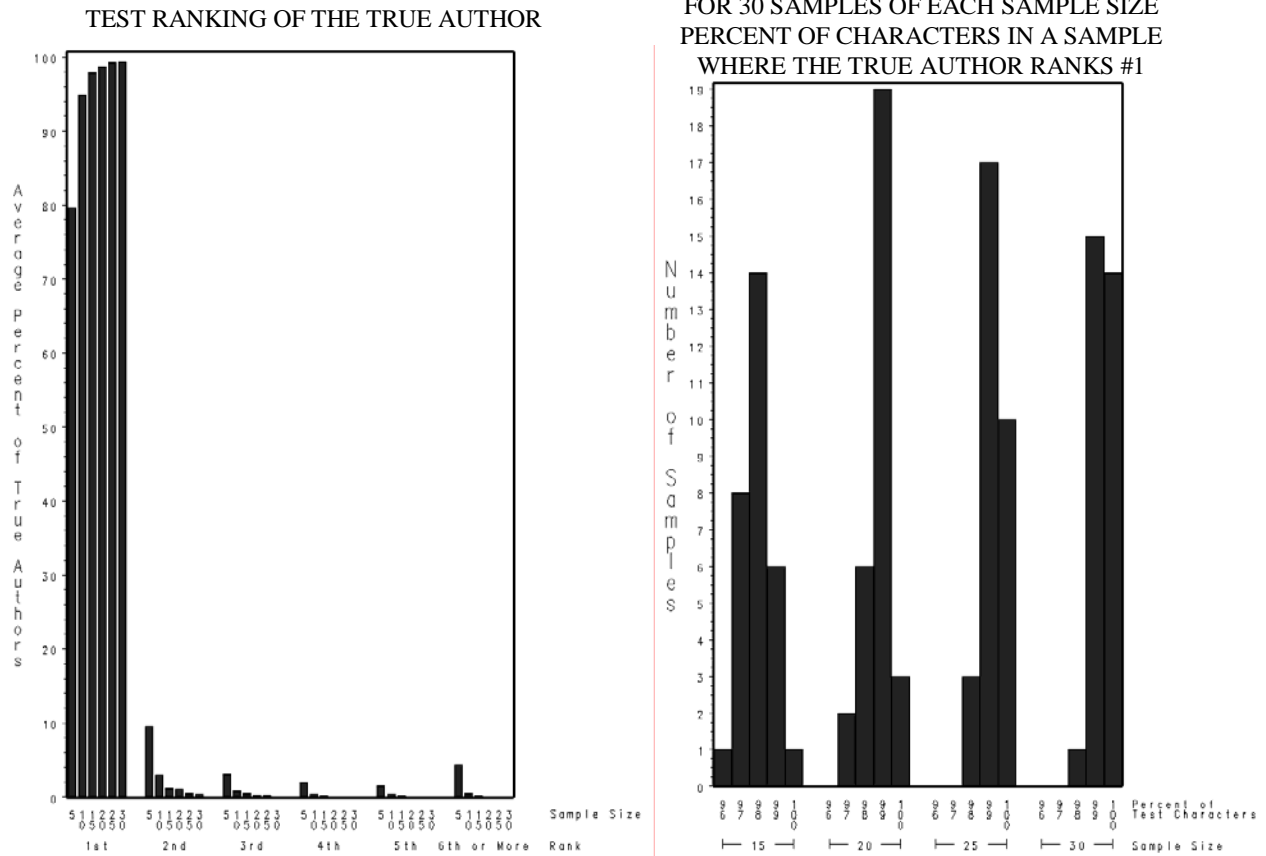


Figure 13: The accuracy of the Biometric Identification

## 5 Conclusion

We have demonstrated that the quantification of handwriting by the assignment of a graph to each character provides a very powerful basis for biometric identification. The concepts of Biometric Kernels and Competitive Matrices are used to build a biometric identification database for use in determining the identity of an unknown author of a writing sample.

We are applying the technology described in this article to applications in the areas of Individual Identification, Document Clustering and Forensic Document Examination.