GANNON
TECHNOLOGIES GROUP

NIST

FLASH ID

Handwriting Derived Biometric Analysis Software

June 4, 2013

NIST

Measurement Science and Standards in Forensic Handwriting Analysis Conference

Donald T. Gantz, PhD
Mark A. Walch

FLASH ID: *Introduction*

FLASH ID: Forensic Language Independent Analysis System for Handwriting Identification

We describe the workings of the FLASH ID system in the following steps:

1. **Image Processing**
2. **Writer Modeling**
3. **Writer Scoring**

We then present the results of our tests of the system

FLASH ID: *Framework*

FLASH ID has a well-defined frame for comparing handwriting; it combines

- **a method for segmenting writing into graphs;**

- **an isomorphic classification for graphs;**

Isomorphism Class 4;112.0

- **a meaningful shape definition for graphs.**

Shape Class

This frame allows that writings be compared in a "like to like" manner.

FLASH ID: *Image Processing*

Image Processing: FLASH ID takes a scanned image (300 dpi) of a document, segments it, and skeletonizes each segment, yielding a one pixel wide skeleton. The writing can be regarded as a combination of graphs with nodes and edges.

Original Word:

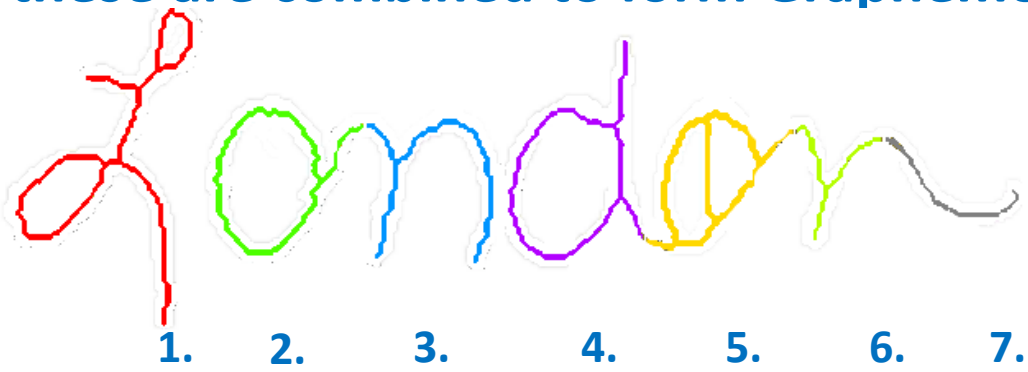
Original Word with skeleton exposed:

Extracted Skeleton (graph):

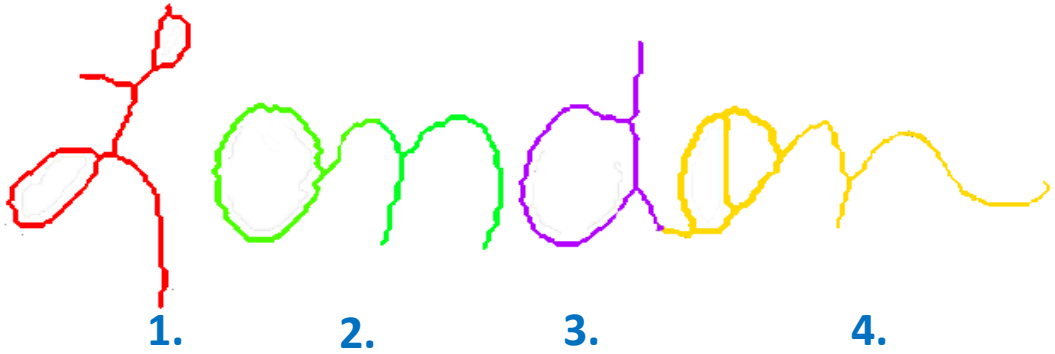
FLASH ID: Segmentation into *Proto-graphemes*

***Proto-graphemes*: these are combined to form Graphemes.**

Original Proto-graphemes:



Combined Proto-graphemes: Forms Graphemes



Embeddings from selected Proto-graphemes:

FLASH ID: *Encoding Handwriting as Data*

The connectivity pattern of the nodes and edges defines the isomorphism class of graphs.

Neither bending the edges, changing the length of the edges, reorienting the image, nor reflecting the image changes the isomorphism class.

Each graph has a *Feature Vector* of physical measurements.

The shape class of a graph is determined by the geometric relationships between parts of the graph.

Handwritten text

Isomorphism Class 4;112.0

Shape Class for handwritten text

FLASH ID: *Encoding Handwriting*

Graphemes:

Isomorphic Classes:

Class Feature Vector:

_2001_2_1	_2001_4_1	_2001_4_2	_2001_4_3	_2001_4_4	_2001_8_1	_2001_8_2	_2001_8_3	_2001_8_4	_2001_16_1
4.4721	20.2781	2.2361	24.7133	2.2361	36.4005	2.2361	40.8044	2.2361	22.4939
5.099	11.2154	2.6352	16.3144	2.5055	22.561	3.1623	27.6586	2	13.8339
4.1231	13.1432	2.1541	17.2658	2.01	24.5153	2.2361	28.6356	2	15.2845
22.4722	12.6549	24.2389	30.1249	10.6367	24.1868	32.8024	40	18.0278	35.0105
4.1231	12.6549	2.0881	16.6182	2.0396	24.1868	2.2361	28.0713	2	14.6227
27.7849	17.1363	15.1691	31.5691	14.5809	31.6228	17.4929	40.4475	15.0333	25.8671

FLASH ID: *Fully Automated*

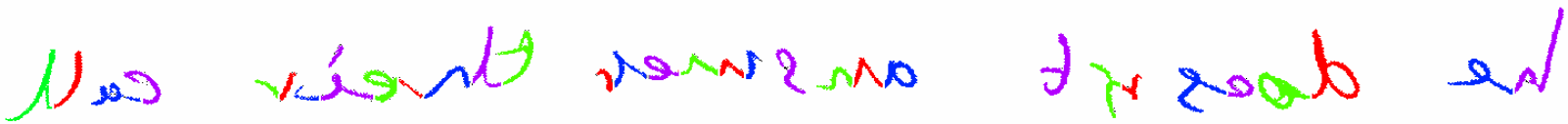
FLASH ID contains proprietary software which automatically segments writing and assigns an isomorphism class and a shape class to each graph.

FLASH ID: *Language Independence*

Handwritten Cursive English and Handwritten Arabic are segmented into graphemes by a common algorithm.

English Script:

English "Mirror Writing":



Arabic:

FLASH ID: *Writer Modeling*

Modeling based on Writers' known writings:

The combination of an isomorphism class and a shape class defines a *Grapheme Type*.

***Categorical Modeling* at the document level: The Table classifying the Grapheme Types for a Writer's known writings characterizes that Writer, and it is used to identify the Writer of Questioned Writing.**

Writing is also modeled at the *Grapheme Type lev*el.

> ***Feature Space Modeling*: The physical measurements within the Feature Vector of the graph are used to form pairwise comparisons of Writers whose known writings exhibit instances of the Grapheme Type.**
>
> **Stored results from the pairwise comparisons of Writers are used to identify the Writer of Questioned Writing.**

FLASH ID: *Feature Space Modeling*

FLASH ID incorporates Feature Space Modeling at the Grapheme Type level.

The known writings for Writers in a database of interest are compared to writings from an unrelated Reference Set of Writers each of whom has a large known writing corpus.

The comparisons provide a stored scoring database that is used to evaluate potential writership of a questioned writing by members of the database of interest.

FLASH ID: *Writer Scoring*

Scoring of Questioned Writing:

The Questioned Writing is preprocessed yielding *Analysis Graphemes*.

Each Writer in the database of interest whose known writings contain instances of the Grapheme Type of the Analysis Grapheme receives a *Reward Score* for the Analysis Grapheme.

Writers' Reward Scores are accumulated for all Analysis Graphemes in the Questioned Writing.

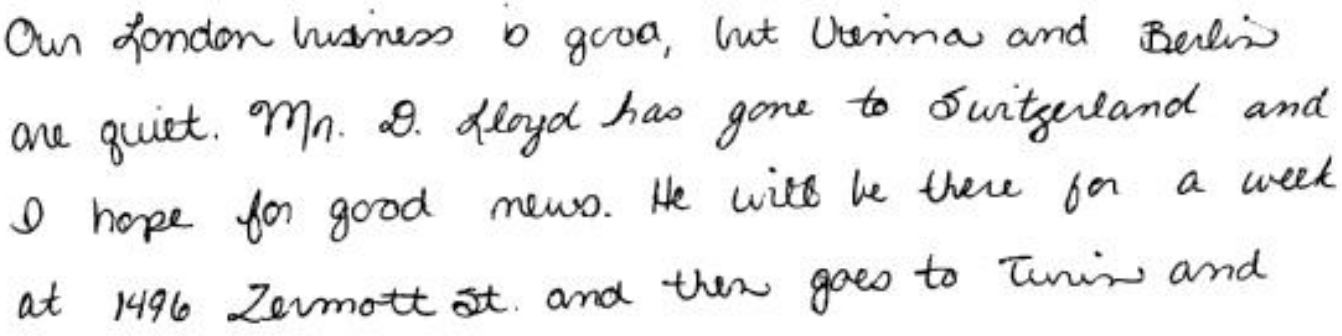
Writers in the database of interest are *ranked* according to their cumulative Reward Scores for the Questioned Writing.

FLASH ID: *Writer Scoring*

The histograms on the next slide show the scores for two test documents written by the writer illustrated below. Language-independent segmentation was used.

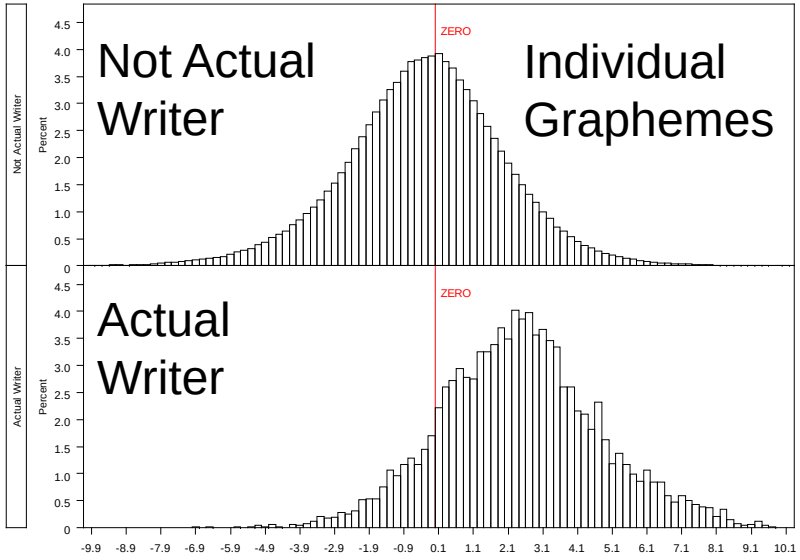
Note that the individual grapheme scores for the actual writer tend to be positive while the scores for the non-writers tend to center closer to zero.

When *pseudo documents* are created by sampling the graphemes, the scores over the pseudo documents move to the right for the actual writer (as the size of the pseudo-document increases) and do not for the non-writers.

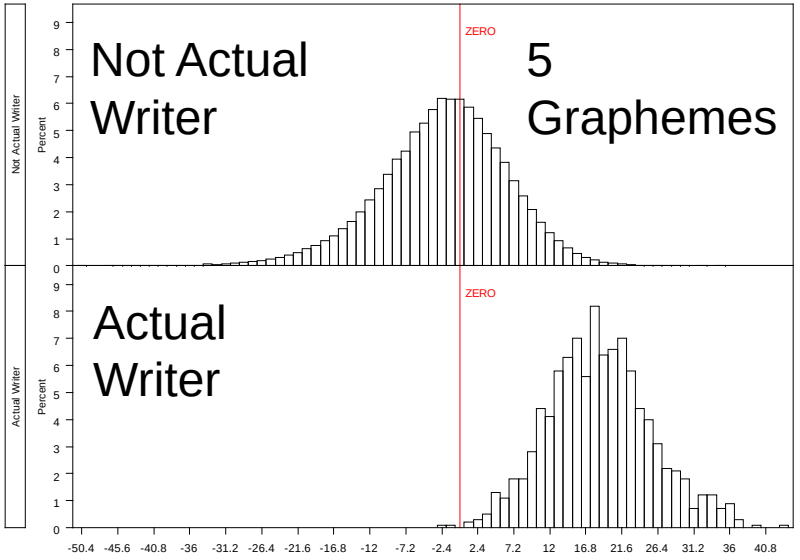

Our London business is good, but Vienna and Berlin
are quiet. Mr. D. Lloyd has gone to Switzerland and
I hope for good news. He will be there for a week
at 1496 Zermott St. and then goes to Turin and

FLASH ID: *Writer Scoring*

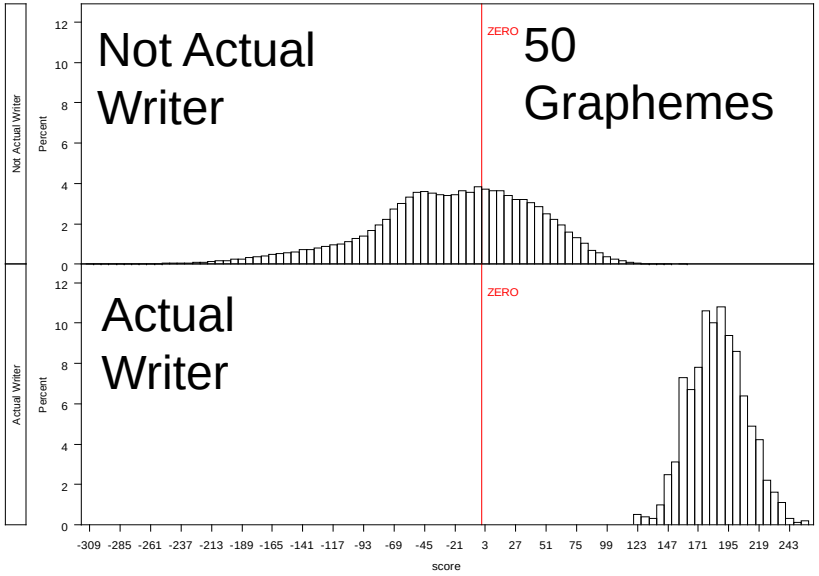
Actual Writer vs. Not for Individual Graphemes



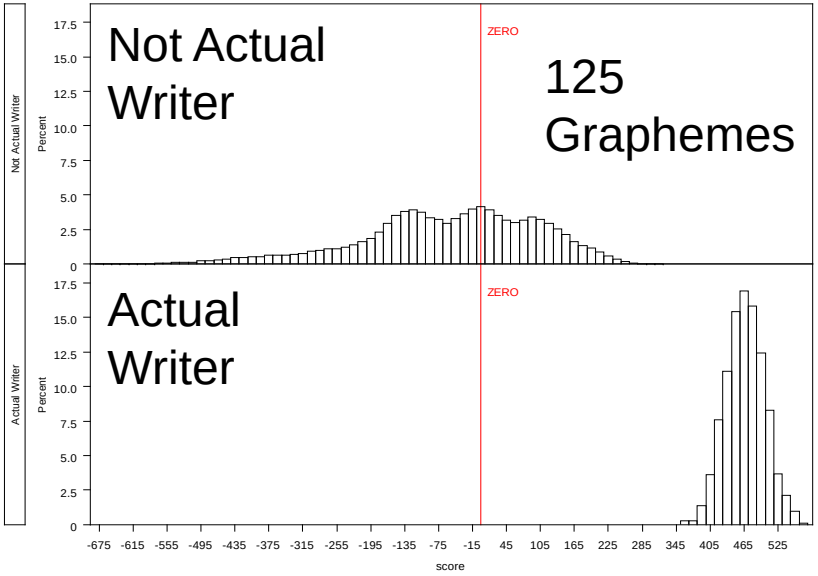
Actual Writer vs. Not for Pseudo Documents of Size 5 Graphemes



Actual Writer vs. Not for Pseudo Documents of Size 50 Graphemes



Actual Writer vs. Not for Pseudo Documents of Size 125 Graphemes



FLASH ID: *Testing the Technology*

[AAFS 2009] Our testing results use writings from the "FBI 500." The FBI Lab acquired handwriting samples from about 500 writers (both script and print). We will give results for 449 writers for whom we have access to script samples. Each writer wrote about five "London paragraphs."

The first 300 writers formed the database of interest. We used three paragraphs for modeling and two for Closed Set testing. The remaining 149 writers were "not in the database of interest" and they were used in the Open Set testing.

In each study, we created pseudo-documents by random sampling from available graphemes. In the character-based studies, we created for each test writer 1000 pseudo-documents of sizes 10 to 40 characters and 500 of sizes 50 to 80. In the language-independent studies, we created 500 pseudo-documents for each writer for each pseudo-document size.

FLASH ID: *Testing the Technology*

For each study, we computed an accuracy rate for the "closed set problem" which is to identify the writer in the database of interest most likely to have written the test document. (For our studies, "correct" means that the *actual writer was ranked first*.)

We also calculated an *Equal Error Rate* for the "open set problem" which is to decide whether or not the writer of a test document is in the database of interest. The equal error rate results from adjusting the *Inclusion Criterion* (in our case, the difference in score between the first and second ranked writers) so that the rates of false inclusion and false exclusion are the same.

The results are summarized in Tables 1 and 2 below.

FLASH ID: *Testing the Technology*

The Latin character-based studies used ground truth segmentation into actual Latin Characters as graphs. For testing, we used only graphemes of the "T" (4;112.0 code) isomorphism representing lower case letters *n, e, r, u, a, l, s, h, o, c, l*.

Table 1: Character-based Data		
Pseudo-document Size	Accuracy Rate	Equal Error Rate
10 characters	89.49%	17.37%
20 characters	98.20%	10.51%
30 characters	99.56%	7.43%
40 characters	99.84%	5.80%
60 characters	99.94%	4.91%
60 characters	99.97%	4.37%
70 characters	99.99%	3.93%
80 characters	100.00%	3.66%

Different writers used different numbers of these characters to write their paragraphs, but on average the particular characters used in this study comprised less than 16% of the available characters in the London paragraph.

FLASH ID: *Testing the Technology*

The language-independent studies used all grapheme types.

Table 2: Language-independent Data				
Pseudo-document Size	Including Document 62		Excluding Document 62	
	Accuracy Rate	Equal Error Rate	Accuracy Rate	Equal Error Rate
50 graphemes	97.26%	12.86%	97.69%	12.63%
100 graphemes	99.03%	7.51%	99.49%	7.24%
200 graphemes	99.37%	4.86%	99.85%	4.58%
300 graphemes	99.41%	4.10%	99.91%	3.82%
400 graphemes	99.44%	3.63%	99.94%	3.35%
500 graphemes	99.47%	3.34%	99.97%	3.06%
600 graphemes	99.48%	3.16%	99.98%	2.90%
700 graphemes	99.49%	3.06%	99.99%	2.80%
800 graphemes	99.49%	2.98%	99.99%	2.71%
900 graphemes	99.49%	2.92%	99.99%	2.66%
1000 graphemes	99.50%	2.87%	100.00%	2.63%
1500 graphemes	99.50%	2.79%	100.00%	2.55%
2000 graphemes	99.50%	2.72%	100.00%	2.51%

FLASH ID: *Testing the Technology*

[AAFS 2010] The Open Set Problem Revisited:
We introduced a new inclusion criterion, that is, the score used to evaluate the difference between the first and second ranked writers in order to make the open set decision. We used our "Vector of Counts" methodology, which has demonstrated great potential to improve open set results.

The "Vector of Counts" (VOC) methodology requires a reference set of documents from writers neither in the database of interest nor likely to have written a questioned document.

FLASH ID: *Testing the Technology*

We create a Grapheme Type level, Feature Space Modeling system from documents of writers in this new reference set. This modeled system is different than the modeled system first used to score the questioned document.

For each grapheme in a questioned document, the writer in the reference set most likely to have written that grapheme is recorded. A *frequency Vector of Counts* is thus created for the questioned document.

The pattern of proportions of graphemes assigned to writers in the reference set characterizes the writer of the questioned document.

FLASH ID: *Testing the Technology*

Our finding is that documents written by the same writer have similar (or at least "close") VOCs. We can measure closeness using a *chi-squared statistic.*

By treating the known writings for the database of interest writers as questioned documents, we create a VOC for each writer.

We take the difference between the closeness of the first place and second place VOCs to the questioned document VOC.

If the questioned document is actually written by the first place writer, the difference should be relatively large. If the document is not written by a writer in the database of interest, the difference should be relatively small.

FLASH ID: *Testing the Technology*

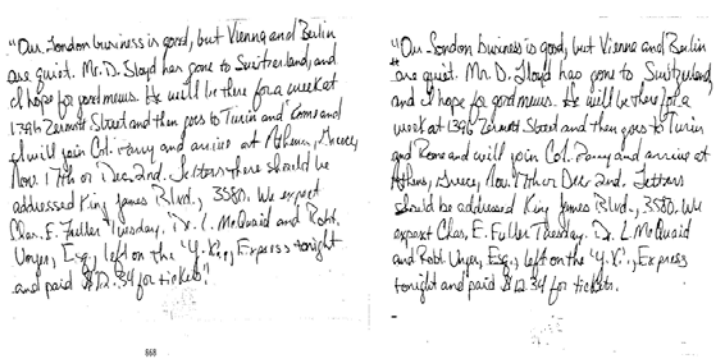
[AAFS 2010] Results:

The equal error rate (EER) improved between 10% and 50% by using the VOC information in addition to the score difference information.

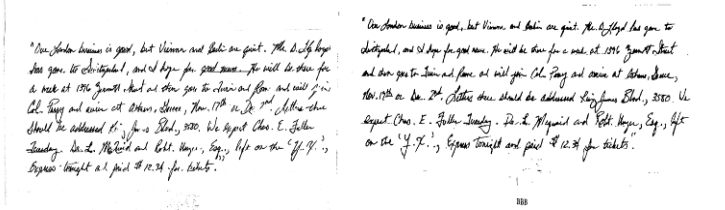
For instance, if the score difference EER had been 5 percent, the improved EER would have been between 2.5% and 4.5%.

FLASH ID: *USACIL 1,000 Writer Blind Test*

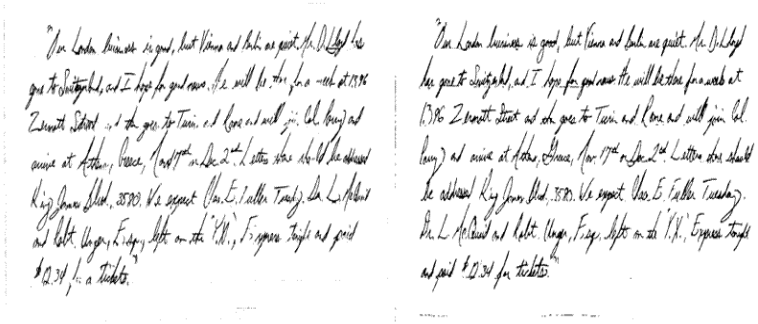
Reference match 868 and unknown image AAA



Reference match 892 and unknown image BBB



Reference match 360 and unknown image AA



USACIL Closed Set Blind Test: 56 London paragraphs had a match among a data base of 1,000 London paragraphs. 55 were top ranked by FLASH ID.

Rank	Results		
	Number	Percent	Cumulative
1	55	98.2	98.2
2	0	0	98.2
3	0	0	98.2
4	0	0	98.2
5	0	0	98.2
6-10	1	1.8	100.0

FLASH ID: *FBI End User Enhancement Final Report*

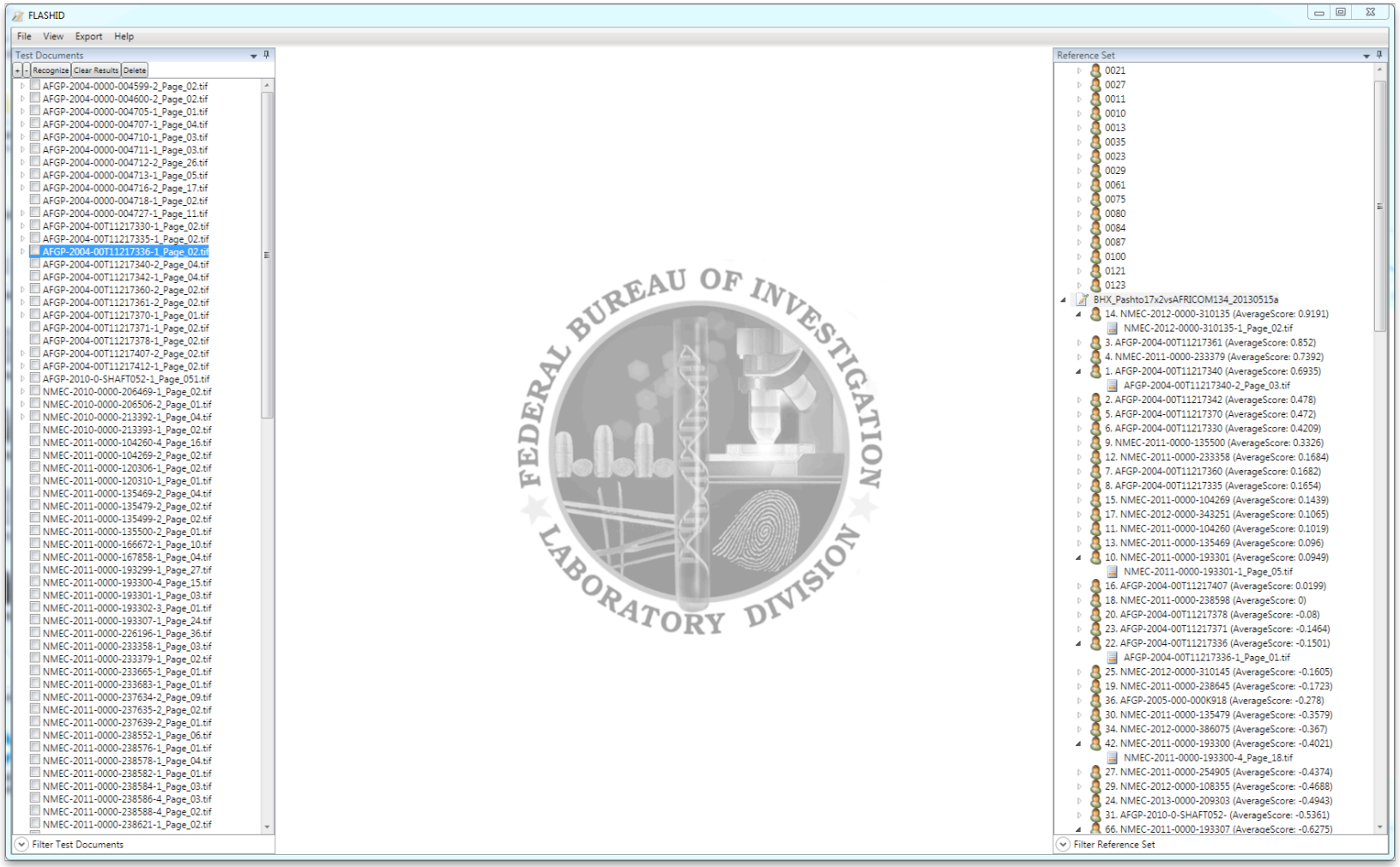
FBI End User Enhancement Final Open Set Test: 1,025 Test Writer Documents were compared to a data base of 1,025 writers.

An Inclusion Criterion was used to decide if the top ranked data base writer was the writer of a test document.

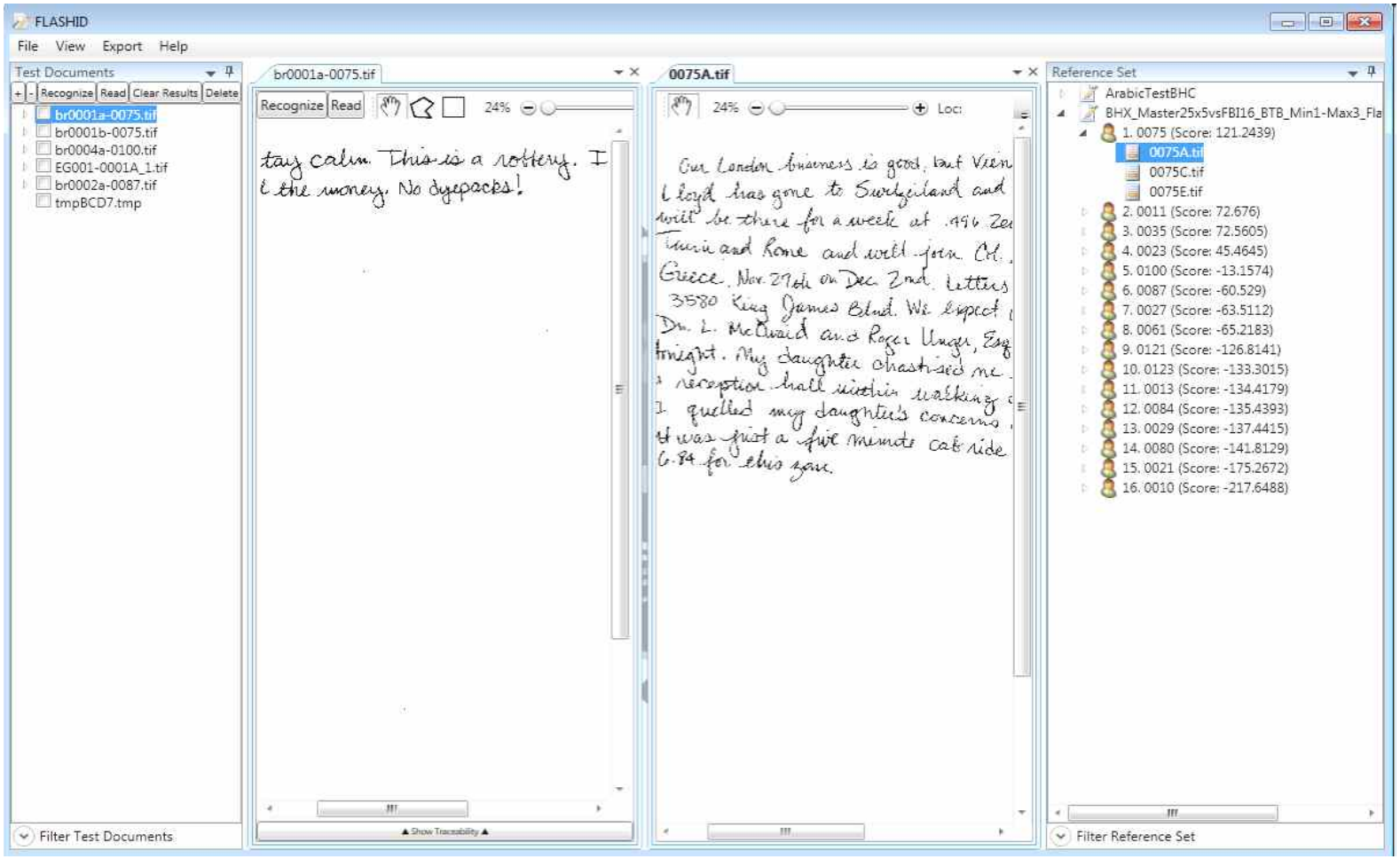
Of the 525 writers with a match, the matching writing was ranked first for 523, but 5 of these were determined to be not a match by the inclusion criterion. (Error rate is 7/525 = 1.3%)

Of the 500 writers without a match, by the inclusion criterion 493 top ranked data base writers were declared not to have a match and 7 were declared to be a match. (Error rate is 7/500 = 1.4%)

FLASH ID: *Implementation—the FDE Workstation*

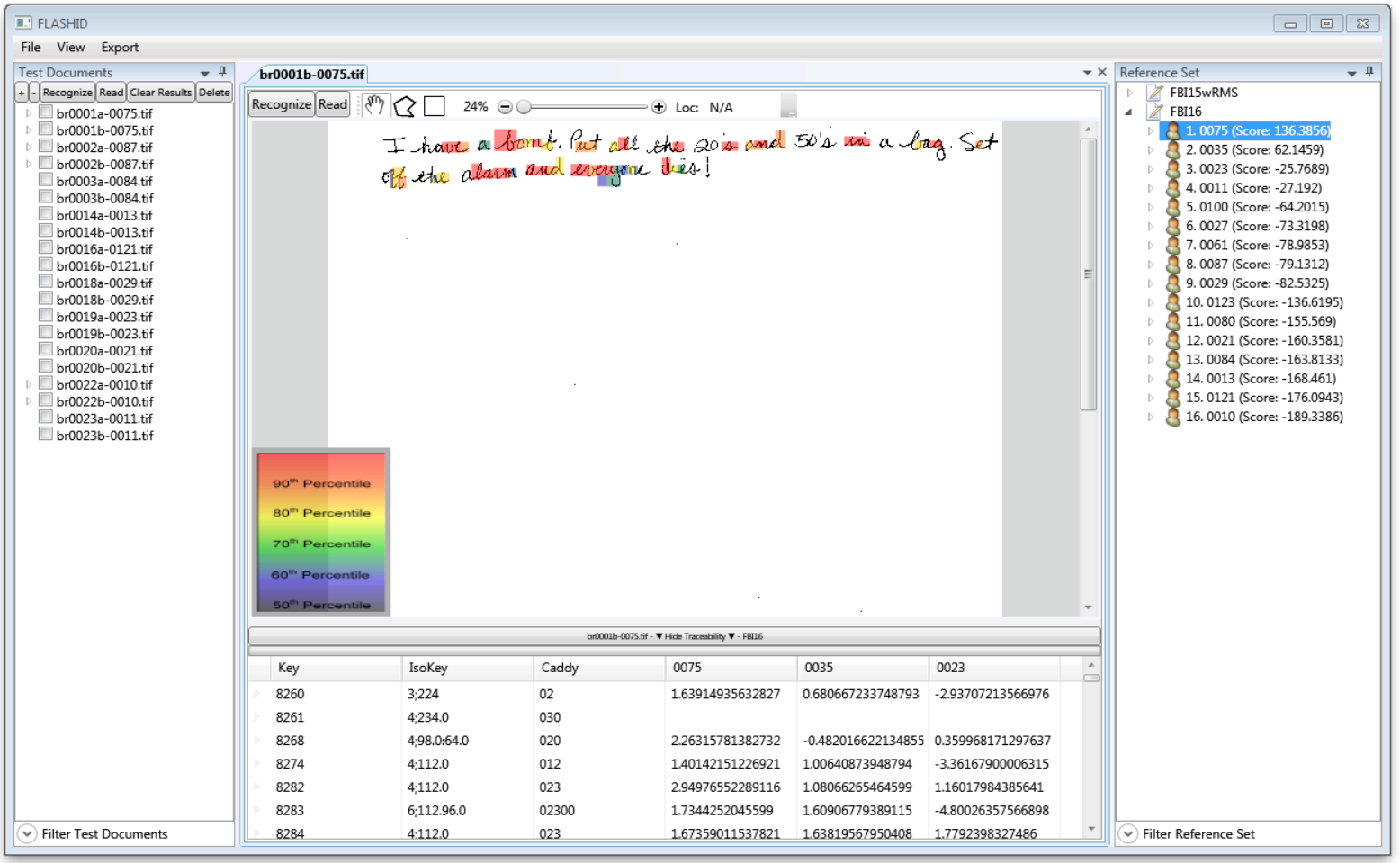


FLASH ID: *Side-by-Side View*



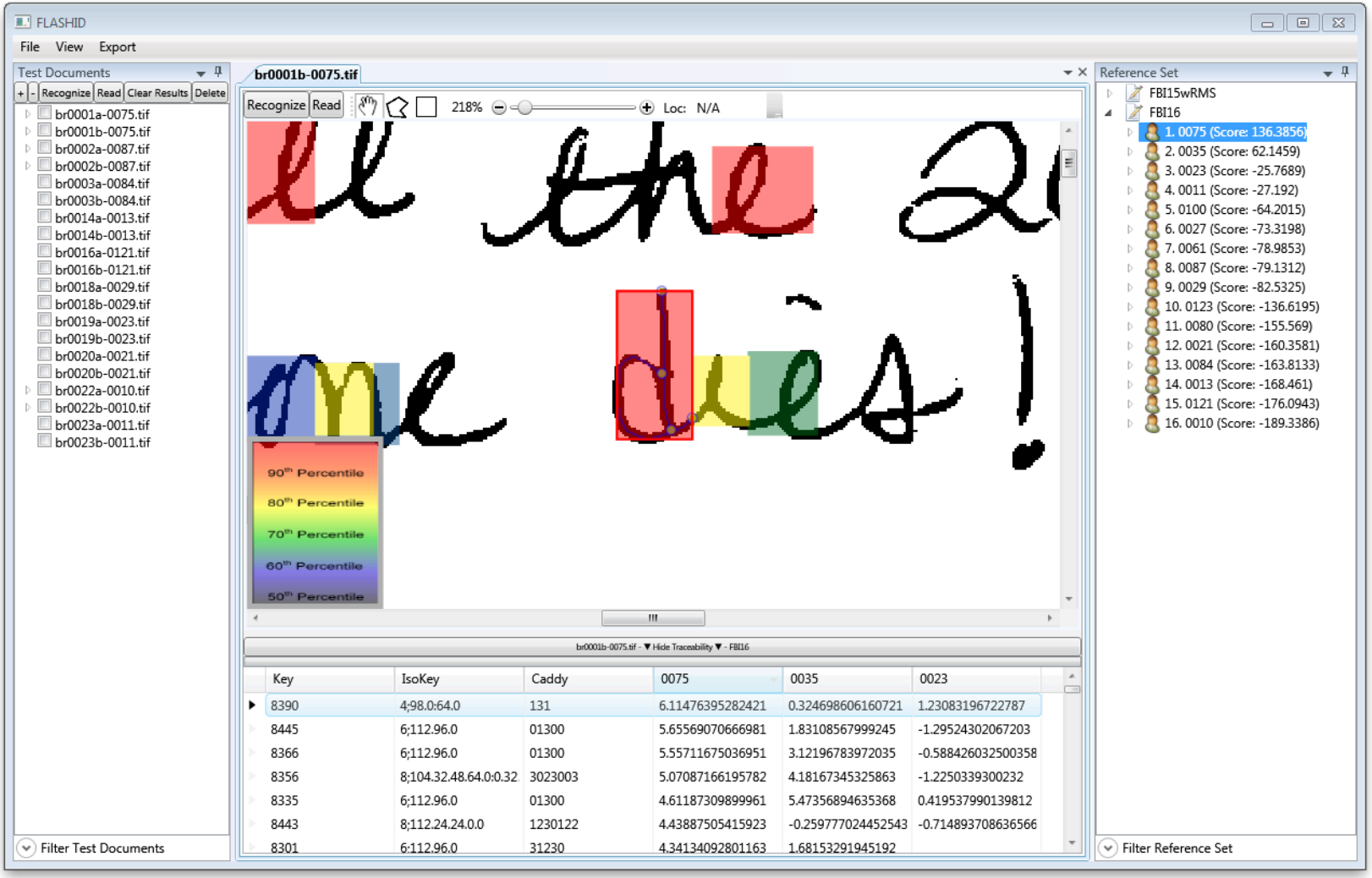
FLASH ID: *Identity Heat Map View*

FLASH ID Forensic Document Examiner workstation at FBI Laboratory



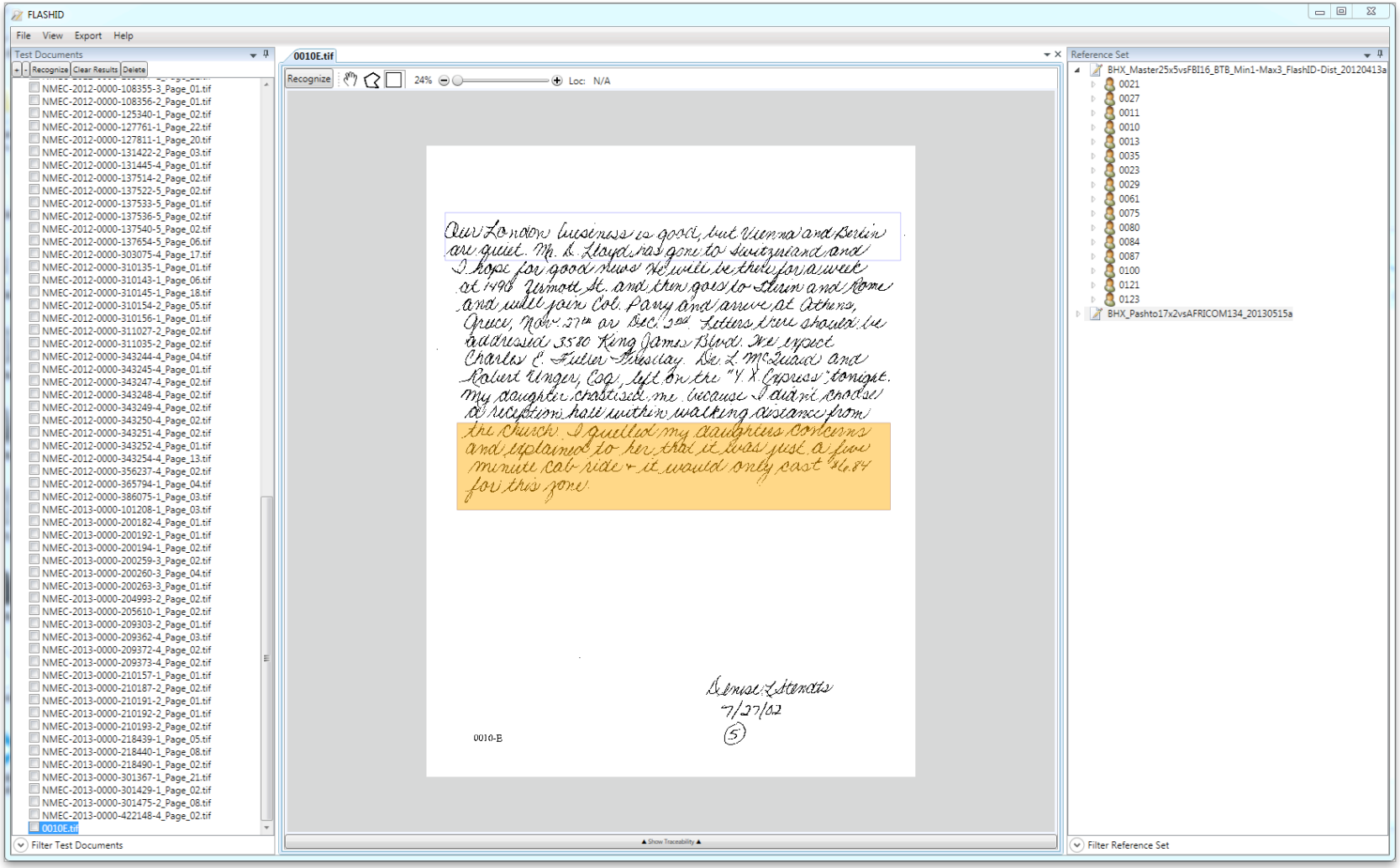
FLASH ID: *Trial Exhibit Support*

Outline of writing characteristics showing strong "biometric signal"



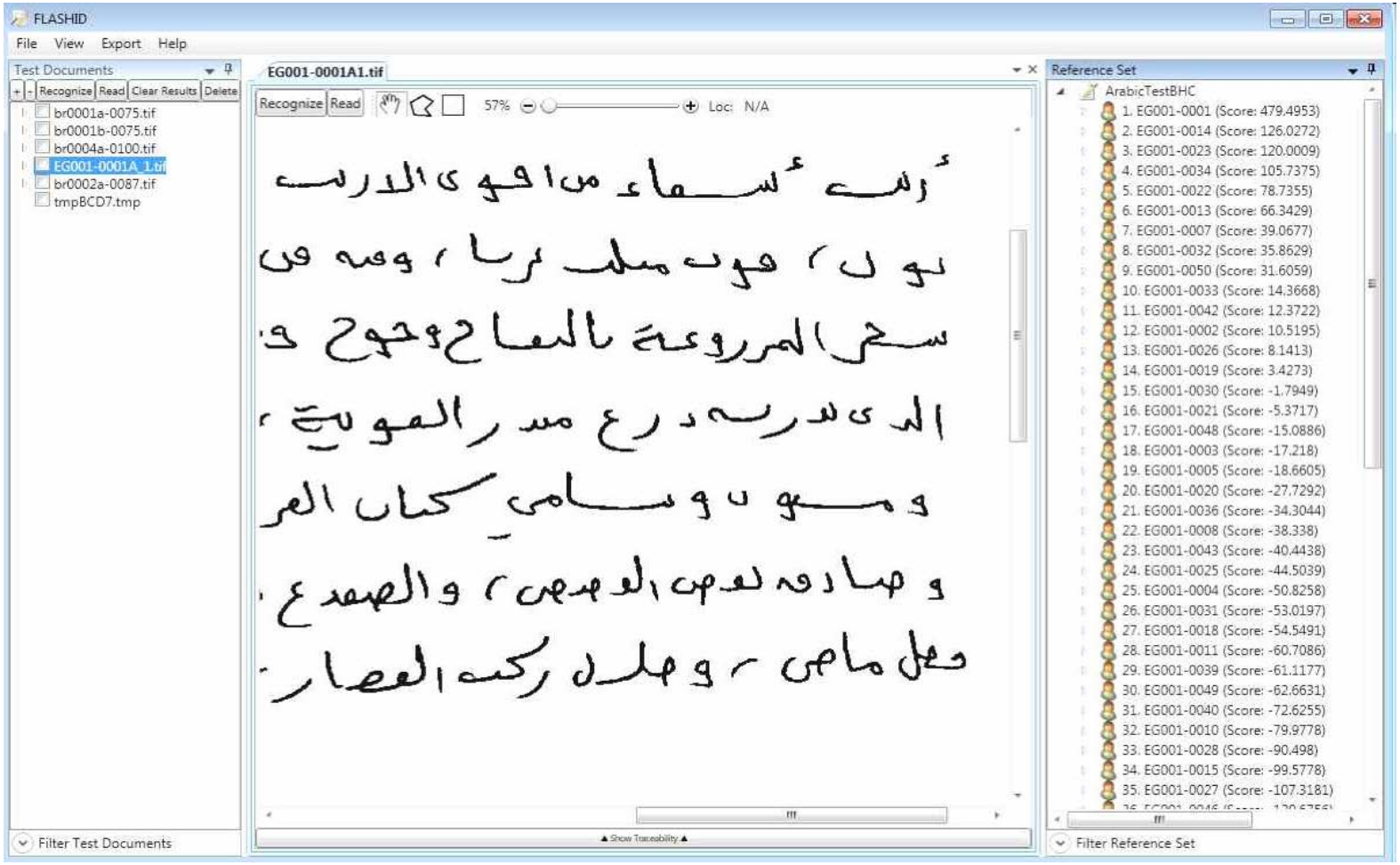
FLASH ID: *Text Selection Capability*

Ability to select text of interest from within documents.



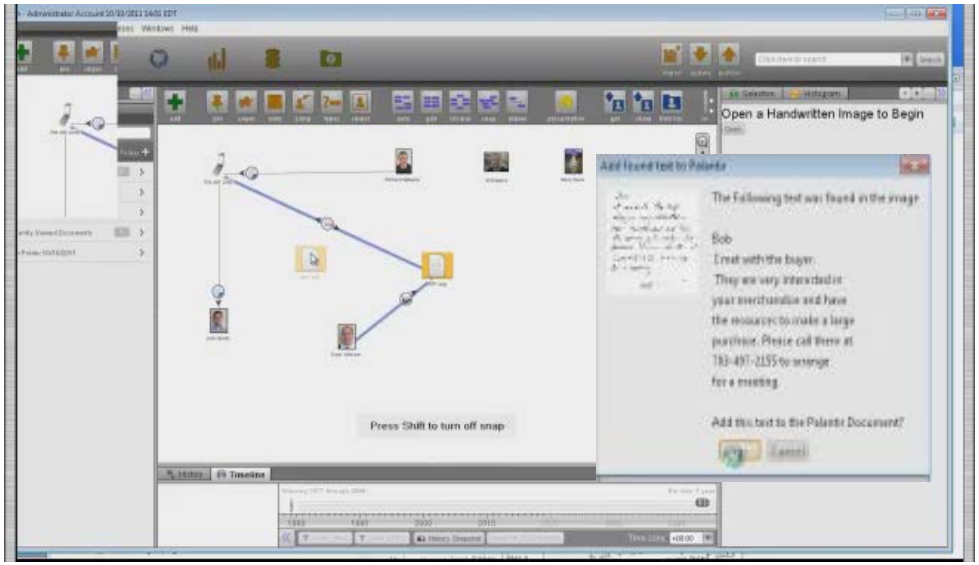
FLASH ID: *Language Independent*

Works across multiple languages and scripts.



FLASH ID: *Palantir Integration and Cluster View*

Integrated as a "helper" in Palantir Analytical Platform.



Cluster view provides "big data" visualization perspective to handwritten document collections.