Statistical Analysis of Letter Importance for Document Examination

Amy Crawford and Dr. Alicia Carriquiry
Iowa State University

csafe Center for Statistics and Applications in Forensic Evidence

NIST National Institute of Standards and Technology U.S. Department of Commerce

1. Background

The *principle of individuality* in handwriting analysis says that, given sufficient quantity and quality of writing to compare, every person has unique writing characteristics [1]. So while our writing has natural variation from one line of text to the next, there are still individualizing features that reside there.

We take a context (and language) independent, automated approach to statistically assess whether a piece of questioned writing lies within the natural variability of any one of our writers (and outside the natural variability of the rest).

2. Data

- Images of words were skeletonized and broken into small manageable pieces called "graphemes" by the Flash ID® software [2].

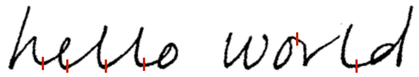
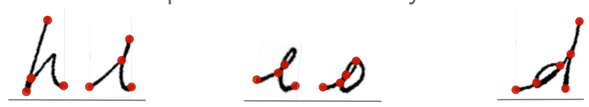


Figure 1: Example of locations where the software might split words to generate graphemes.

- A **node** is a terminal point, a point where three or more lines cross, or a point where lines take sharp turns.
- Graphemes were grouped and labeled according to the number of nodes present and the way those nodes connect.



(a) (b) (c)

Figure 2: **(a)** graphemes with 4 nodes and connectedness code 112 (label *4_112*). **(b)** graphemes with 4 nodes and different connectedness codes. **(c)** a grapheme with 6 nodes.

- For 9 writers, 6 written paragraphs from the Computer Vision Lab (CVL) Database [3] were processed using Flash ID®. For each writer, 5 paragraphs were used for modeling and 1 was kept out for use as questioned writing.
- For a subset of graphemes, we calculated the relative frequency of occurrences for each writer (Table 1).

	Grapheme								
Writer	2_192	3_224	4_112	4_120	4_98	5_97	6_112	8_112	Total
writer #1	76	107	367	13	100	1	81	7	752
	0.10	**0.14**	**0.49**	**0.02**	**0.13**	**0.00**	**0.11**	**0.01**	1
writer #2	66	57	241	33	90	3	70	12	572
	0.12	**0.10**	**0.42**	**0.06**	**0.16**	**0.01**	**0.12**	**0.02**	1
⋮									
writer #9	89	72	899	40	51	0	417	82	1650
	0.05	**0.04**	**0.54**	**0.02**	**0.03**	**0.00**	**0.25**	**0.05**	1

Table 1: An example of the data. For a given writer, values in the top row are grapheme counts, and values in the bottom row are the relative frequencies of graphemes. Bold values serve as data for subsequent analysis.

3. Methods and Results

Step 1: Which graphemes are important?

- We used the *randomForest* package in R to gather graphemes in order of importance for predictive analysis of authorship.

Step 2: How many graphemes to include in analysis?

- A Bayesian hierarchical model (see handout) was successively fit to increasing subsets of graphemes, added in the order dictated by Step 1. We then calculated the posterior probability of origination from the known true author for each questioned writing.

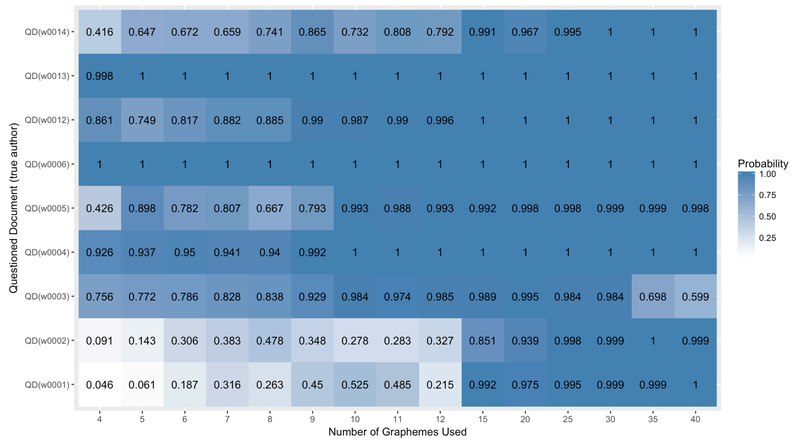


Figure 3: A plot of the posterior probability of origination from the true author for nine questioned documents by number of graphemes used in Bayesian hierarchical modeling.

- When 15 graphemes were used in modeling, over .50 probability was assigned to the true author for each questioned document.
- When the model does not assign most (or all) of the probability to the true author, we can investigate where it is accumulated by other authors. Probability assignments by the model trained with 10 graphemes are shown in Figure 4 below.

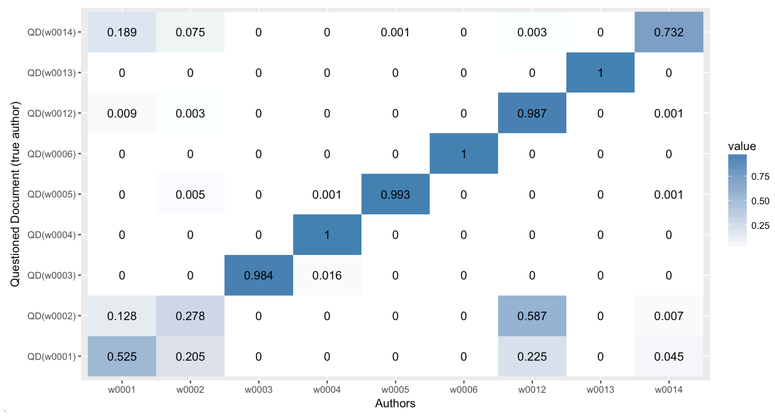


Figure 4: A plot of the posterior probability placed on the known authors for each of the questioned documents from the Bayesian hierarchical model with 10 graphemes.

4. Conclusions

Important Graphemes

- According to Step 1, the graphemes shown below are among the most important for predictive analysis.

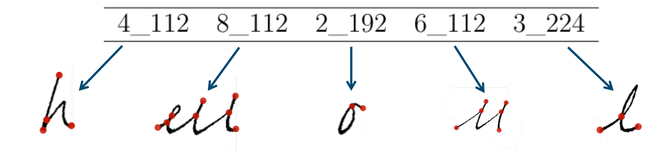
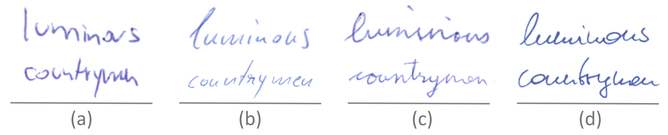


Figure 5: Five of the graphemes deemed most differentiating by the random forest model and examples of each.

- Letters with ascenders like "h" and "l", along with letters with loops like "e", "o", and the cursive "l" are considered to be informative by document examiners. Graphemes chosen as most differentiating by our random forest align nicely with these commonly suggested letter structures.
- The method is free from context of writing and human input, yet still arrives independently at important features and conclusions that align with those of examiners.

Writing Style

- Writers "0001", "0002" have print writing styles, while writers "0006" and "0013" have more connected cursive styles. The others fall somewhere in between.



(a) (b) (c) (d)

Figure 6: Writing samples from **(a)** writer "0001", **(b)** writer "0002", **(c)** writer "0006", and **(d)** writer "0013".

- Our analysis reflects the notion that it takes more information (graphemes in modeling) and is difficult to infer whether simplistic questioned writing falls within the natural variation of the true author, and out of others.

5. References

[1] Harrison, D., Burkes, T.M., Sieger, D.P.: *Handwriting Examination: Meeting the challenges of Science and the Law*, Forensic Science Communications 11(2009)
[2] FLASH ID®, Sciometrics LLC, Chantilly, VA, USA
[3] Florian Kleber, Stefan Fiel, Markus Diem and Robert Sablatnig, *CVL- Database: An Off-line Database for Writer Retrieval, Writer Identification and Word Spotting*, In Proc. of the 12th Int. Conference on Document Analysis and Recognition (ICDAR) 2013, pp. 560-564, 2013.

6. Acknowledgements

This work was partially funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement #70NANB15H176 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, University of California Irvine, and University of Virginia.