Developing a Handwritten Letter Database and Classifier for use in Forensic Document Analysis

Carley McConnell[1], Da'Monie McRae[2], Yolonda Miller[3], Samantha Tyner[4]

Iowa State University[1], Albany State University[2], Fayetteville State University[3], CSAFE[4]

csafe Center for Statistics and Applications in Forensic Evidence

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

Project Rationale and Goals

Courts often question the admissibility of handwritten documents as evidence, even when analyzed by handwriting experts. Analyzing handwriting involves a comprehensive comparative analysis between a questioned document and known handwriting of a suspected writer. Human errors can result in false positive or false negative identification of authorship. We want to use statistical analysis to reduce these errors.

Our main research goals are to create a database of handwritten letters and develop a classifier for identifying handwritten letters trained on our database.

Data

We used the Computer Vision Lab (CVL) Database [2] to retrieve a total of 162 documents, consisting of 6 handwritten English texts written by 27 different writers.

Using a Shiny App [3], we labeled each individual handwritten letter to build a database for training a statistical model to identify the letter.

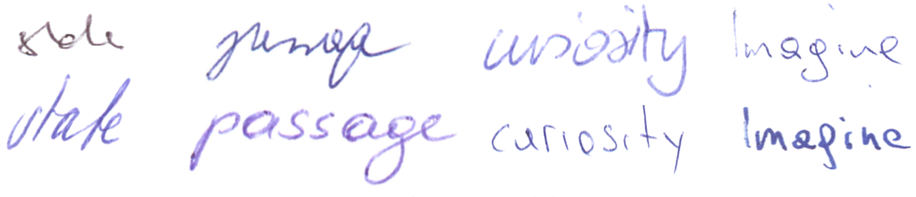

Figure 1: Variation of written words between writers

- We created a database consisting of 52,000+ classified letters, including all lowercase/ uppercase versions of each letter in the alphabet except for J, and combining letters such as P, U, V, W, Y, and Z.
- Next, we used this database in a statistical model to predict the accuracy of letter classification by a computer program.
- Data was split into testing and training sets, where the test set was 20% of the entire data set.

Methods and Results

First, we used Principal Component Analysis (PCA), a dimension-reduction tool, to reduce a large set of variables to a small set that still contains most of the information in the large set.

Next, the statistical computer program R [1] was used to classify the handwritten letters using random forest analysis. Random forests [5] are a supervised statistical learning method that uses image features to sort handwritten letter images into groups.

- 5000 trees were "grown" from the training data
- 200 features sampled for splitting at each "branch" of trees
- 63.2% of the training data and 26.3% of the testing data were accurately predicted

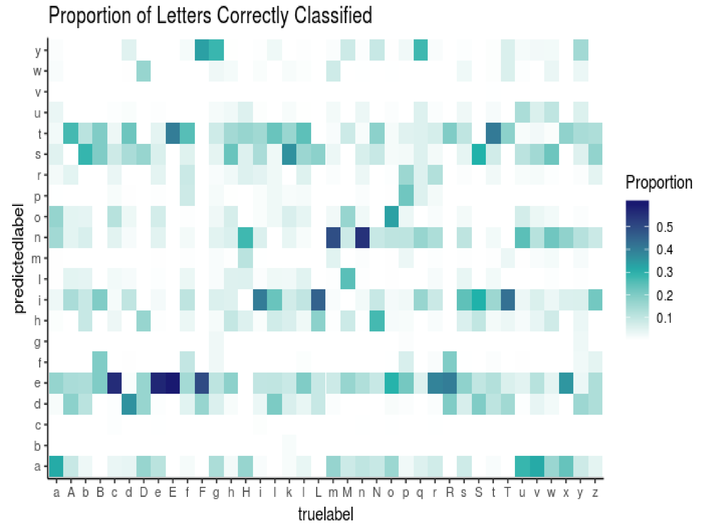

Figure 2: Random forest[5] analysis of true and predicted labels for the testing data set[4]

Conclusions

An extension of this study will include collecting more data for a larger database, considering factors such as:

- letters that were more frequent were better predicted
- complexity of letters when written in print vs. in cursive
- different letters that have the same characteristics when written by the same authors

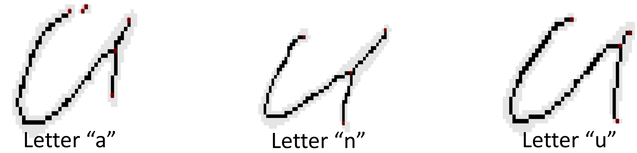

Figure 3: Different letters with similar form written by the same author

The ultimate goal of our project is to incorporate our classifier into a more complex statistical model to identify the author of a questioned document. Eventually, we hope that this project will be used to make the field of document examination more statistically sound.

References

1. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
2. Sablatnig, Robert. (2015 February 15) An Off-line Database for Writer Retrieval, Writer Identification, and Word Spotting. Retrieved from https://cvl.tuwien.ac.at/research/cvl-databases/an-off-line-database-for-writer-retrieval-writer-identification-and-word-spotting/
3. Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2018). shiny: Web Application Framework for R. R package version 1.1.0. https://CRAN.R-project.org/package=shiny
4. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
5. A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.
6. Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

Acknowledgements

We would like to thank:

- The REU students that contributed to data collection: Alese Brown, Badiah Hannon, and Malisha Jones.
- Dr. Samantha Tyner for her guidance in statistics.
- CSAFE Staff

IOWA STATE UNIVERSITY Albany State University University System of Georgia FAYETTEVILLE STATE UNIVERSITY

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.