Incorporating a Statistical Model into Forensic Shoeprint Analysis

James E. Kruse, Andrew Kimble III, Samuel Rew, Marion Gray-Lion,
Guillermo Basulto-Elias Ph.D., and Martin Silerio-Vazquez M.Sc
Iowa State University Research Experience for Undergraduates

Background

The primary mission of the Center for Statistical Application in Forensic Evidence (CSAFE) is to work on developing statistical models to be used in the interpretation of forensic evidence. Through the utilization of statistical tools such as random forests and sole finder we look to develop a model to identify shoe prints found at crime scenes.

We compute Hu moments and use them as unique summaries of each shoe sole. Hu moments are a set of eight summaries that are invariant to scale, rotation, and translation. In theory, when these points are incorporated into a model, they will be representative of the shoe itself.

Materials and Methods

Materials Used:

- Shoes
- R software
- Shoe Scanner
- Test Subjects
- R-packages

Methods:

- We used five different types of shoes with identical soles between two test subjects (G & M).
- Subjects stepped heel to toe on a scanner while shifting weight to ensure a uniform image
- Each shoe was scanned five times (100 scans total) to generate our data frame
- R software utilized to organize, visualize, and interpret data
- Scanned files were converted from the raw image to a negative image
- Performed preliminary analysis utilizing density plots and scatter plots
- The final model utilizes Hu moments as features for classification
- We built our model with the “randomForest” R package

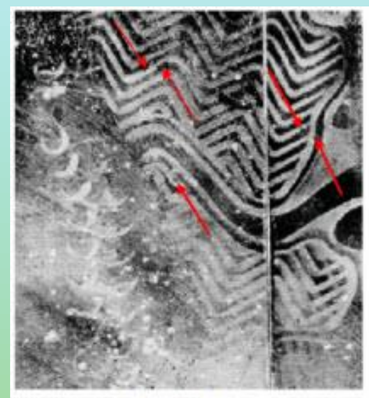

Figure 2(A)
Questioned Footwear Impression

Figure 2(B)
Known Shoe

Results: Data Visualization

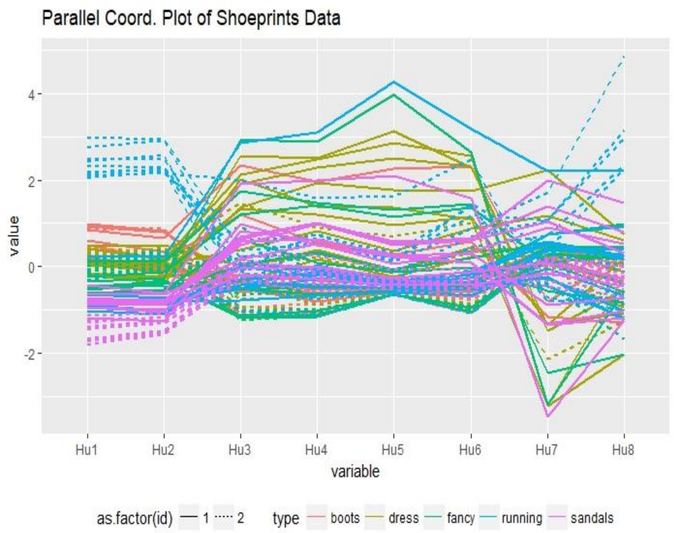

Fig. 1: Parallel Coordinate Plot: A separation of the shoes and test subjects.

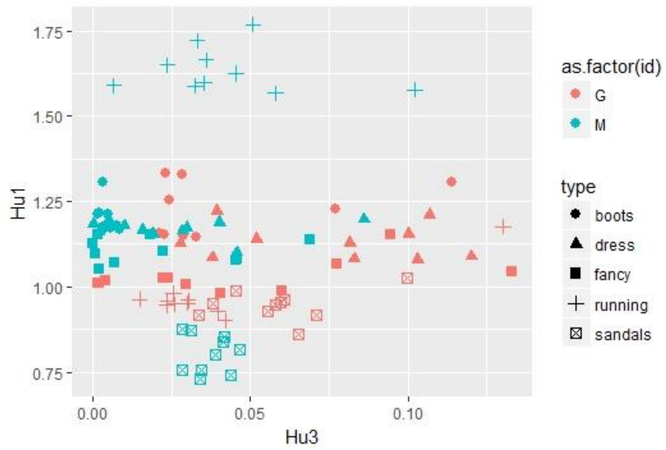

Fig. 2: Scatter Plot: Compilation(G&M): Different types of shoes between test subjects.

Results: Model

Interpreting the Data

Part I: Working with R Packages

- Images resulting from utilizing R software
- Similarities in data and distinguishing the differences
- Consistency in data indicates reproducibility of data collection

Part II: Random Forrest Model

- Identity determination between subjects (G & M)
- Train model: 60% of each test subject’s shoe data
- Test model: 40% of each test subject’s shoe data
- Accuracy of test model: 90% (correct identification – **Fig. 3**)

Pred.	G	M
G	17	1
M	3	19

Fig. 3: Model Prediction Table

Conclusion

In the future, we hope that our model can be applied to larger data sets while maintaining a high accuracy. The preliminary results lend support to the idea that statistical methods for forensic shoe print analysis can be a valuable tool when conducting forensic investigations. Going forward, we will begin to compile a larger data set (consisting of identical shoes for comparison) which we will then use to further test and train our model for verification and reproducibility.

References

- Hu, M. (1962). Visual pattern recognition by moment invariants. IEEE Transactions on Information Theory, 8(2), 179-187. doi:10.1109/tit.1962.1057692
- Lionel Henry and Hadley Wickham (2017). purrr: Functional Programming Tools. R package version 0.2.2.2. https://CRAN.R-project.org/package=purr
- Gregoire Pau, Florian Fuchs, Oleg Sklyar, Michael Boutros, and Wolfgang Huber (2010): EBImage - an R package for image processing with applications to cellular phenotypes. Bioinformatics, 26(7), pp. 979-981, 10.1093/bioinformatics/btq046
- Hadley Wickham, Romain Francois, Lionel Henry and Kirill Müller (2017). dplyr: A Grammar of Data Manipulation. R package version 0.7.0. https://CRAN.R-project.org/package=dplyr
- Barret Schloerke, Jason Crowley, Di Cook, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg and Joseph Larmarange (2017). GGally: Extension to ‘ggplot2’. R package version 1.3.1. https://CRAN.R-project.org/package=Ggally
- Carson Sievert, Chris Parmer, Toby Hocking, Scott Chamberlain, Karthik Ram, Marianne Corvellec and Pedro Despouy (2017). plotly: Create Interactive Web Graphics via ‘plotly.js’. https://plot.ly/r, https://cpsievert.github.io/plotly_book/, https://github.com/ropensci/plotly.
- Hadley Wickham (2017). tidyr: Easily Tidy Data with ‘spread()’ and ‘gather()’ Functions. R package version 0.6.3. https://CRAN.R-project.org/package=tidyr
- Guillermo Basulto-Elias and Martin Silerio-Vazquez (2017). solefinder: solefinder. R package version 0.0.1. https://github.com/CSAFE-ISU/solefinder
- Hadlay Wickham (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
- Smith, Michael B. 2009. “The Forensic Analysis of Footwear Impression Evidence.” Forensic Science Communications 11 (3). https://archives.fbi.gov/archives/about-us/lab/forensic-science-communications/fsc/july2009/review/2009_07_review02.htm.

Acknowledgements

We would like to thank:

- Guillermo Basulto-Elias Ph.D., Martin Silerio-Vazquez M.Sc, Sam Tyner M.Sc, and Joe Papio M.Sc for their guidance and assistance during this project.
- CSAFE and NIST for their financial contributions and for making this project possible.
- Iowa State University for the facilities and additional training given through the REU program.

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.