

A Clustering Method for Graphical Handwriting Components and Statistical Writership Analysis

Lead Researchers: Amy M. Crawford, Nicholas S. Berry and Alicia L. Carriquiry
Journal: Statistical Analysis and Data Mining | **Publication Date:** August 2020

OVERVIEW

Researchers developed and tested a statistical algorithm for analyzing the shapes made in handwriting to determine their source. Unlike other programs that analyze *what* words are written, this algorithm analyzes *how* the words are written.

THE GOALS

- 1 Develop a semi-automated process to examine and compare handwriting samples from questioned and reference documents.
- 2 Improve upon the existing methodology for determining writership.

APPROACH AND METHODOLOGY

Key Definitions:

Graphs: Simple structures with nodes and edges to represent shapes that constitute handwriting

Writership: The set of graphs a person typically makes when writing

K-means Algorithm: An iterative algorithm that separates data points into clusters based on nearest mean values

In this study, researchers scanned and analyzed 162 handwritten documents by 27 writers from the Computer Vision Lab database, a publicly available source of handwritten text samples, and broke down the writing into 52,541 unique graphs using the processing program *handwriter*. From there, a K-means algorithm clustered the graphs into 40 groups of similar graphs, each anchored by a mean or center graph. To allocate graphs to groups, researchers developed a new way to measure the distance between graphs.

Then, researchers processed an additional document from each of the 27 writers — held back as a “questioned document” — to test if the algorithm could accurately determine which document belonged to which writer. The new method for clustering graphs appears to be an improvement over the current approach based on adjacency grouping, which relies only on edge connectivity of graphs.

using adjacency clustering

above 50% probability of a match on 23 documents

using dynamic K-means clustering

above 90% probability of a match on 23 documents

correctly matched 26 documents



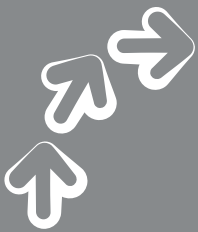
KEY TAKEAWAYS FOR PRACTITIONERS

- 1 The new approach shows promise, as it allows practitioners to more objectively analyze handwriting by studying the way letters and words are formed.
- 2 When compared to the more readily available but more volatile adjacency grouping method, the K-means clustering method contributed to greater accuracy when trying to identify the writer of a questioned document from among a closed set of potential writers.

FOCUS ON THE FUTURE

The new method favors certain properties of handwriting over others to assess similarities and can be extended to incorporate additional features.

The mean of a group of graphs is often a shape that does not actually occur in the document. Instead of centering groups using a mean graph, researchers are exploring whether using an exemplar graph as a group's anchor will simplify calculations.



NEXT STEPS

Try the *handwriter* algorithm by downloading it at forensicstats.org/tools/.

Investigate publicly available databases of handwritten documents:

- [CSAFE Handwriting database](#)
- [Computer Vision Lab database](#)

Additional relevant publications and presentations:

- [Statistical Analysis of Handwriting for Writer Identification \(2019\)](#)
- [Database of Handwriting Samples for applications in forensic statistics \(2020\)](#)
- [Statistical Analysis of Handwriting: Probabilistic Outcomes for Closed-set Writer Identification \(2020\)](#)

Explore more at forensicstats.org/insights/.

FUNDING



CSAFE is a publicly funded organization headquartered at Iowa State University. The National Institute of Standards and Technology (NIST) is one of the center's providers, supporting CSAFE as a nationally recognized Center of Excellence in Forensic Sciences, NIST Award # 70NANB15H176.

