J9 Statistical Characterization of Writers for Identification

Donald T. Gantz, PHD, George Mason University, Department of Applied Information Technology, MS 1G8, 4400 University Drive, Fairfax, VA 22030; John J. Miller, PhD, George Mason University, Department of Statistics, MSN 4A7, 4400 University Drive, Fairfax, VA 22030; Christopher P. Saunders, PhD, George Mason University, Document Forensics Lab, 4400 University Drive, MS, 1GB, Fairfax, VA 22030; Mark Lancaster, PhD, George Mason University, 4400 University Drive, MS 1G8, Fairfax, VA 22030; and JoAnn Buscaglia, PhD, Federal Bureau of Investigation Laboratory, Counterterrorism and Forensic Science Research Unit,, Building 12, Quantico, VA 22135*

3

After viewing and discussing this presentation, attendees will be familiar with the statistical techniques behind a writer identification tech- nology that is based on an innovative quantification of handwritten text and computationally intense statistical methods. Attendees will learn how a writer is characterized through quantitative analysis of a writing sample and they will see how a data base of writers is scored and sorted to identify the writer of a questioned document from among the population of writers in a data base. An oral presentation at this conference provides a high level summary of the biometric identification results that have been attained by applying various statistical algorithms to the quantification of handwriting. This presentation provides lower level details showing how statistical methods are used to characterize writers.

This presentation will impact the forensic community by making attendees aware of the mathematical quantities that underpin the biometric identification with handwriting and the efforts of the three cooperating organizations to assist forensic document examiners (FDEs) in exploiting these quantities to support their practice. These organizations are Gannon Technologies, Inc., the Document Forensics Laboratory at George Mason University and the FBI Laboratory.

Other oral and poster presentations at this Conference address the hand- writing quantification technology developed by Gannon Technologies, Inc. A recognized character in a document is associated with a mathematical graph, which is an array of curves that intersect or end in vertices. The frequency pattern of graph types observed in a document for each separate alphabetic character is a very powerful biometric identifier of the writer of the document. However, this poster presents statistical details for the minutia level biometric identification, which is based on physical feature measure- ments on the graph associated with a character.

There are hundreds of separate feature measurements tied to even the simplest graph with only a few curves and vertices. Reducing the very large number of feature measurements down to a very few that characterize the way the writer writes the given character is a challenge that statisticians call the curse of dimensionality. Two approaches have been taken to meet this challenge. The first approach that was implemented performed separate head-to-head statistical comparisons between the writer’s sample and a sample from each writer in a test bed data base. For each test bed writer, a discriminant analysis procedure selects a subset of feature measurements that separate the samples of the two writers. The entire procedure is repeated starting with a reduced set of feature measurements that were broadly selected to discriminate the writer from other writers in the test bed. Several repetitions reduce the feature measurements to ten or fewer, called a biometric kernel, which characterize the writer for the particular character and graph pair. The full characterization of the writer is the full collection of biometric kernels corresponding to each character-graph pair that was observed in the writer’s sample documents.

A second approach focuses on the importance of a single feature meas- urement on a character-graph pair for a particular writer. For that writer and single feature measurement, a t-test is performed against each other writer in a test bed, and the p-values from the t-tests are recorded. After transforming and combining these p-values, a decision is made on the importance of this single feature measurement in distinguishing the particular writer from all other writers in the test bed. Next, information is combined from all feature measurements on the same character. For the particular writer, each decision made on the importance of a feature measurement for a character-graph pair gives rise to a p-value. Techniques used in the analysis of microarray data are especially suited for determining the weight that should be placed on each feature measurement for a character-graph pair. It is this collection of weights that determines the biometric kernel that characterizes the writer for the particular character-graph pair.

Based on minutia level biometric kernel-based identification alone, the true writer of a questioned document will be retrieved from a data base of known writers with high accuracy if that writer’s minutia characterization is stored in the data base.

Handwriting, Statistics, Biometrics