J5 New Results for Addressing the Open Set Problem in Automated Handwriting Identification

Donald T. Gantz, PhD, George Mason University, Department of Applied Information Technology, Mail Stop 1G8, 4400 University Drive, Fairfax, VA 22030; John J. Miller, PhD*, George Mason University, Department of Statistics, Mail Stop 4A7, 4400 University Drive, Fairfax, VA 22030; Christopher P. Saunders, PhD, George Mason University, Document Forsenics Lab, 4400 University Drive, Mail Stop, 1GB, Fairfax, VA 22030; Mark A. Walch, MPH, The Gannon Technologies Group, 7600 Coleshire Drive, Suite 600, McLean, VA 22102; and JoAnn Buscaglia, PhD, FBI Laboratory, CFSRU, FBI Academy, Building 12, Quantico, VA 22135*

After attending this presentation, attendees will be updated on the current development status of a powerful tool for automated open set handwriting identification. Forensic document examiners will be more aware of the handwriting identification tools that are being developed to assist them in their practice.

The presentation will impact the forensic science community by increasing awareness of substantial advances being made in adding scientific underpinnings to the practice of forensic document examination.

The Open Set Problem involves making a two-stage decision when attempting to ascertain whether a questioned document was written by some individual in a reference collection (for which training material exists for each writer in the reference collection). The first step is to decide whether the document was written by any writer in the reference collection and the second step is to decide which writer in the reference collection is the most likely writer of the questioned document (or to give a "short list" of likely writers), presuming that the decision is that some writer in the reference collection was the writer of the questioned document. At the AAFS 2009 Annual Meeting, results were presented for this problem that were generated using the FLASH ID software system. Those results used the difference between the aggregated score (totaled over all graphemes in the questioned document) for the first place writer and the aggregated score for the second place writer as the basis for the "in the reference collection" decision. In this paper, results will be given for an improved open set decision based on a combination of the original criterion with a new criterion based on a "Vector of Counts" (VOC) methodology described below.

The VOC methodology is a way to obtain categorical type feature data by using the FLASH ID system with continuous feature data. It works in the following manner. First, a "base set" of writers is obtained, who are not in the reference collection or likely to be among writers of any questioned documents we observe. Writing samples, from these individuals and using FLASH ID create a trained system of the same sort as is used for the reference collection. This base set is used to analyze any document by recording for each grapheme in that document, which writer in the base set is most likely to have written that grapheme. In this way, a vector of counts for the document can be developed by counting how many graphemes are assigned to each writer in the base set.

Next, the training writings is taken for each writer in the reference collection and obtain a VOC for each of those writers. When a questioned document is analyzed, its VOC is obtained as well. Then, the VOC can be compared for the questioned document with the VOC for the first place writer when writers in the reference collection are assigned questioned document scores by FLASH ID. One way to do this comparison of VOCs is using a chi-squared statistic. Since large values of chi-squared would indicate a relative mismatch between the questioned document and the first place writer and since small values of the previously used difference of first and second place writer scores would also indicate a poor match, taking the ratio of these two criteria can be an effective tool for improvement of the open set decision. Numerical results are given based on extensive simulations to illustrate the improvement.

Handwriting Identification, Open Set Problem, Vector of Counts

 ** Presenting Author*