Proceedings of Student-Faculty Research Day, CSIS, Pace University, May 2nd, 2008

Longitudinal Keystroke Biometric Studies on Long-Text Input

Elizabeth Wood, Julio Zelaya, Eric Saari, Kenneth King, Mike Gupta, Nicola Howard, Sadia Ismat, Mary Angela Kane, Mark Naumowicz, Daniel Varela, Mary Villani
Seidenberg School of CSIS, Pace University, White Plains, NY, 10606, USA
(eawood@med.cornell.edu, jz16811n@pace.edu, es71900n@pace.edu, dr_klove@hotmail.com, mikegupta@gmail.com, howard.nicola@gmail.com, si76337n@pace.edu, kodie@verizon.net, mark@bkln.net, dv77210n@pace.edu, villanmv@farmingdale.edu)

Abstract

Pace University's keystroke biometric system is able to identify individuals with a high degree of accuracy. A relatively new dichotomy model system has been developed to support studies of keystroke biometric authentication, and preliminary results with this system have shown promise. The current study extends those authentication experiments and examines the consistency of an individual's keystroke patterns over periods of week and years, expanding upon a small preliminary longitudinal study carried out on four subjects in the fall of 2007. Results of the preliminary study suggest that an individual's typing patterns do not change significantly over a period of 2 to 4 weeks. The current study repeats the fall 2007 experiments with a larger sample size, and begins to examine the consistency of typing patterns over a longer period of approximately two years. The study also provides new data on keystroke patterns in a set of identical twins, shedding light on the extent to which keystroke patterns may be biologically determined.

1. Introduction

Biometric systems deal with the task of establishing the distinctiveness of an individual in a population based on a set of biometric measurements that are unique to an individual. Applications of biometrics include both *identification* and *authentication*. Identification systems attempt to identify an individual from a group. Authentication systems attempt to verify whether an individual is who he or she claims to be.

Biometrics can be roughly divided into physiological biometrics (such as fingerprint and iris) and behavioral biometrics (such as keystroke patterns, mouse movements, and signature) [1], although certain biometrics may be both behaviorally and physiologically determined. A person's physiological characteristics tend to change little over time, whereas behavioral characteristics are more subject to change based on testing conditions. As such, physiological measurements are more likely than behavioral measurements to be similar or identical when measured at different points in time in the same individual.

When used in biometric systems, a disadvantage of physiological measurements is that they typically require special equipment for accurate measurement. Keystroke biometrics is the field of biometrics that deals with an individual's unique typing patterns. A distinct advantage of keystroke biometrics is the ubiquity of keyboards in human-computer interactions. With the growing use of Web-based authentication [2], special equipment becomes impractical is it is difficult to predict a user's location at the time of identification or authentication.

Among biometrics systems, keystroke biometrics is distinct in its potential to address both *multi-modal verification* (also addressed by many other biometrics) and *continuous monitoring.* Multi-modal verification deals with verification on multiple levels (such as a password plus a biometric); continuous monitoring deals with the need to ensure that, once a user is authenticated, some else does not "hijack" the user's session. For keystroke-based sessions (i.e., when a user is communicating with a computer system via the keyboard), the keystroke biometric is an ideal and natural way to perform continuous monitoring without requiring any explicit response from the user.

Studies in keystroke biometrics have been ongoing at Pace University since 2004 and include studies in both identification and authentication. The studies are faculty-initiated and have been carried out by graduate students in the Master's and Doctoral programs in the Department of Computer Science and Information Systems.

Using the Pace system, prior studies have established a high-level of accuracy in identifying individuals from a group. Recent studies, using the dichotomy model, have shown potential in the area of authentication. Relatively few studies, however, have looked at consistency of an individual's keystroke patterns over periods of weeks, months or years.

Preliminary longitudinal studies carried out using the Pace system in the Fall of 2007 examined identification and authentication accuracy when an individual is re-tested at intervals of approximately two- and four-weeks [3]. These studies showed average identification accuracies of 95% and 98% over two- and four-week intervals, respectively. Average authentication accuracies were about 94% for the two-week interval and 95% for the four-week interval. Although these results are promising, the sample size for this study was quite small (four subjects). Therefore, the accuracy obtained in these studies may not be representative of a real-world application where the larger pool of potential matches might increase the likelihood of incorrect matches.

In addition to the small sample size, a limitation of the Fall 2007 study is the relatively short period of time over which keystroke patterns were studied in the same individuals. While a short-term study might provide good insight into the impact of day-to-day changes due to factors

such as mood, stress-level, or sleep, different factors are likely to come into play over longer periods of time. The most obvious of these are age-related changes, changes resulting from major shifts in life circumstances, and even personality changes (which are more likely over a long period of time than over a period of weeks). Other possible long-term changes include those resulting from new health-related issues or disabilities.

The studies described in this paper expand upon the Fall, 2007 studies by pooling the Fall, 2007 data with new data on an additional 9 subjects also collected at two- and four-week intervals. In addition, we begin to explore the consistency of typing patterns over longer periods of time by comparing data on eight subjects who participated in a 2006 study to new data entered by the same subjects two years later. Finally, we present identification accuracy data on a set of identical twins studied over a short period of time in order to shed light on the impact of biology on keystroke patterns.

2. Authentication and the Dichotomy Model

In biometrics, there are two important models that can be used for establishing the individuality of a person: identification (polychotomy, one-of-many decision) and authentication or verification (dichotomy, binary decision). In identification applications, a user is identified from within a population of, say, n users (one-of-n response), and is usually considered to be the more difficult of the two problems. In contrast, authentication involves the process whereby an individual is verified as being the person he/she claims to be or not. Therefore, in authentication applications a user is either accepted or rejected, that is, the output is a binary response, yes or no. It has been argued that the authentication (verification) problem is more suitable than the identification model for establishing the individuality of a person, specifically in cases where the number of classes is too large to be completely observed [5], for example, the population of an entire nation.

In order to establish the distinctiveness of an individual, i.e., to authenticate that individual, we start by transforming the many-class problem into a dichotomy. This is done by using a distance measure between two samples of the same class and between samples of two different classes. The usefulness of this model lies in that it allows for the inferential classification of individuals without the need to sample all the classes. Also, it provides reliability when performing classification of all classes based on the data obtained from a small sample of classes representing the whole population. Using this model, all sample pairs are categorized as either the same class or different class. Given two biometric data samples, the distance between the two samples is first computed. This distance measure is used as data to be classified as positive (intra-variation, within person or identity) or negative (inter-variation, between different people or non-identity) [5].

In addition to building upon preliminary longitudinal studies, the current work further tests the feasibility of the dichotomy model as a foundation for keystroke biometric authentication. Although the use of this model in biometric authentication is fairly new, it has great potential and has proven to be a powerful technique in iris authentication [5][6].

3. Keystroke Biometric System

The keystroke biometric system at Pace consists of three separate interoperable modules: one for entry of raw keystroke data by the research subject, one for feature extraction, and one for pattern classification. The raw data-entry module utilizes a Java applet that collects a long-text sample (approximately 600-700 keystrokes) from the subject and stores it in a text file that is named according to the subject's name, the keyboard type (desktop or laptop) and typing task (copy operation or free-text). The data stored in the text file includes the actual keystrokes entered by the subject, the amount of time each key was pressed (duration), and the elapsed time between adjacent keystrokes (latency).

The feature extraction module inputs a set of raw data files and converts them to a "feature file". In the feature file, the raw data is collapsed into a set of 239 "features" [7]. These features includes measures such as average duration of selected letters, average duration of all left-hand letters, duration of specific letter-to-letter transitions, percentage of keystrokes that are left mouse clicks, and overall input rate. The feature files are then fed into a pattern classifier to yield accuracy results for identification or authentication.

For pattern classification, there are two systems, one based on an identification model and used for identification studies, and another based on the dichotomy model and used for authentication studies. Each classifier has a corresponding version of the feature extractor. The specific version used for a given experiment depends upon the classifier to be used, which is determined by the nature of the experiment (identification vs. authentication). The two versions of the feature extractor operate identically, but output data in different formats in accordance with the two classification systems (identification model based system and dichotomy model-based system).

3.1 Raw Data Collection

The raw data collection system consists of a set of HTML/PHP screens and a Java applet that together walk subjects through the process of registering (new users) and entering keystroke samples (registered users). When a new subject enters the system, s/he is asked to register by providing his/her name and a small set of additional information relevant to the keystroke entry task (type of computer, whether the individual is right- or left-handed, person's typing method and speed, etc.).

Once a subject has registered, he or she can proceed with entering keystroke samples. Samples can be entered in one sitting, or the subject can come and go until all required samples have been entered. Subjects are asked to provide all samples for a given "complete data set" within a two-day period. A *complete data set* consists of five samples in each of the following four "quadrants":

Laptop Free-Text	Laptop Copy
Desktop Free-Text	Desktop Copy

For longitudinal studies, each subject is asked to provide multiple complete data sets at specified intervals.

Before entering a sample, the user is prompted to indicate the current keyboard type (desktop or laptop) and typing operation (free-text or "copy fable"). The "Copy fable" task consists of copying a paragraph that tells a well-known fable about the bat and the weasels. In the free-text tasks, the user is asked to select from a set of 10 topics, and write freely on the topic. The user is required to type a minimum of 677 characters for each free-text sample, and is permitted to enter up to 1000 characters. The Java applet, which inputs the text typed by the user, maintains a running counter of keys pressed and will not permit the sample to be submitted until the minimum number of keystrokes have occurred. The keystroke entry form is depicted in Figure 1, below.

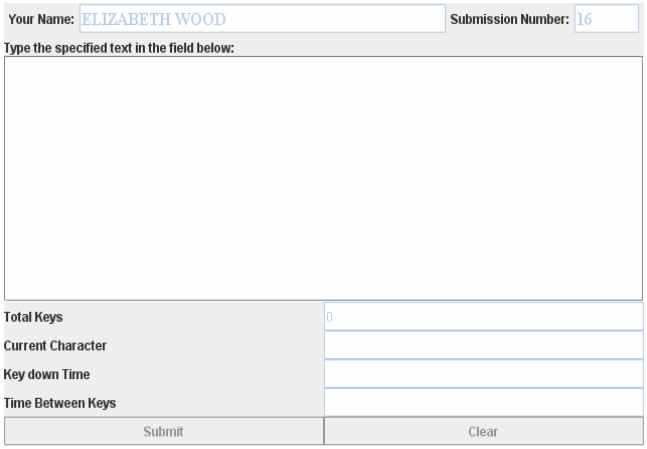

Figure 1. Keystroke entry screen (Java applet).

3.2 Feature Extraction

Feature extraction utilizes one of two Java-based feature extractors developed by past graduate student teams at Pace. The old version of the feature extractor, which is compatible with the identification-model-based pattern classifier, is used for identification studies. A newer version is used for dichotomy-model-based authentication studies (see [8]). The format of the resulting feature file differs depending upon the specific feature extractor version used, but both versions produce summary information reflecting the 239 features used in pattern classification. The feature files are then fed into the appropriate pattern classifier for identification and authentication studies.

In order to optimize pattern classification, the feature extractors can perform several compensatory operations on the raw data before producing feature files. These include the removal of outliers (which could occur, for example, if a subject was interrupted during a typing task, resulting in a long delay between two successive keystrokes) and fallback. Fallback is used within the feature extractor when there is an insufficient amount of data about a particular feature. It involves the substitution of other related data for the missing data. Two fallback methods have been studied at Pace and are available in the Pace systems: linguistic fallback and touch-type fallback. The first performs fallback based on grammatical structures, and the second based on geography of keys on a standard computer keyboard. Studies have shown the linguistic fallback method to be superior to the touch-type method when used within the Pace systems. Therefore, linguistic fallback is used throughout the current studies.

A screen shot of the feature extractor is provided below. The settings shown for items 2 through 8 are those used for the current study.

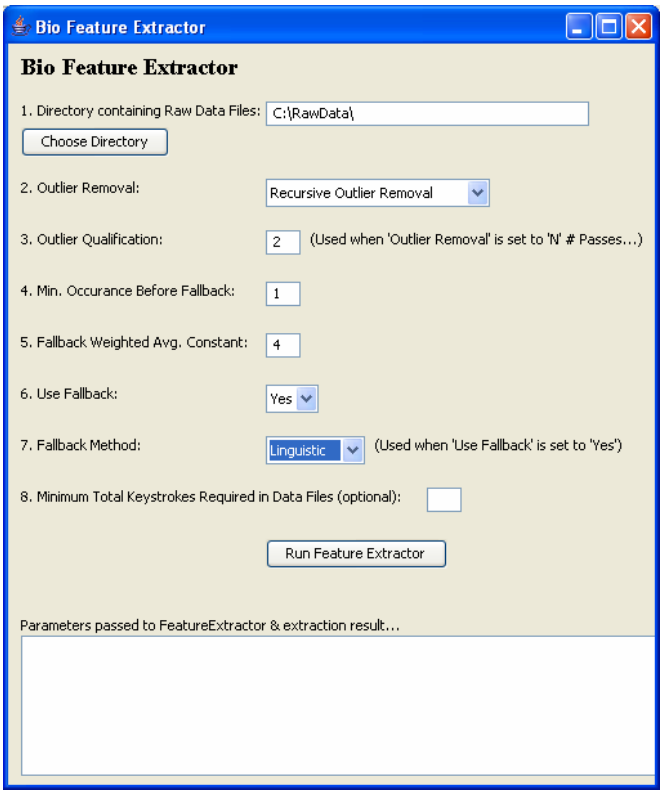

Figure 2. Feature extractor

3.3 Pattern Classification

The pattern classification systems at Pace use a train-test model in which a classifier is "trained" on one set of data and tested with another. This can be done in one of two ways. The first is the "leave one out" approach, in which a single features file is used. In this approach, one sample is pulled out of the features file, the system is trained on the remaining samples, and testing occurs on the sample that was pulled out. The test is successful if the best match is with another sample from the same subject. In the second approach, "train-on-one/test-on-another", two features files are used. The system trains on one file, and each sample in the second file is tested against the training file. The test is successful for a given sample if the best match is with another sample from the same subject. The train-on-one/test-on-another approach is appropriate when samples are collected from the same subjects under different conditions, as in the current study where data are collected at different time points.

3.3.1 Identification Classifier

The classifier used for keystroke biometric identification studies at Pace University is a Java application with a user-

interface shown in Figure 3. In the current study, item 1 (fallback) was set to "linguistic" (to match the method used by the feature extractor) and item 9 (classification method) was set to "train & test", which requires two input files and performs a "train on one/test on the other" analysis of the two feature files. Different pairs of feature files were provided to the classifier (items 12 and 13) and the classifier was run to obtain testing accuracy between each pair of feature files.

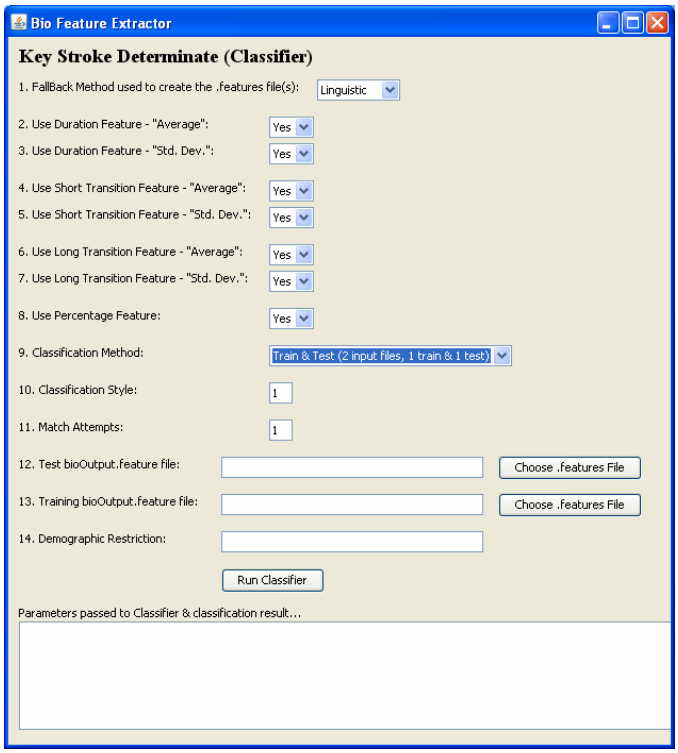

Figure 3. Identification classifier

3.3.2 Authentication Classifier

The Biometric Authentication System, which includes the dichotomy model-based classifier, consists of two Java-based components: a simple standalone GUI and dichotomy transformation utilities.

3.3.2.1 The GUI

The GUI provides a simple interface to facilitate authentication. The user specifies training and testing feature files for an experiment, then clicks "Apply Dichotomy Model" to perform the dichotomy transformation (Figure 4).

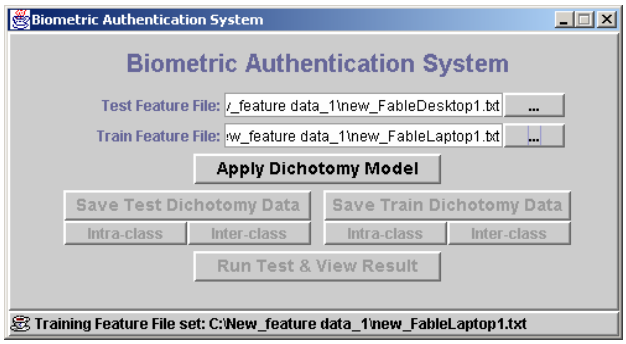

Figure 4. The GUI of the dichotomy model-based classifier

The user can choose the maximum number of inter- or intra- class samples to create, thereby allowing for a reduced set of randomly selected inter-class samples being used for experimentation, as the number of potential inter-class samples tends to be very high (Figure 5). The Pace work done in the fall of 2007 set the maximum number of inter- or intra-class sizes to 500. In the current work, this number was increased to 1000.

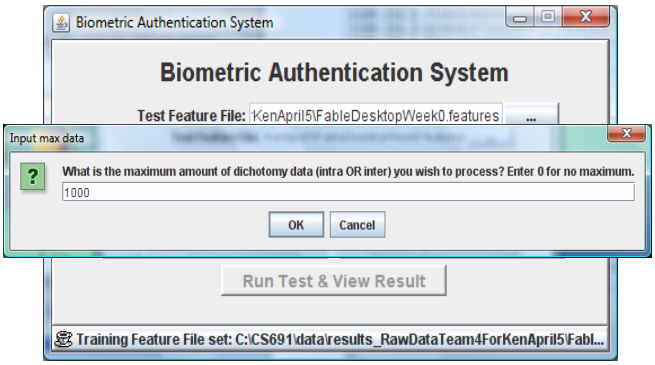

Figure 5. Choose maximum size of intra/inter-class data

The user can then save the testing and training dichotomy data in individual files either as a combination of intra and inter class samples or as separate intra- and inter-class files (Figure 6). A sample dichotomy file is shown in Figure 7.

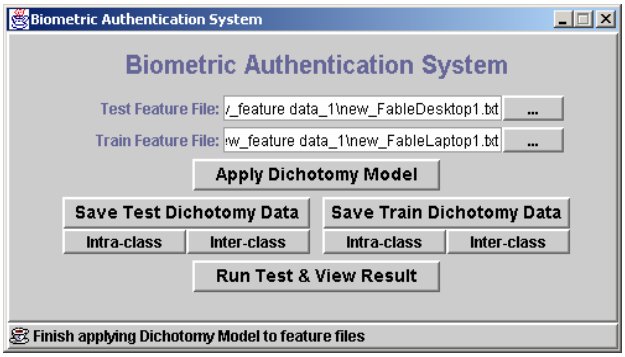

Figure 6. Save dichotomy data, run test, and view result

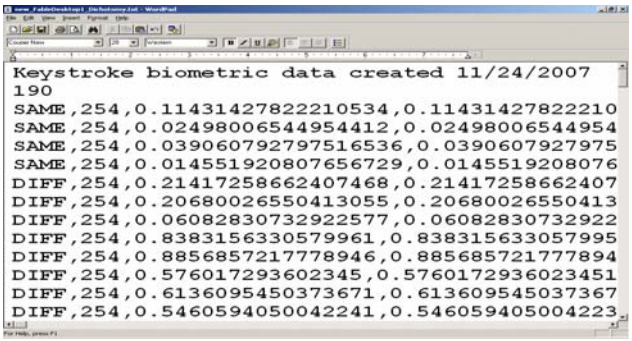

Figure 7. Dichotomy data file

Sample experimental results are shown in Figure 8. These can be saved to an .html file for future use.

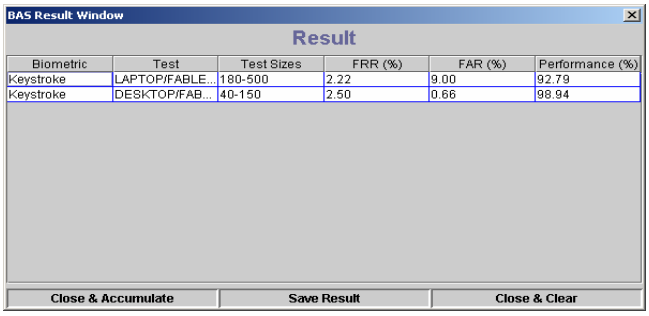

Figure 8. View results

The system maintains a log file, which records all system activities. This log file can be viewed by clicking on the status bar of the main window, and cleared by clicking on the "Clear Log" option at the bottom of the screen. A sample log file is shown in Figure 9, below.

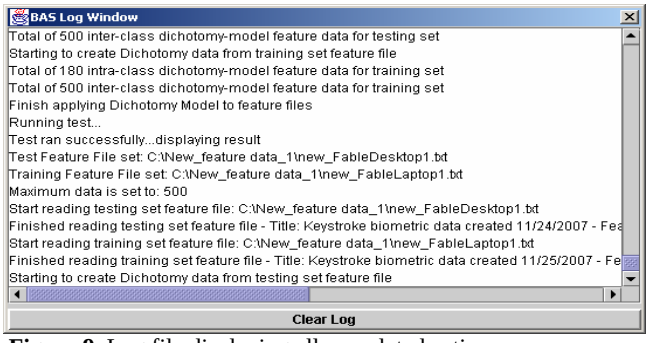

Figure 9. Log file displaying all completed actions.

3.3.2.2 The Dichotomy Transformation Utilities

The dichotomy utilities include the functionality required for the dichotomy transformation and the verification of the individuals. A detailed description of the dichotomy model transformation and underlying theory can be found in [8].

4. Experimental Results

The primary focus of the current work is on the accuracy of identification and authentication systems when samples are obtained from the same individuals at different points in time. The first study looks at accuracy when three data sets are collected from the same individuals at approximately two- and four-week intervals. The second looks at accuracy over a two-year period in a different set of subjects.

The specific experiments performed, and results obtained, are described below. All data obtained were divided into training and testing sets as required by the individual experiments. For authentication experiments, the dichotomy conversions were performed. Finally, classification experiments were conducted using a "train-on-one/test-on-another" approach where the training and testing data sets represented keystroke data collected on the same individuals at the same and different points in time.

4.2 Same Subjects at 2- and 4-Week Intervals

The 2- and 4-week studies extend preliminary work carried out in the fall of 2007 in which the same 4 subjects entered keystroke data at weeks 0, 2, and 4 [3]. In the current work, we improve the power of these studies by collecting new data on an additional 9 subjects and pooling the new data with that of the original four subjects for a total of 13 subjects. The results of these studies are presented in the Tables 1 and 2.

For same-week studies, we performed experiments under different conditions (different keyboard, different typing task, or both) both within a time point and across time points. For "same-week" studies, identification and authentication results were obtained at week 0 (W0-W0), week 2 (W2-W2) and week 4 (W4-W4). These results were then combined for overall same-week accuracies. (Note that the authentication table shows only combined results, but the same process was used to obtain these results.) For two-week intervals, results were obtained by training on week 0 and testing on week 2 (W0-W2) and by training on week 2 and testing on week 4 (W2-W4). These results were similarly combined for overall two-week identification and authentication accuracies. In Table 1 a downward trend can be observed over time.

Experiment	Same Week				Two-Week Interval			Four-Week Interval
	W0 W0	W2 W2	W4 W4	Avg	W0 W2	W2 W4	Avg	W0 W4
LapCopy/ DeskCopy	98.5	100	92.3	96.9	96.9	98.5	97.7	87.7
DeskCopy/ LapCopy	92.3	93.9	84.4	90.2	90.8	87.5	89.2	87.5
LapFree/ DeskFree	98.5	96.9	93.8	96.4	95.4	87.7	91.6	87.7
DeskFree/ LapFree	86.2	84.6	86.2	85.7	87.7	70.7	79.2	76.9
DeskFree/ DeskCopy	98.5	95.4	98.5	97.5	89.2	86.2	87.7	90.8
DeskCopy/ DeskFree	90.7	92.3	100	94.3	69.2	86.2	77.7	78.5
LapFree/ LapCopy	90.8	96.9	87.5	91.7	93.8	84.4	89.1	89.1
LapCopy/ LapFree	96.9	95.4	90.8	94.3	92.3	93.8	93.1	90.8
LapFree/ DeskCopy	92.3	87.7	89.2	89.7	81.5	78.5	80	83.1
DeskCopy/ LapFree	87.7	78.5	78.5	81.7	75.4	73.8	74.2	70.7
DeskFree/ LapCopy	69.2	80	75	74.7	83.1	71.9	77.5	75
LapCopy/ DeskFree	81.5	96.9	76.9	85.1	83.1	89.2	86.2	81.5
Average				**90**			**85**	**83**

Table 1. Identification accuracy on thirteen subjects studied at week 0, week 2 and week 4 (same subjects, different conditions, same or different time points). Five samples were collected from each subject in each quadrant, for a total of 65 samples in each test file (with the exception of the 4-week laptop copy file, which was missing one sample for a total of 64 samples). Percentages shown are percent of the 65 samples from which the subject was correctly identified in the associated training file.

Authentication (Table 2) was only slightly more successful than identification at two- and four-week intervals, and also showed a downward trend over the four-week period. Moreover, all authentication accuracies were lower than in the prior longitudinal study, perhaps due to the increase in sample size from 4 (prior study) to 13 (current).

Experiment	Train-Test					
	Same Week		Two-Week Interval		Four-Week Interval	
	FAR/ FRR	Acc	FAR/ FRR	Acc	FAR/ FRR	Acc
DeskCopy/ LapCopy	2.05/3.30	96.8	2.31/4.15	96.1	6.92/1.79	97.6
LapCopy/ DeskCopy	6.74/25.13	77.0	5.77/33.35	69.8	4.61/35.60	68.0
DeskFree/ LapFree	5.64/7.00	93.2	2.31/7.75	92.9	10.00/9.60	90.4
LapFree/ DeskFree	7.18/6.77	93.2	1.92/10.15	90.8	1.53/21.20	81.1
DeskCopy/ DeskFree	2.82/4.07	96.1	2.96/4.30	95.9	3.84/4.39	95.7
DeskFree/ DeskCopy	3.85/7.57	92.9	3.08/10.75	90.1	3.07/16.10	85.4
LapCopy/ LapFree	9.59/7.97	83.0	6.92/24.10	77.9	5.38/34.90	68.5
LapFree/ LapCopy	4.10/9.27	91.3	3.08/9.40	91.3	2.30/10.5	90.4
DeskCopy/ LapFree	4.36/6.57	93.7	5.00/7.85	92.5	2.30/7.10	93.5
Lap Free/ DeskCopy	3.33/4.23	87.0	3.46/15.6	85.8	2.30/26.40	76.4
LapCopy/ DeskFree	6.48/9.73	81.8	5.77/30.05	72.7	7.14/21.60	80.0
DeskFree/ LapCopy	3.33/9.33	91.4	3.08/6.95	93.5	7.69/2.90	96.5
Average	5.0/10.9	**89.8**	3.8/13.7	**87.4**	4.6/16.0	**85.3**

Table 2. Authentication accuracy on 13 subjects studied at weeks 0, 2 and 4. As with identification studies, each feature file contained 5 samples from each of 13 subjects, for 65 samples in all (with one exception noted in legend for Table 1).

4.1 Same Subjects at a Two-Year Interval

In order to study the accuracy of identification and authentication systems over longer periods of time, we contacted each of 36 subjects who had participated in early identification experiments carried out by Villani et al. in 2006 [9]. Subjects were asked to return to the Web-based raw data collection system and enter a new complete data set (5 samples in each of the four quadrants).

Not all subjects were able to respond to our request. However, we were successful in obtaining complete new data sets from 8 of these individuals.

As with the two- and four-week studies, "same year" data were obtained using both Y0-Y0 and Y2-Y2 experiments. The results of these individual experiments were then combined for overall accuracies. The Y0-Y0 and Y2-Y2 data are shown explicitly in the identification table (Table 3) and (due to space constraints) only combined results are shown in the authentication table (Table 4).

Experiment	Same Year			Y0-Y2
	Y0-Y0	Y2-Y2	Combined	
LapCopy/DeskCopy	75	77.5	76.3	65
DeskCopy/LapCopy	90	85	87.5	67.5
LapFree/DeskFree	75	77.5	76.3	80
DeskFree/LapFree	82.5	77.5	80	80
DeskFree/DeskCopy	95	90	92.5	75
DeskCopy/DeskFree	95	100	97.5	57.5
LapFree/LapCopy	100	100	100	57.5
LapCopy/LapFree	100	97.5	98.8	60
LapFree/DeskCopy	65	80	72.5	52.5
DeskCopy/LapFree	80	80	80	65
DeskFree/LapCopy	77.5	75	76.3	57.5
LapCopy/DeskFree	75	75	75	82.5
Average			**84.4**	**66.7**

Table 3. Identification accuracy on eight subjects studied in 2006 and again in 2008. Each subject submitted five samples in each of four quadrants, for a total of 40 samples in each test file. The table shows percent of the 40 samples that were accurately identified when run against the specified training file.

As shown in Table 3, identification accuracies across the two-year interval range from 52.5% to 80%, with an overall average of 66.7%. If the data were entirely random, with eight subjects represented in the data, one would expect about one in 8 (or 12.5%) of identification attempts to be successful by chance. Therefore, the results indicate that individuals remain identifiable by their keystroke patterns after a two-year interval, but at a lower level of accuracy than at a 2-week or 4-week interval.

Authentication results (Table 4) are somewhat better, with an average accuracy of almost 92% with a two-year interval between the train and test sets. This is slightly lower than previous authentication results on four subjects studied over a two-week interval (about 94%) and four-week interval (about 95%) [8], but is substantially better than authentication accuracies obtained at two- and four-weeks in the current study. The reason for this discrepancy may be the smaller sample size in the two-year study.

Experiment	Same Year		Two-Year Interval	
	FAR/FRR	Acc	FAR/FRR	Acc
DeskCopy/LapCopy	5.00 / 8.29	92.1	5.00 / 9.71	90.8
LapCopy/DeskCopy	1.25 / 4.14	96.2	0.00 / 8.42	92.4
DeskFree/LapFree	0.64 / 8.33	92.4	3.94 / 5.82	94.4
LapFree/DeskFree	1.56 / 2.21	97.7	2.50 / 3.42	96.7
DeskCopy/DeskFree	8.75 / 7.50	92.4	10.00/12.00	88.2
DeskFree/DeskCopy	5.13 / 10.61	89.9	2.63 / 9.37	91.3
LapCopy/LapFree	0.00 / 2.86	97.4	0.00 / 3.71	96.7
LapFree/LapCopy	1.25 / 3.29	96.9	0.00 / 7.28	93.5
DeskCopy/LapFree	5.63 / 10.29	90.2	6.25 / 12.85	87.8
Lap Free/DeskCopy	3.13 / 8.86	91.7	2.50 / 9.00	91.7
LapCopy/DeskFree	5.00 / 5.00	95.0	1.25 / 6.57	94.0
DeskFree/LapCopy	0.00 / 10.40	90.6	1.31 / 13.49	87.7
Average	3.11 / 6.82	**93.5**	2.95 / 8.47	**92.1**

Table 4. Authentication accuracy on eight subjects studied in 2006 and again in 2008. Each feature file contained 40 samples in total.

4.3 Studies in a pair of identical twins

The current two- and four-week studies include one pair of 14-year-old identical twins, presenting a unique opportunity to explore the impact of biology on keystroke patterns. Toward this end, studies were conducted on the identical twins, using their father (also a subject) and one random individual from the others who completed the short-term studies as controls.

Since the twins and the father had all used the exact same keyboards, the selection of the father as a control allows us to separate the impact of the specific keyboard from the impact of biology. However, since the father would be expected to be biologically similar to the twins, separate studies were also conducted with an unrelated subject as an additional control.

For these studies, data for each twin was pooled across all three time points (weeks 0, 2 and 4) and tested against similarly pooled data for (a) the other twin, (b) the father, and (c) an unrelated subject. Feature files were tested using "train on one/test on another" where the train and test data sets used a different keyboard and a different typing task (considered to be the "least optimal" conditions under which to identify an individual). Since only two individuals were represented in each data set, the likelihood of an accurate match "by chance" for any given identification attempt would be expected to be 50%. Thus, under these conditions one would expect high identification accuracy.

The results of the twin studies are shown in Table 5. The overall results (rightmost column) show extremely high accuracy in distinguishing an unrelated subject from either one of the twins, fairly high accuracy in distinguishing either twin from the twins' father, and much lower accuracy in distinguishing the twins from one another. These results support a strong influence of genetic make-up on keystroke patterns, although larger studies would be needed to confirm this finding. Moreover, additional factors such as the twins' gender, age, and even the impact of a similar environment could also play a role in the apparent similarity of their keystroke patterns.

Pooled Data	Train/Test				Avg
	DeskCopy LapFree	LapCopy DeskFree	LapFree DeskCopy	DeskFree LapCopy	
Twin 1/ Twin 2	50%	100%	50%	93.3%	**73.3%**
Father/ Twin 1	93.3%	96.7%	93.3%	100%	
Father/ Twin 2	100%	100%	100%	96.7%	
Average Father/ twin	96.7%	98.4%	96.7%	98.4%	**97.6%**
Unrelated/ Twin 1	100%	100%	100%	100%	
Unrelated/ Twin 2	100%	100%	96.7%	100%	
Average Unrelated/ twin	100%	100%	98.4%	100%	**99.6%**

Table 5. Identification accuracy in twin studies

5. Enhancements to the Authentication System

The authentication application was enhanced with two additional columns that display the train file used and the train sizes created, as shown in Figure 8. In addition, the saved HTML file was also modified to display the appropriate newly created columns.

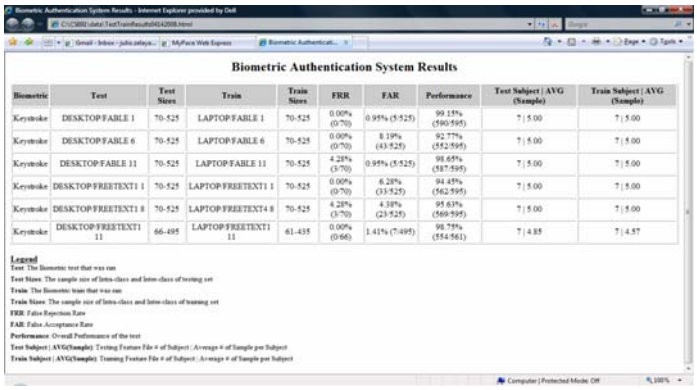

Figure 10. The new train and train sizes columns.

6. Enhancements to the Keystroke Monitoring System

These studies involved a number of geographically dispersed subjects who were asked to submit a total of 20 keystroke samples within a short time frame. A variety of factors can make this task difficult for subjects, and clear results depend upon clean and complete data collection. Regular monitoring by the study team is essential to ensuring good results. With such monitoring, there can be immediate follow-up to address obstacles or to ensure that incomplete data sets are finished before too much time elapses.

For this purpose, Web-based monitoring tool (written in PHP) was created in 2006 when Dr. Villani did her original keystroke biometric studies. This tool provides a quick list of all keystroke biometric subjects (past and present) and the total number of samples entered in each quadrant.

To provide for more accurate monitoring, the current project team improved upon the monitoring tool in several important ways. First, the tool was enhanced to provide not only sample counts, but also "date of last keystroke entry". This date is presented as a hyperlink which, when clicked, provides a listing of all samples entered by an individual, grouped by quadrant (desktop free text, laptop copy, etc.). Since subjects sometimes enter extra samples, a sample count alone provides limited information, and the availability of a sample list is essential in monitoring subjects who are returning to the system to enter new data for longitudinal studies after entering complete data sets in the past. Secondly, e-mail addresses that were previously simply displayed in the monitoring tool are now presented as "mailto" hyperlinks, facilitating the process of e-mailing a subject to resolve issues. Finally, an additional link has been added to provide for the zipping and download of all raw data files for a selected subject. This "zip and download" process had previously been carried out on an *ad hoc* basis through modification of an existing PHP program, requiring a time-

consuming manual process each time there was a need to obtain a different subject's data.

7. Conclusions

The current studies show a progressive decline in both identification and authentication accuracies over a four-week period, with a larger decrease after four weeks than over two weeks. In identification studies, an even larger decrease was observed over a two-year interval. These results suggest that identification accuracy does degrade over time. Two-year authentication results, surprisingly, were better than four-week authentication results. However, it must be noted that the two-year subject group also had better "same year" authentication accuracies. It is possible that the better authentication results in the two-year study are simply a by-product of the smaller sample size in that study (8 subjects) compared to the two- and four-weeks study (13 subjects).

When comparing identification results over the two- and four-week period vs. a two-year period, several factors need to be considered. In the two-year study, there is a much greater likelihood that different computers were used for year 0 vs. year 2. (In fact, subjects were asked to provide information on the keyboards used and most reported using different keyboards for the recent data set than for the original data set.) On the other hand, most subjects in the two- and four-week studies used the same laptop and the same desktop computer at all three time points. However, it is interesting that these factors did not similarly affect the two-year *authentication* results.

It should also be noted that the two- and four-week studies included a pair of identical twins, and the separate twin studies confirmed significant similarity in typing patterns between these two subjects. This would be expected to decrease matching accuracy in data sets that included both twins. Most likely, overall identification and authentication results for the two- and four-week studies would have been a bit higher if only one of the twins had been included. However, the removal of one twin would not be expected to change the longitudinal trends, an important finding of this study.

These studies should be repeated with larger sample sizes. Although the impact of a larger sample on the accuracy of keystroke biometric systems (i.e., the scalability of these systems) has not yet been clearly demonstrated [2], one could imagine that accuracy might decrease as sample sizes grew. This follows from the fact that the larger the group size, the greater the likely diversity of keystroke patterns within a training sample. With a larger pool of potential matches, the likelihood of an incorrect match would be expected to increase. The current and prior longitudinal studies [3] support this concept, showing decreased accuracies across-the-board in going from four to eight subjects, and from eight to thirteen. It is unclear whether a "leveling off" point would be reached at some point with a much larger sample. Scalability is a recognized problem in keystroke biometrics, and one that remains to be resolved [2].

Future studies should also explore improvements in analysis techniques that could lead to better accuracies over time. The Pace system analyzes data based on a set of 239 features. It is likely that additional features could be identified that, together with the currently used features, might significantly improve the ability to characterize an individual uniquely for either identification or authentication.

However, "good" accuracy may be sufficient from some real-world authentication applications. For example, in password hardening, typing patterns can be used in combination with other data (such as the actual username/password entered by a user) for authentication, so the typing pattern is not the only determinant of authentication success. Moreover, system adaptations could be made in real-world applications to lower the likelihood of a false rejection. For example, a user entering a correct username/password combination but failing keystroke authentication could be asked to re-enter the username/password. Another possible adaptation involves the criteria for authentication success. Instead of requiring that the best match in the training set be one of the user's training samples, a system could require simply that a correct match be found within the five "best matches". (The number 5 is arbitrary, and this number could also be scaled in proportion to the number of users in the training set.) Further studies are needed to assess the functional accuracy (i.e., the false acceptance and false rejection rates) in a situation that more closely resembles a real-world application.

Changes in authentication accuracy over time can be further addressed through dynamic updating to the training data as a user continues to log in over time. Monrose *et al* [10] report just such an approach to password hardening. Assuming a user logs in with some frequency, a solution like this would help to address likely age-related changes, which would tend to occur gradually. Even in the absence of time-related changes, such a system would gradually enlarge the training data set, improving its knowledge about each user and, presumably, its accuracy in authentication.

A major obstacle in this study was the difficulty in convincing the "original 36" subjects to return to the system and submit new keystroke samples. Future studies should consider incentives, if possible, to improve subject "compliance". For some subjects, a promise to send a copy of the final technical paper might spark some interest. For others, a more concrete incentive (e.g., an electronic gift certificate to Barnes and Noble) might prove to be a strong motivator.

However, several subjects expressed motivation to participate but simply did not have sufficient time. For a typical subject with an average typing speed, the entry of twenty samples takes approximately two hours. This is a significant time commitment for a busy person with no specific incentive to participate. Future authentication studies could explore the use of shorter text entries using a purely copy operation. This scenario would closely mimic a real-world "password-hardening" application where a user re-types the same password and the system tracks patterns in the way that it is typed, although the scenario would have less applicability to continuous monitoring, where the exact text typed by the user varies from session to session.

7. References

[1] F. Bergadano, D. Gunetti, and C. Picardi, "User authentication through keystroke dynamics," *ACM Transactions on Information and System Security (TISSEC)*, volume 5, issue 4, November 2002.

[2] A. Peacock, X. Ke, and M. Wilkerson, "Typing patterns: a key to user identification," *IEEE Security and Privacy*, volume 2, issue 5, pp. 40-47, September-October 2004.

[3] T. Buch, A. Cotoranu, E. Jeskey, F. Tihon, and M. Villani "An enhanced keystroke biometric system and associated studies," *Proc. CSIS Research Day*, Pace University, May 2008.

[5] S.-S. Choi, S. Yoon, S. K. Cha, and C. C. Tappert, "Use of histogram distances in iris authentication", http://www.csis.pace.edu/~ctappert/srd2004/paper06.pdf.

[6] S. Yoon, S.-S. Choi, S.-H. Cha, Y. Lee, and C. C. Tappert, "On the Individuality of the Iris Biometric," *ICGST-GVIP Journal*, volume 5, issue 5, May 2005.

[7] M. Ritzmann, "Strategies for managing missing or incomplete data in biometric and business applications", http://www.csis.pace.edu/~ctappert/dps/d861-07/pres-ritzmann.ppt

[8] S. Bharati, R. Haseem, R. Khan, M. Ritzmann, and A. Wong, "Biometric authentication system using the dichotomy model," *Proc. CSIS Research Day*, Pace University, May 2008.

[9] M. Villani, C. Tappert, G. Ngo, J. Simone, H. St. Fort, and S. Cha, "Keystroke biometric recognition studies on long-text input under ideal and application-oriented conditions," *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop*, February 2006.

[10] F. Monrose, M. Reiter, and S. Wetzel, "Password hardening based on keystroke dynamics", *Proceedings of the 6th ACM Conference on Computer and Communications Security*, Kent Ridge Digital Labs, Singapore, pp. 73-82, 1999.