A Longitudinal Study Investigating Pressure Related Characteristics for Keystroke Analysis

Christopher S. Leberknight
Department of Computer Science
Montclair State University
leberknightc@montclair.edu

Abstract

One factor impeding the acceptance of biometric security is the lack of actual classification rates based on usage in a real world setting. While previous results from controlled lab experiments have led to significant advancements in the field, they cannot be generalized nor do they accurately provide implications for practical deployment. This research investigates classification rates that can be achieved using a keystroke analysis prototype in an uncontrolled setting. We introduce a new pressure related feature, vpdelta, for classifying typing patterns and conduct a 5 week long field experiment to examine the variability of typing patterns over time. A classification rate of 87% was obtained during actual use and little variability in typing patterns was observed during the experiment. This finding helps to validate and extend previous lab studies, and provides increased support for the application of keystroke analysis in a real world environment.

Introduction

Most security systems used to control either physical or logical access lack the ability to verify an individual's identity. Biometric research attempts to address this limitation using different technologies and algorithmic approaches. Despite significant progress in the field, biometric technologies have still not gained wide-scale acceptance. Out of the small number of biometric access control systems in use today, most are primarily based on physiological characteristics. These systems are often quite expensive to develop and present several privacy concerns such as usage and ownership of the data. In addition, such systems are not well suited for providing transparent security controls that may be desired under certain conditions such as silent alarm notification. To balance this trade-off between security and usability an alternative approach to improve physical security using behavioral characteristics is proposed that offers several key advantages. Unlike physiological biometrics, a behavioral biometric can be observed and analyzed covertly without the intruder's knowledge. Another advantage of behavioral biometrics is the potential for lower manufacturing and integration costs. For example, keystroke analysis is a behavioral biometric that can be used to examine the uniqueness of different typing patterns. The most predominant types of access control systems in use today consist of magnetic or proximity based electronic keypads. These keypads could easily be retrofitted to incorporate simple electronic circuitry enabling keystroke analysis. Lastly, even though previous research indicates there are several challenges impeding the acceptance and adoption of biometric technologies (Chandra and Calderon, 2005) the results do not differentiate between physiological and behavioral biometrics. Due to the predominance of physiological-based biometric systems on the market, it may be more appropriate to assume that many of the issues surrounding the acceptance of biometric technologies (Chandra and Calderon, 2005) are directed more at physiological biometrics versus behavioral biometrics. This is an especially important factor when considering the impact of privacy issues on technology

acceptance. For example, perceptions of privacy may be different depending on whether the biometric is behavioral or physiological. The combination of high integration costs, privacy concerns with physiological biometric systems, as well as the desire for a covert security mechanism, necessitate further investigation into new approaches for behavioral biometric recognition systems. Therefore, this research aims to evaluate the efficacy of a behavioral biometric prototype based on keystroke analysis, in the context of perimeter defense and facility access in a 5 week long field study. This research builds upon prior work in keystroke analysis for computer security and using the field study methodology answers the call for an increase in information systems based approaches to security (Siponen and Oinas-*Kukkonen* , 2007).

The main contributions of this research are:

- presents empirical results from a field study that provide insight into actual performance of keystroke analysis using a biometric keypad prototype with new pressure features
- presents one of the very few papers to substantiate biometric classification performance based on human subject trials with a small sample text

The presence of real world data supporting the performance of a biometric is often absent in most of the literature and is one of the major inhibitors for wide-scale adoption of the technology.

Related Work

The review of research on keystroke analysis based on temporal characteristics and more recent work that has investigated the possibility of analyzing additional features such as pressure applied during each key press event, has provided valuable insight that has influenced the approach taken in this work. Unlike previous research, this paper examines the combined effect of keystroke durations and pressure features on the classification rate in a real world setting using an embedded device and a 4 digit sample text. Typing characteristics for keystroke analysis are derived by analyzing patterns comprised of keystroke durations (key hold time, keywords) and digraph latencies (time between successive keystrokes) are illustrated in Fig. 1 (Lin, 1997). A side by side comparison of temporal characteristics indicate that keystroke durations are better suited for identifying unique typing patterns compared to keystroke latencies. Overall, keystroke durations outperformed keystroke latencies in every instance (Robinson et. al., 1998). In addition, an experiment conducted by Bleha and Obaidat (Bleha and Obaidat, 1992) pointed out that misclassification rates increase steeply as the recognition string length drops to as few as ten characters. With respect to two consecutive keystrokes or digraphs, the digraph sample size is more critical for dynamic keystroke analysis versus static keystroke analysis (Leggett et. al., 1991).

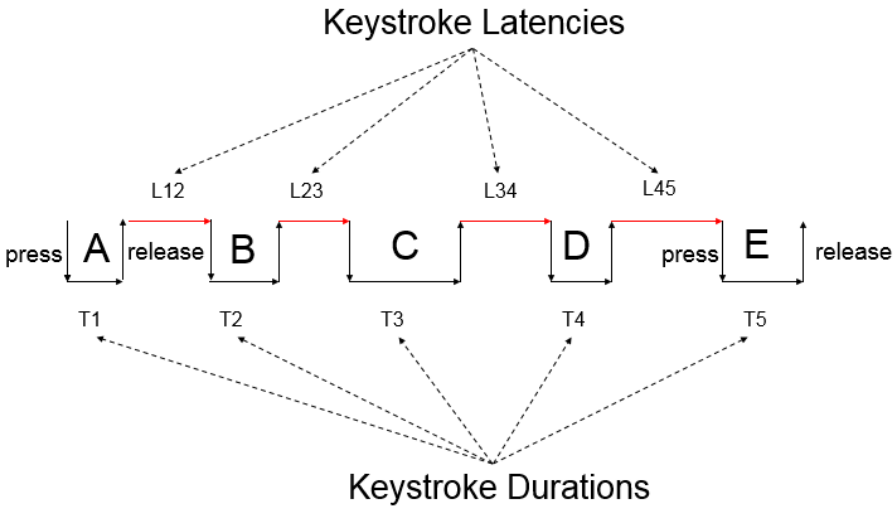

Figure 1: Temporal Characteristics

While the performance of keystroke analysis has primarily been evaluated by analyzing keystroke durations and keystroke latencies, pressure or force has also shown to be a promising feature for classifying typing patterns. Previous work investigates the use of combined keystroke pressure and latency using neural networks (Loy et. al., 2005, Loy et. al., 2007. In their earlier work (Loy et. al., 2005), the authors report a false accept rate (FAR) of 87% and a false reject rate (FRR) of 4.4%. They had 10 users who were not informed that their pressure patterns were collected from a keyboard equipped with pressure sensors. In addition, 10 samples based on a sample text consisting of 8 alphanumeric characters were collected in a single instance. This does not mimic the same behavior and typing patterns that would be generated if the samples were collected over a longer period of time. Research in our paper differs in the following ways. First, we analyze keystroke durations instead of keystroke latencies as they have shown to provide superior results (Robinson et. al., 1998). Second, due to Institutional Review Board (IRB) regulations, informed consent is required from the subjects participating in our study. This may have complicated the ability to achieve high classification rates. Since subject's knew their patterns were being captured it may have influenced how they entered their input on the keypad. However, this is most representative of a real world scenario in which users have to be informed about any personal data that is being collected. Third, a stand-alone embedded device similar to a proximity card reader was used to collect data from 14 subjects over a 5 week period. Literature shows that data collected from a keyboard is not likely to generate accurate typing patterns due to OS latency and the configuration of the keys on the keypad (Nonaka and Kurihara, 2004). As can be seen in Fig. 2, the orientation of the keys for the personal identification number (PIN) used in this research poses greater challenges for classifying typing patterns. The distance between keys greatly reduces the variability in keystroke durations that is required to distinguish typing patterns between different individuals. Lastly, we use a 4 digit sample text as the input for classifying typing patterns. Previous work that is most similar to our research also uses a keypad embedded with pressure sensors to collect keystroke samples consisting of a 4 digit number to classify typing patterns (Kotani and Horii, 2005). They achieved better results than what is reported in this research. However, their experiment was conducted in a controlled environment where subjects were instructed to"…assimilate melody notes of a well-known song might be helpful for recalling the dynamic keystroke patterns rather more efficiently than to memorize each keystroke" (pg. 293). This may have actually coerced the subjects into creating a consistent typing pattern and could be a significant factor contributing to their overall results. Unlike their work, this paper examines the performance of keystroke analysis in an uncontrolled setting where subjects are free to type their PIN any time during the experiment. This greatly increases the complexity associated with classifying unique typing patterns. Eltahir et al., also demonstrate the benefit of pressure sensors for keystroke analysis (Eltahir et. al., 2008). However, unlike our research their experiment uses a standard keyboard and analyzes keystroke latencies and peak amplitudes for sample text containing 6 alphanumeric characters. One advantage of their work compared to ours, is that they provide the equal error rate (EER) which gives a better indication of classification performance where our paper only reports the false accept rate (FAR). However, the FAR at the EER reported in Eltahir et al., 2008 is higher than what was achieved in our study.

The rest of this paper is organized as follows. Section 3 presents details of the field study followed by results in Section 4 and associated limitations in Section 5. Concluding remarks are provided in Section 6.

3 Field Study

The main purpose for this experiment is to evaluate the robustness of the biometric keypad prototype and classification performance of a 4 digit PIN under real world conditions. This is accomplished by analyzing the effect of typing patterns over time on the classification rate for several subjects. This study is aimed at rigorously verifying the lab results from previous research (Leberknight, 2015), providing a more meaningful and practical explanation of how the biometric keypad prototype would perform in everyday use. In a previous study (Leberknight, 2015), several factors influencing the classification rate for keystroke analysis were investigated. However, one factor that has not been extensively examined in previous research is the challenge of obtaining consistent results due to different behaviors exhibited during the experiment. Even though this experiment is not designed to specifically examine the impact of different behaviors on typing patterns, it is assumed that different behaviors are more likely to occur over an extended period of time compared to an experiment conducted over an extremely short interval, as was done in most previous

lab studies. Consequently, it is assumed that the field experiment accounts for the effect of different behaviors on the classification rate over time as opposed to the five minute interval used to capture patterns in previous experiments (Leberknight, 2015). Capturing typing patterns over an extended period of time are more indicative of individual's true typing rhythm compared to mass enrollments in which several typing samples are collected within a one to three minute intervals. Due to changes in behavior over time, understanding the length of time required for

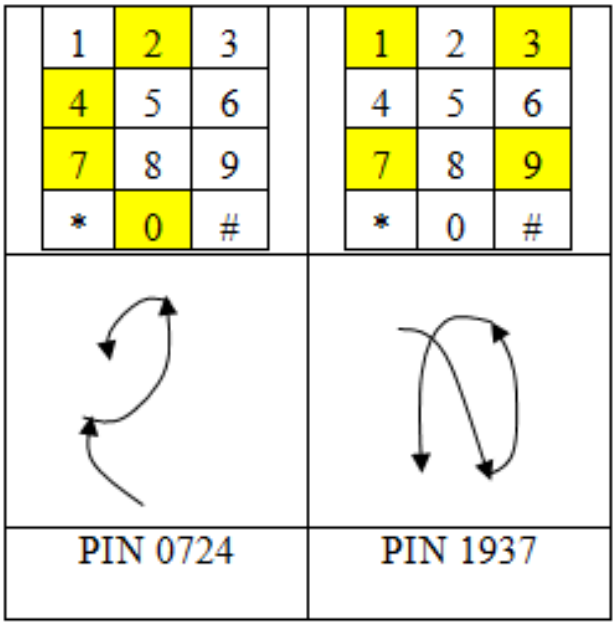

Figure 2: PIN Configuration

individuals to attain a stable typing pattern is critical for assessing the reliability of the security system. The length of the data collection process helps to explain the amount of data necessary for each subject to achieve a consistent typing pattern. Previous research has indicated that keystroke patterns stabilized after 10 weeks of experimental sessions (Kotani and Horii, 2005). Several other examples have also provided guidance for designing the field experiment with respect to the length of the data collection process (Bleha et. al., 1992). However, the results offer no conclusive evidence that would point to some standard or baseline for the minimum number of typing samples required to achieve a stable pattern under real world conditions using temporal and pressure related characteristics. A review of literature on keystroke analysis revealed that a severe limitation inhibiting the actual use of biometric technologies is the lack of real world data. Of the limited examples which attempt to simulate a real world test case, none have evaluated the addition of pressure and discussed the time required for patterns to stabilize (Robinson et. al., 1998). In addition to the challenge of obtaining consistent patterns due to changing behaviors over time, previous research has also indicated that keystroke analysis is not accurate enough to be used as a unimodal biometric (Umphress and Williams, 1985). However, this finding was based on classification rates obtained by analyzing only temporal characteristics. Based on previous results, we claim that using pressure characteristics as well as temporal features that are extracted from an embedded device, will yield higher classification rates than any one characteristic analyzed on a personal computer. More characteristics may produce better discrimination quality, and performing the computations on an embedded device versus a personal computer, less latency induced inaccuracies will be produced. Despite the support for keystroke analysis as a unimodal biometric, the approach taken in this research is to combine password and biometric security. This design provides tighter controls by requiring two layers of security: authorization and authentication. The main reason for taking this approach is that the additive effect of both authorization and authentication methods increases over time. While password security decreases exponentially over time, biometric security increases as a function of standard error. Essentially, the power of any biometric is based on the number of samples used to compute a profile. As the number of samples, or N, increases the

standard error decreases. Therefore, as the number of samples increases, the biometric security increases and the additive effect over time of using both systems ultimately leads to increased security.

3.1 Experimental Design

The requirement to evaluate the application of the biometric keypad in a real world setting precluded the manipulation of any control variables. Therefore, the design of the experiment was entirely predicated upon the natural setting and actions of the subjects involved in the experiment. The one exception was that all subjects were assigned the same PIN. The main focus of the field study involved the examination of the classification rate over time. The time period consists of the time and number of samples collected during different weekly intervals over a 5 week period. The purpose of the time

Table 1: Field Study Experimental Design

Dependent Variable	Level	Condition
Classification level	Continuous	0-100%
Independent Variable	**Level**	**Condition**
Time Period	2	Based on number of samples collected over a 5 week period the levels are the first 2 weeks and last 3 weeks
Feature	3	Combined features (duration, peak, and vpdelta) based on results from the lab study
PIN	1	PIN 0724
Algorithm	2	kNN and T-DIST(Semi) based on results from the lab study

variable was to examine how the classification rate may change or be affected during different data collection periods. Therefore, the independent variables for the field study were: (1) time period, (2) feature, (3) PIN, and (4) algorithm, and the dependent variable was the classification rate. A summary of all the variables in the field study is provided in Table 1. The features include the peak amplitude, (the highest point in the signal) the vpdelta (the time between the highest and lowest amplitude), and the duration (the total length of time for the entire signal for a key press). The combination of these three features produced the highest classification rate compared to the classification rate of any single feature. The levels for the time periods were determined based on the number of samples collected at the end of the field study.

3.2 Subjects

A total of 14 subjects were enrolled into the field study. The subjects included faculty, staff, and students from a major public research university in the Northeastern United States. However, due to dropout and sampling effects only data from 6 subjects were used in the analysis.

4 Results

A total of 437 typing samples were collected from all of the subjects. The average number of samples collected per subject was 33.6 with a standard deviation of 26.3. The spread of the data corresponding to each subject's' usage of the biometric keypad is very large. The highest usage was observed with subject 6 who used the biometric keypad 90 times over the 5 week period, while the lowest usage was observed with subject 9 who only used the biometric keypad three times. The greatest number of access attempts occurred

on Monday and Thursday during the entire 5 week period. Preliminary analysis of the access attempts suggests that access patterns based on the time and day may also be used in conjunction with the combined

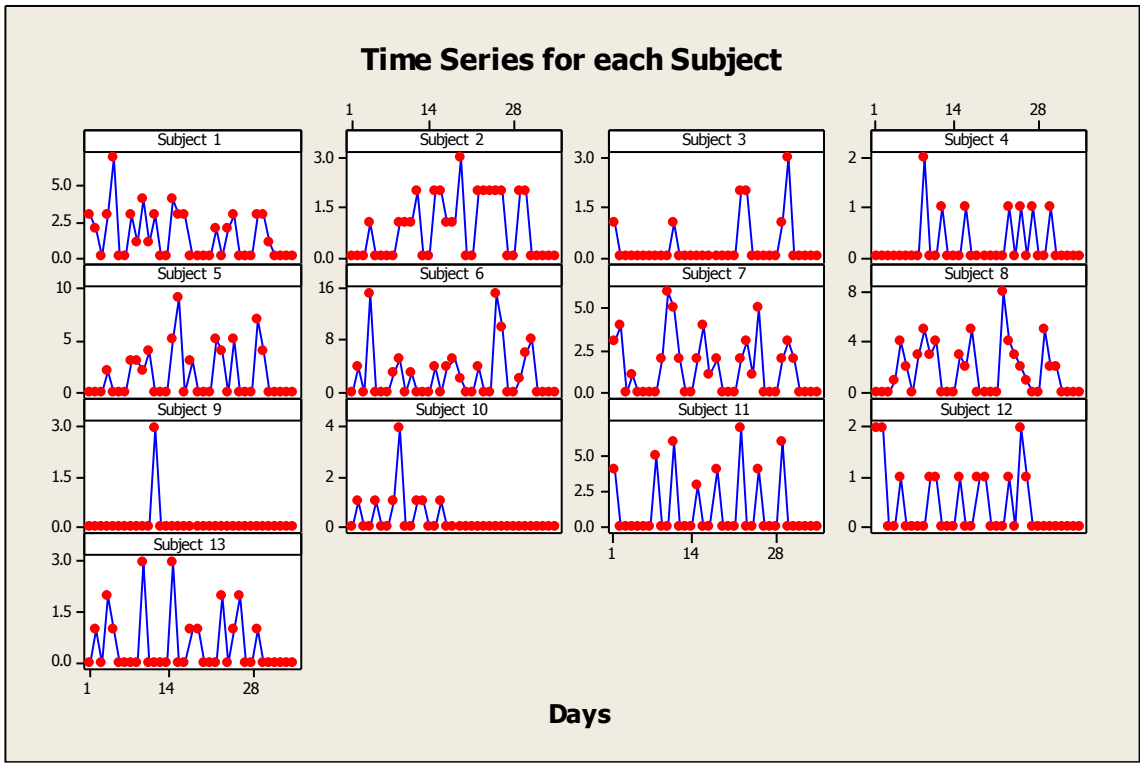

Figure 3: Biometric Keypad Usage (x-axis) for each day (y-axis)

features to improve classification accuracy. This information could be used to identify possible access anomalies and dynamically adjust the threshold for the classification algorithm in order to increase or decrease security on different times and days. For example, it can be observed in Fig. 3, that several distinct patterns emerged over the duration of the field study. Each graph in Fig. 3 represents the number of times a particular subject used the biometric keypad to access the restricted location. Considering the total amount of times each subject used the biometric keypad combined with results in Fig. 3, it is possible that different usage patterns might also be used as a behaviometric feature to improve security. For example, the graph for subject 9 shows that this subject only used the biometric keypad to access the restricted location three times all on the 12th day over the entire five week study. This is quite different from subject 1 who more frequently accesses the location. Using this information to categorize which individuals seldom and frequently access a particular location may provide additional value for discriminating between different individuals. After the keypad usage for each subject was computed the typing patterns were analyzed during different time periods using the combined feature. The decision to use the combined feature to classify typing patterns was based on the superior results obtained in the lab study (Leberknight, 2015) compared to other features. The main point of interest for this study is to investigate whether there is a point in time during the 5 week period in which the patterns stabilized. Unfortunately, since this experiment was not controlled, in order to mimic a real world application and usage of the keypad, no hard limit was set on the amount of data captured. As a result, there were many instances where an insufficient amount of data was present to classify many of the subjects. Consequently, seven subjects out of the 13 subjects were eliminated, and only six subjects were retained for analysis. In addition, there were many instances in which a small amount of data was captured during one of the time periods. The lack of a large and balanced data set restricted the analysis to two time intervals. The first interval contains the combined features for all six subjects that were collected during the first two weeks (T1), and the second interval (T2) contains the combined features for all six subjects that were collected during the last three weeks. There were a total of 33 samples for each of the six subjects. T1, consists of the first two weeks of the study and

contains 15 samples, and T2 consists of the last three weeks of the study and contains 18 samples. While the number of subjects in the field study is much smaller than the lab study (Leberknight, 2015) each subject in the field study has 32% more samples. The classification

Table 2: Field Study Mean Classification Rate (%)

	T1	T2
kNN	86.67	86.33
T-DIST (Semi)	73.83	78.83

rate for each pair of subjects was computed for both time periods using the kNN and T-DIST (Semi) algorithms. The choice to use these two algorithms was based on the results from the lab study (Leberknight, 2015) that suggests the kNN and T-DIST (Semi) produce the best classification accuracy. TDIST (Semi) algorithm is an implementation of Cauchy Classifier. The standard Cauchy(0,1) distribution is a special case of the Student's t distribution with one degree of freedom. With six subjects, a total of 15 comparisons or classification rates were computed for each time period. The mean classification rates, in percentages, for the 15 comparisons between the six subjects are presented in Table 1. The first column lists all type of algorithms used for classification; kNN and T-DIST (Semi) algorithms. The column labeled T1 contains the mean classification rates for all paired comparisons during the first two weeks while the column labeled T2 contains the mean classification rates for all paired comparison during the last three weeks. The highest average classification rate for all of the subjects is 87% using the kNN algorithm, and the lowest average classification rate is 74% using the T-DIST (Semi) algorithm. Both of these rates were observed during T1 and the kNN algorithm produced the best classification rate. The results are consistent with results obtained in a previous lab study (Leberknight, 2015). Overall, the average classification rate, between each algorithm, for all subjects during time period T1 and time period T2 are very similar. The average classification rate for all of the subjects using the kNN algorithm is 87% during time period T1 and 86% during time period T2. The T-DIST (Semi) algorithm performed significantly lower. The average classification rate for all of the subjects using the T-DIST (Semi) algorithm is 74% during time period T1 and 79% during time period T2.

The kNN algorithm appears to have performed more consistently over the two time periods with only a 1% difference, while the T-DIST (Semi) differed by 5% between the two time periods. In the next section, the results of the individual features and classification rates are presented to examine the differences between the typing patterns over time.

4.1 Analysis of Typing Patterns over Time

To visualize the difference between the individual features of all six subjects, the mean connect line for each feature is used to provide a cleaner illustration of the differences across key, feature, and time. For example, the mean value of each key and feature for subject 11 using PIN 0724 is illustrated in Fig. 4. The main point of interest is the existence of a significant difference in the typing pattern between the two time periods. Results in Fig. 4 suggest that the different typing patterns across all features for both time periods are fairly consistent. The top panels in the diagram correspond to the mean feature value of each key during the first time period T1. The bottom panels correspond to the mean feature value of each key during time period T2. The features from left to right are the duration, peak amplitudes and vpdeltas. The y-axis corresponds to the mean value for each key measured in voltage for the peak amplitude and time for the vpdeta and duration feature. The x-axis corresponds to each key in the PIN 0724. In this diagram, by comparing any two like features over the two time periods it can be observed that the mean connect lines are approximately equal. For example, the duration values in the first top and bottom panels are at the same height and follow a very similar pattern for both T1 and T2. In addition, the peak amplitude and vpdelta features in the second and third panels are also at the same height and follow a similar pattern during both time periods. Overall, by inspecting the individual features, there is no significant difference between typing patterns for each subject over the two time periods. The analysis for the five other subjects also shows relatively small differences suggesting similar patterns for the two time periods. These results suggest typing patterns using

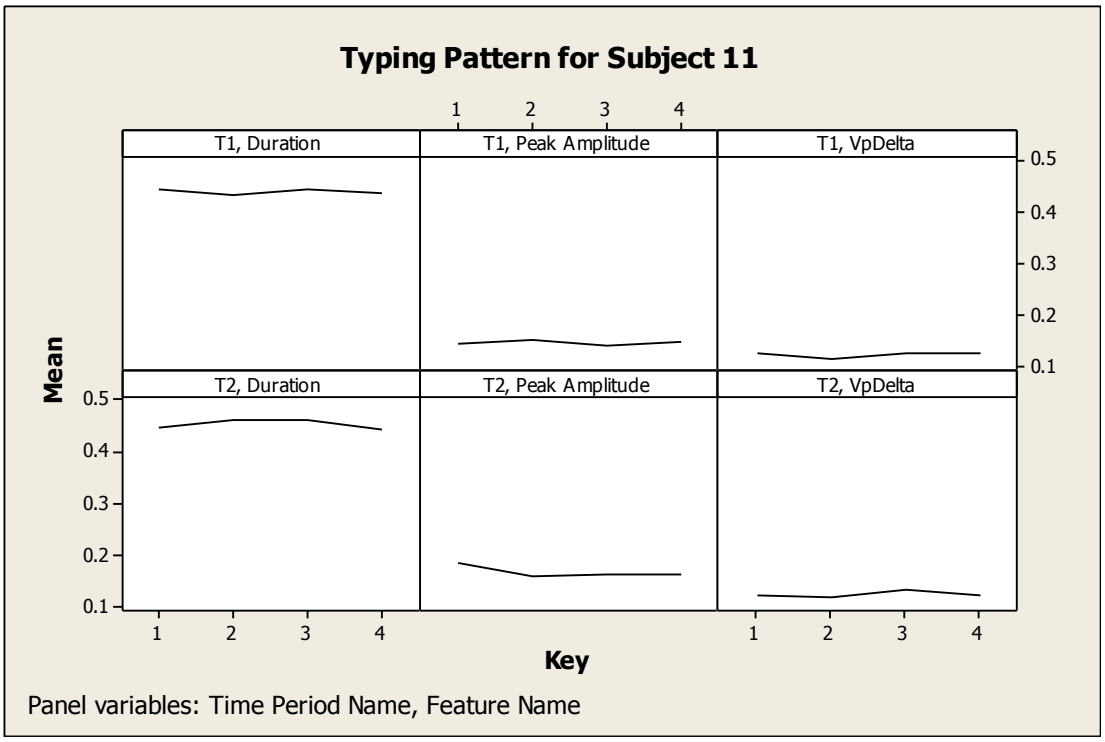

Figure 4: Typing Pattern for Subject 1. Mean voltage level (y-axis) for each key (x-axis)

the three features with the biometric keypad prototype stabilized after 15 samples and the 18 typing samples collected from each subject during the last three weeks were not required. The assumption that there was little to no improvement classifying individual typing patterns after collecting an additional 18 samples over three weeks based on the results in Fig. 4 is verified against the mean classification rates provided in Table 2. The results in Table 2 support the results observed in Fig. 4 that suggest the patterns stabilized after the first two weeks. Inspecting the classification rates for each algorithm shows there was a 1% and 5% difference in the classification rates between the first two weeks and the last three weeks for the kNN and T-DIST (Semi) algorithm, respectively. Therefore, no significant improvement in classification accuracy was observed by providing additional time to use the keypad.

5 Limitations

The field study was entirely conducted indoors and an investigation into the robustness of the hardware when exposed to outdoor environmentals should be conducted. Also, while it is assumed that during the 5 week study subjects may not always have been in the same mental, physical, or emotional state, these specific effects on the classification rate were not tested. In addition, due to the limited amount of keypad usage by several subjects during the 5 week experiment, data for only 6 out of the 14 subjects could be used for the analysis. Therefore, to further verify the results obtained in the field study, future research should include a larger sample size which examines the effect of stress or fatigue on classification accuracy. In addition, the results are based on the analysis of 9 algorithms from a previous lab study (Leberknight, 2015). Future research will investigate the performance against different algorithms such as support vector machines (SVM). Lastly, this research reports the performance of the biometric in terms of false accept rates. Future research will examine false reject rates and more importantly equal error rates to provide greater details on the performance of keystroke analysis using pressure-related characteristics.

6 Conclusion

This research highlights a new pressure feature, vpdelta, and the optimal algorithm that can be used with a small sample text to generate the highest classification rate under real world settings. During a 5 week field

experiment the effect of time on classification accuracy was examined and a classification rate of 87% was achieved during each of the two time periods. Overall, it is difficult to compare these results with other studies that have investigated pressure as an additional feature (Kotani and Horii, 2005, Eltahir et. al, 2008, Loy et. al., 2005, Loy et. al., 2007) since there is no standard reporting of a performance metric and different time frames, sample text, and devices are used to capture the data.

References

Bleha, S., Knopp, J., & Obaidat, M. (1992). Performance of the perceptron algorithm for the classification of computer users. Paper presented at the Proceedings of the 1992 ACM/SIGAPP symposium on Applied computing: technological challenges of the 1990's.

Bleha, S. A., & Obaidat, M. S. (1991). Dimensionality reduction and feature extraction applications in identifying computer users. Systems, Man and Cybernetics, IEEE Transactions on, 21(2), 452-456.

Chandra, A., & Calderon, T. (2005). Challenges and constraints to the diffusion of biometrics in information systems. Communications of the ACM, 48(12), 101-106.

Kotani, K., & Horii, K. (2005). Evaluation on a keystroke authentication system by keying force incorporated with temporal characteristics of keystroke dynamics. Behaviour & Information Technology, 24(4), 289-302.

Leberknight, C. (2015). An Embedded System for Extracting Keystroke Patterns Using Pressure Sensors (in press). International Journal of Biometrics.

Leggett, J., Williams, G., Usnick, M., & Longnecker, M. (1991). Dynamic identity verification via keystroke characteristics. International Journal of Man-Machine Studies, 35(6), 859-870.

Lin, D.-T. (1997). Computer-access authentication with neural network based keystroke identity verification. Paper presented at the Neural Networks, 1997. International Conference on.

Loy, C. C., Lai, W., & Lim, C. (2005). Development of a pressure-based typing biometrics user authentication system. ASEAN Virtual Instrumentation Applications Contest Submission.

Loy, C. C., Lai, W. K., & Lim, C. P. (2007, November). Keystroke patterns classification using the ARTMAP-FD neural network. In Intelligent Information Hiding and Multimedia Signal Processing, 2007. IIHMSP 2007. Third International Conference on (Vol. 1, pp. 61-64). IEEE.

Nonaka, H., & Kurihara, M. (2004). Sensing Pressure for Authentication System Using Keystroke Dynamics. Paper presented at the International Conference on Computational Intelligence.

Robinson, J. A., Liang, V., Chambers, J. M., & Mackenzie, C. L. (1998). Computer user verification using login string keystroke dynamics. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 28(2), 236-241.

Siponen, M. T., & Oinas-Kukkonen, H. (2007). A review of information security issues and respective research contributions. ACM Sigmis Database, 38(1), 60-80.

Umphress, D., & Williams, G. (1985). Identity verification through keyboard characteristics. International Journal of Man-Machine Studies, 23(3), 263-273.

Eltahir, W., Salami, M. J. E., Ismail, A., & Lai, W. (2008). Design and evaluation of a pressure-based typing biometric authentication system.EURASIP Journal on Information Security, 2008(1), 345047.