Literature Review
Keystroke Analysis for Remote Authentication

Gabriel Padis
School of Computing
Dublin City University
gabriel.padis2@mail.dcu.ie

Rim Zaafouri
School of Computing
Dublin City University
rim.zaafouri2@mail.dcu.ie

Dr Darragh O'Brien
Supervisor
School of Computing
Dublin City University

Date of publication: October 1, 2020

I. Topic of proposed practicum

Real-time keystroke analysis for password authorization is a dynamic biometric based authentication approach used to limit a user's access to an application or a database. Keystrokes dynamics is a type of behavioural biometrics that measures attributes of a user including for example latencies between keystrokes, keystroke durations, finger placement and pressure, typing speed, or even the order in which the keys are pressed. This new technology is a non-invasive and user-friendly promising approach to reinforce the security of systems, as it requires no specific hardware equipment other than a keyboard. The authentication could be done at an entry-level, or as a continuous verification.

In the context of our practicum, we hope to assess the applicability of keystroke dynamics for detecting intruders over the network.

We will first be going through a literature review of five of the papers that we read, and an evaluation of how that existing work will relate to our research. Then we will discuss the research questions we aim to address, including the choices of softwares, datasets and experiments to conduct.

II. Literature review

A. Static Authentication

Static keystroke dynamics is the study of fixed static data used in the purpose of authentication, such as passwords.

The first survey paper we will be discussing gives an overall idea about the general work done in the area of keystroke dynamics. It is called *Keystroke dynamics for biometric authentication — A survey* and was written by Shanthi Bhatt and T. Santhanam for the 2013 *International Conference on Pattern Recognition* [7]. The paper first describes the concept of biometrics which is a pattern recognition scheme that limits a user's access to applications or systems. Keystroke dynamics are behavioural biometrics which collect data that is captured from the user and stored in a dataset. The authentication approach makes it a dynamic biometric based approach. This survey then exposes the various methods of keystroke dynamics capturing. These methods include patents, statistical analysis, neural networks, fuzzy logic, data mining, and mobile phone keystroke analysis.

With statistical methods, some researchers studied the efficiency of structured text analysis compared to free-text by working on a dataset of 63 users and using Euclidean measure. With neural networks, other researchers highlight the results of digraphs and trigraphs on 154 users in 2002 and achieved a performance with a False Acceptance Rate (FAR) of 0.01 and False Rejection Rate (FRR) of 4. At last, this paper also gives an insight on the datasets collected for keystroke dynamics studies. Prior to 2009, almost all researchers collected their own data and created their own datasets for study purposes. During the Dependable Systems and Networks (DSN) 2009 conference, a requirement for common datasets was set in place by scientists. These common datasets include the data of many users with several measured attributes (keystroke latency, keystroke pressure, typing speed, finger placement, ...) , and in different formats (fixed-width format, or free-speech format).

The second paper we have researched is *Keystroke dynamics as a biometric for authentication* by Monrose and Rubin [1], which is the most cited paper of the keystroke dynamics in a security context for user authentication. It is also one of the first to be written in English in 1990. It describes keystrokes as a biometric for identity verification, on the same par as hands, fingerprints and eyes. The pattern that is observed is how each user type, not what they type. From it the different steps of representation, extraction, classification and identification are defined. They used a single feature : latency in digraphs. Instead of using keystroke timings like their predecessors and filtering outliers values, because it does not seem like a good way to have precise results. They extend the research on fixed authentication using covariance matrix of latencies. They did many classifiers (Euclidean, Weighted probability, Bayesian, ...) and for each calculated Equal Error Rate (EER). They had their best result with the Bayesian detector with a 92% correct identification rate. They remarked the difficulty of free text usage because of the fact that it is less precise and harder to have a good model for. Proposed rheir solution as an application for root access verification on servers, an idea that we also had a bit earlier in our research while brainstorming the uses that technology could have.

In this research paper [3], the distance between two samples was computed on the basis of the relative positions of the trigraphs the samples are made of. In this study, the sample was not thrown away because of typing errors, which increased the collected number of trigraphs. According to the number of samples given, the users were classified in 4 different categories. Users who had given 4 samples had a 100% accuracy rate, while users who had given just 1 sample had a 97% accuracy rate. As a result, a 4% False Alarm Rate and an Impostor Pass Rate of less than 0.01% were achieved (less than one successful attack out of 10,000 attempts).

The purpose of this paper [12] was to collect a keystroke-dynamics data set, to develop an evaluation procedure, and to compare the performances of a range of different authentication algorithms. The first step of this research was to collect a sample of keystroke-timing data. The second step was to implement 14 anomaly-detection algorithms that analyze the password-timing data, including algorithms of Manhattan, Euclidean, and Mahalanobis, neural network, fuzzy-logic, etc. As a result, the best equal-error rate was 0.096, obtained by the Manhattan detector, and the best zero-miss false-alarm rate was 0.468, obtained by the Nearest Neighbor (Mahalanobis) detector.

The following paper [14] references the K.S. Killourhy and R. Maxion paper which indicates that the top performing classifiers as feature matching for keystroke dynamics algorithms are the Manhattan distance and the Mahalanobis distance, and proposes a new distance metric that combines the benefits of these two classifiers. This new distance metric ensures that undesirable correlation and scale variations are accounted for, and that the influence of outliers is suppressed. As a result of conducted experiments using the nearest neighbor classifier, the new proposed distance metric was proven to improve the accuracy of keystroke dynamics using static text, as it reduced the EER to 8.7%.

In this paper [13], researchers used a pre-existing dataset of 51 subjects that contains 400 repetitions of passwords. Different times of pressing and releasing the keys were captured and used as labels. The data was then evaluted by the statistical method of Support Vectors Machine (SVM) and by a deep learning model. The accuracy of the results was then analyzed: deep learning (92.6% accuracy) outperforms SVM (71.15% accuracy). This shows that user authentication may be achieved using deep learning, specifically using an optimizer (NADAM) for higher accuracy.

This paper [6] proposes an approach to identify a legitimate user of a mobile phone by an analysis of their keystroke dynamics. As a first step, an application logging mobile keystroke data was developed. The application runs in the background so that it isn't intrusive to the user, and it logs all the keys pressed by the user along with the press and release times of the keys. A dataset of 25 users quantified into profiles of 250 key-hits was collected. The features selected for the analysis of the data were key hold time, digraph time, and error rate. The data was then analyzed with 5 existing algorithms for evalutation, and then implemented with a proposed tri-mode system for identification which includes a learning mode using a Feed-Forward Fuzzy Classifier, a detection mode using dynamic optimizers (Particle Swarm Oprtimization, and General Algorithms), and a verification mode. As a result, the proposed system performed with an error rate of less than 2% and an FRR of close to 0 after the verification mode.

B. Continuous Authentication

Continuous authentication is the process on identifying a typer on a long period of time even after an authentication process. The ryhtm of the text typed is analyzed.

Various methods have been used in continuous authentication. Patrick Bours and Soumik Mondal from the Norwegian Information Security Laboratory did a experience in in 2015 *Continuous Authentication with Keystroke Dynamics* [8], and they used Reinforcement learning algorithms. The focus is aimed at two main requirements: the first is to not interrupt the user in their daily activity, and the second is for the system to use every keystroke to determine the genuineness of the user. In the purpose of this study, researchers Bours and Mondal have collected data of 53 users over the period of 5 to 7 days, and have defined a trust model which evaluates the trust of the user based on a behaviour comparison of the current user with a template of the genuine user. The trust model uses key digraph classification protocols for the verification process, which include scaled Euclidean distance and correlation distance measurements. Depending on the retrieved results from the measurements, the trust model increases the system's trust (called Reward), or decreases it (Penalty). To test their model, the data was investigated with a panel of 52 impostor users and 1 genuine user. The results have then been analyzed and classified in different categories: very good (all 52 impostors are detected and the genuine user is never locked out), good (the genuine user is never locked out but not all impostors were detected or the genuine user is locked out but all impostors are detected), bad (the genuine user is locked out and not all impostors are detected) and ugly (the false detection is drastically higher than the accurate one in all cases). Most of their results fall in the "Very Good" or "Good" categories, which means their study was rather successful. In the future, Bours and Mondal hope to improve their results by applying different classification techniques and different functions.

In the paper *Iterative Keystroke Continuous Authentication: A Time Series Based Approach* by Abdullah Alshehri, Frans Coenen and Danushka Bollegala [11], they took a completely different axis of research by considering the behaviour as a form of time series while comparing a sample to a template,

instead of using vector that requires to have a large amount of data. They highlight the problem of large feature vectors that would happen in a continuous authentication system. It underlines the fact that for big number of n-graphs there is a need of big sets of data per user so as to not have any holes in the n-graph possibilities. It could be completed with Neural Network for the missing values, but has not proven greatly successful. Digraph are more common so more data means more precision. To construct for each person a template, they used only the key flight between a key up and the next key down. A subsequence consists of n time points, and a profile is a set of subsequences. They used the Dynamic Time Wrapping method to find similarities between points of subsequences. With a a matrix of minimum warping distance, its sum is the warping path. The smaller it is, the closest the two samples are. They used a similarity threshold taken from the average of the profile warping distance. They obtained a False Match Rate / False Acceptance Rate of 0.54% and a False Non Match Rate / Fale Rejection Rate of 1.64%. It is better than their comparison to a vector approach with flight times of n-graphs, respectively 8% and 6%.

Encouraging results have been obtained by Daniel Gunetti and Claudia Picardi in 2005 [4]. They tried to authenticate personal identities of 205 individuals with a False Alarm Rate below 5% and and Imposter Pass Rate of less than 0.005%. They used pressed time for duration of n-graph, the time between pressed time of first key and the nth. The mean of each n-graph was calculated for the typed samples. Each n-graph are independent from the text and thus can be compared together. They calculated the distances of two typing samples using n-graphs with n going from 2 to 4.

Yan Sun, Hayreddin Ceker and Shambhu Upadhyaya [9] in 2016 applied a Gaussian Mixture Model (GMM) on raw keystroke data from 157 subjects. It uses the mean and the standard deviation of two consecutive keystrokes, digraphs. Their results show that using a GMM gives a EER of 0.39% whereas using a Gaussian Distribution yields an EER of 0.01%. It also shows their dataset has a greater precision that the one available previously with a GMM EER of 0.08% and a Gaussian distribution EER of 1.3%.

Jiaju Huang, Daqing Hou, Stephanie Schuckers, Timothy Law, and Adam Sherwin [10] compared their algorithm of Kernel Density Location (KDE) against Gunetti & Picardi's and Buffalo's SVM algorithm. KDE is a non-parametric way to estimate the probability density function (PDF) of a random variable Needs 4 or more instances in sample and profile in digraph flight times to have a good enough average after calculating their probability density. It's EER and performance are slightly better than Gunetti & Picardi's method on the tested datasets (0.04095% against 0.055225% EER). But KDE is better in truly uncontrolled environment with a stable and consistent performance, especially in noisy datasets.

III. Relation to existing work

There are many variations in keystroke dynamics analysis depending on formats, purposes, methods, measures of success, datasets and supports. We want to help information systems security so we will use different methods on datasets that we have seen and studied in the literature review. Fixed-length authentication will prove very useful for access-control so we will use the anomaly comparison of detectors [5] to re-implement the most efficient methods found in the studied research papers on a chosen dataset. We also hope to evaluate how feature change and engineering might prove useful for an over the network case, while taking into account that the primary purpose is high security identification.

As the possible context of application is over the network security, we are restricted with the number and complexity of features at our disposal. The work of Monrose and Rubin [1] uses just latency as a feature for authentication, so it can be a good starting point to see how it works and how feature engineering may improve the result they found.

Analysis over the network is also possible. In Dawn Xiaodong Song, David Wagner and Xuqing Tian paper in 2001 [2], they described how SSH packages can leak approximate size of data due to padding issues and how every individual keystroke that a user types is sent to the remote machine in a separate IP packet after the key is pressed. This leaks the keystroke timings and allows to do a Gaussian statistical approach to analyze the users typing rhythms. They then used a Hidden Markov Model and the n-Viterbi Algorithm to try to reduce the time it would took to brute force a password by simply eavesdropping on the network.

On the topic of continuous authentication many algorithms are testable on our datasets like Time Series [11], or other methods in the literature like deep learning that have showed great results. This approach is also restricted on what is possible feature wise on a network usage context, so we will evaluate how feature removal and changes impact the detection capabilities.

References

[1] Fabian Monrose and Aviel D. Rubin. "Keystroke dynamics as a biometric for authentication". In: *Future Generation Computer Systems* 16.4 (2000), pp. 351–359. ISSN: 0167-739X. DOI: https : / / doi . org / 10 . 1016 / S0167 - 739X(99) 00059 - X. URL: http : / / www . sciencedirect . com / science / article / pii / S0167739X9900059X.

[2] Dawn Xiaodong Song, David Wagner, and Xuqing Tian. "Timing Analysis of Keystrokes and Timing Attacks on SSH". In: (2001).

[3] DANIELE GUNETTI FRANCESCO BERGADANO and CLAUDIA PICARDI. In: (2002).

[4] Daniele Gunetti and Claudia Picardi. "Keystroke Analysis of Free Text". In: *ACM Transactions on Information and System Security (TISSEC)* (2005).

[5] Killourhy K. and Maxion R.A. “Comparing Anomaly-Detection Algorithms for Keystroke Dynamics”. In: *IEEE/IFIP International Conference on Dependable Systems & Networks* (Aug. 2009), pp. 125–134. DOI: 10.1109/DSN.2009.5270346.

[6] Syed Ali Khayam Saira Zahid Muhammad Shahzad. In: (2009).

[7] S. Bhatt and T. Santhanam. “Keystroke dynamics for biometric authentication — A survey”. In: (Feb. 2013), pp. 17–23. DOI: 10.1109/ICPRIME.2013.6496441.

[8] Patrick Bours and Soumik Mondal. “Continuous Authentication with Keystroke Dynamics”. In: (Jan. 2015). DOI: 10.13140/2.1.2642.5125.

[9] Yan Sun, Hayreddin Ceker, and Shambhu Upadhyaya. “Shared Keystroke Dataset for Continuous Authentication”. In: *2016 IEEE International Workshop on Information Forensics and Security (WIFS)* (2016).

[10] Chris Murphy et al. “Shared dataset on natural human-computer interaction to support continuous authentication research”. In: *2017 IEEE International Joint Conference on Biometrics (IJCB)* (2017), pp. 525–530.

[11] Abdullah Alshehri, Frans Coenen, and Danushka Bollegala. “Iterative Keystroke Continuous Authentication: A Time Series Based Approach”. In: *KI - Künstliche Intelligenz* 32 (2018). DOI: 10.1007/s13218-018-0526-z.

[12] Roy A. Maxion Kevin S. Killourhy. In: (2018).

[13] Dion Darmawan Yohan Muliono Hanry Ham. In: (2018).

[14] Anil K. Jain Yu Zhong Yunbin Deng. In: ().