Keystroke Dynamics

Mohamed Al-Rasbi, Pranav Gummaraj Srinvas, Srinivasu Seeram

Problem Space

Authentication is the process of identifying the user by applying three common procedures based on what the user knows, what the user has and what the user is. Popular techniques for identification using what the user knows are passwords, PIN information, previously lived city etc. However, this data is already stored somewhere so a simple data breach can reveal everything about the person. So, identification using "what the user has" such as RSA key and/or "what the user is" such as fingerprints, iris pattern became an important second step of authentication for highly confidential data. These procedures require additional hardware which is not suitable for low confidential data applications. In such cases, keystroke dynamics can work as a second step of authentication without adding any additional hardware requirement.

Keystroke dynamics is the detailed timing information of the user's typing pattern. When the user types on a keyboard, he follows a unique stroke pattern. The timing information in the pattern can be used to classify and identify the user with a certain confidence interval. Even if the user uses a different type of keyboard such as optical one or keyboard on a touch device, we should still be able to identify the pattern by transforming the data. The timing information contains mainly key duration time and key latency time. Various features will be extracted from this timing data for the classification purpose.

Dataset

We are planning to primarily collect data on our own and also use "Free vs. Transcribed Text for Keystroke-Dynamics Evaluations" dataset. In this dataset, keystroke timing information for 20 users is recorded when they are transcribing and typing free text. We need timing information of transcribing text for our project. For each user, approximately 4500 entries of keystrokes are recorded. Below are the samples of the dataset:

Downtime-to-downtime data:

subject	sessionIndex	screenIndex	index	key1	key2	time
s019	1	3	1	Shift	Shift.t	0.2330
s019	1	3	2	Shift.t	h	0.3116
s019	1	3	3	h	e	0.1392
s019	1	3	4	e	space	0.0811
s019	1	3	5	space	s	0.2738
s019	1	3	6	s	c	0.1869
s019	1	3	7	c	e	0.2397

Holdtime data:

subject	sessionIndex	screenIndex	index	key	time
s019	1	3	1	Shift	0.3442
s019	1	3	2	Shift.t	0.1514
s019	1	3	3	h	0.0844
s019	1	3	4	e	0.1127
s019	1	3	5	space	0.1201
s019	1	3	6	s	0.1409
s019	1	3	7	c	0.1082

The following features will be extracted from the raw data that we collect and from the dataset above:

- Characters per minute
- For each finger on each hand: (dd is the downtime to downtime, from one key to another key)
    - dd for other hand to this finger
    - dd for same hand other finger to this finger
    - dd for same finger to this finger
    - Hold time of keys with that finger
    - Capitalize a letter or hit special char of this finger after normal key from other hand
    - Capitalize a letter or hit special char of this finger after normal key from same hand other fingers
    - Capitalize a letter or hit special char of this finger after capital letter or special character that requires shift key from other hand
    - Capitalize a letter or hit special char of this finger after capital letter or special character that requires shift key from same hand
- Accuracy:
    - 1 - (number of backspaces used / Total number of keystrokes)

- Space bar:
    - dd from any key to space bar
    - Hold time of space bar

And some of the most important features are:

- Characters per minute: Typing speeds differ a lot from person to person.
- Classifying keys based on fingers used to type: Difference in proficiency in using different fingers to type could be used to identify a person.

The outcome will be trying to authenticate a user and give confidence level of the prediction.

Planned Approach

In general, there are two groups of research studies in this field. The first group focused on long and free text strings. They focused more on typing speed and used several distance authentication methods, such as Euclidean and Manhattan distance. Joyce, Rick, and Gopal [1] used 1-norm distance between the data and the reference for the key latency for classification. Monrose, Fabian, and Aviel [2] used a similar technique but added key duration time as another feature.

Tappert et al. [3], Villani et al. [4], Zack et al. [5] are part of the second group in this field. They focused more on various types of input devices. In the experiments used on those studies, two keyboards (desktop and laptop) and two input modes (copy the long text strings and type free text strings) were used. One of the results was that if the same keyboard was used, authentication performance was highly reliable. Otherwise, the error rate increased.

Since most of the features we are extracting are different from the feature used in the research fields, we are going to build our implementation on what is already been done. We will focus more on the impact of different machine learning algorithms, such as KNN, Logistic Regression, SVM, Neural Networks, etc.

The initial steps of our project are going to be as follows:

1. Building keylogger using pynput package in Python using Raspberry Pi to collect raw data. We are trying to be consistent in using only one set hardware.
2. Extracting features from the collected data.
3. Exploring and using different machine learning algorithms.
4. Testing and measuring accuracies and find the best results.

References

[1] Joyce, Rick, and Gopal Gupta. "Identity Authentication Based on Keystroke Latencies."Communications of the ACM, vol. 33, no. 2, Jan. 1990, pp. 168–176., doi:10.1145/75577.75582.

[2] Monrose, Fabian, and Aviel D. Rubin. "Keystroke Dynamics as a Biometric for Authentication."Future Generation Computer Systems, vol. 16, no. 4, 2000, pp. 351–359., doi:10.1016/s0167-739x(99)00059-x.

[3] C.C. Tappert, S.H. Cha, M. Villani, R.S. Zack, A keystroke biometric system for long-text input, Int. J. Inform. Security Privacy 4 (2010) 32–60.

[4] M. Villani, C. Tapert, G. Ngo, J. Simone, H.S. Fort, S.H. Cha, Keystroke biometric recognition studies on long-text input under ideal and applicationoriented conditions, in: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW 2006), New York City, NY, pp. 39–47.

[5] R.S. Zack, C.C. Tappert, S.H. Cha, Performance of a long-text-input keystroke biometric authentication system using an improved k-nearest-neighbor classification method, in: Proceedings of the Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS 2010), Washington, DC, pp. 1–6.