Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 135 (2018) 564–569

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Computer Science and Computational Intelligence 2018, ICCSCI 2018, 07-08 September 2018, Jakarta, Indonesia

Keystroke Dynamic Classification using Machine Learning for Password Authorization

Yohan Muliono[a,*], Hanry Ham[b], Dion Darmawan[b]

[a] *Cyber Security Program, Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*
[b] *Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*

Abstract

Many methods used to perform a password authentication using user's biometrics such as fingerprint recognition, retina recognition, voice recognition, etc. However, additional sensors needed to perform most of biometric recognition methods and will be invasive to users caused by additional tools needed to perform a password authentication. Keyboard Dynamics is one of the solution to perform password authentication without adding any tools which being disruptive to some users. The biometric keystroke dynamic system is relatively unexplored compared to other behavioral authentications discipline. Coupled with the limited number of studies that have been done compared with other biometric systems. Several machine learning research has been conducted but few of them applying deep learning for solving this problem. This research will be focusing in deep learning using optimizer to beat the previous research which using another machine learning techniques. This research shows a better result using optimizer in deep learning resulting in 92.60% accuracy.

Keywords: Machine learning; Deep Learning; Dynamic Keyboard; User Authentication

1. Introduction

In this fast-paced technology era, a security is one of the important issue when it comes to the authentication system. The experts implements various approaches of authentication system: using human's biometric information such as fingerprint, retina, voice and so on. They aim at creating a recognition system by using traditional, machine learning, and deep learning methods. Each systems have their own benefits and drawbacks, currently fingerprint approach its maturity, proved by the commercialized product that are available in the market. However, in order to

[*] Corresponding author. Tel.: +62-21-534-5830 ext 2188 ; fax: +62-21-530-0244.
E-mail address: ymuliono@binus.edu

10.1016/j.procs.2018.08.209

implement such systems, additional sensors are needed as an input reader to the system, the cost of those sensors are vary in the market.

Our proposed method is that to implement an authentication system using keystroke typing behaviour[1]. Keystroke typing behaviour can be defined as the unique characteristic of each user in typing each keystroke provided in the keyboard. We assume that, by only extracting the features from the typing characteristic of each user can be applied as promising authentication system that has a low cost compared to the biometric one due to the system does not require any addition sensors and ease for user to perform it. The usage of features such as keystroke typing behaviours shows that it can identify how many user's hand to type the password, gender, age[2],[3], emotions[4], detecting user's fatigue[5].

In addition to the keystroke typing behaviour approach, some researches have been conducted in the last few years in order to find the best algorithm to perform the authentication task. Moreover, the keystroke dataset may vary in each subject due to inconsistency of human performance each time and time variable, meaning that the same keyword typed in previous may generate different features in the next day due to the variables mentioned above. Dataset used[6] 's work. Furthermore, the dataset used to train our proposed model for classifying the subject based on his/her typing behavior. The datasets already come with a clean data. Thus, this research didn't explain about how to clean the noise of the data and the cleansing process of the data. However, how this research using the dataset to fulfill the experiment will be explained in section experiments and results. This research will be focused on experimenting a deep learning algorithm and try to outperform the conventional machine learning that were conducted in some research Ali et al.[6], Revett et al.[7], Yu and Cho[8]. Nadam optimizer in deep learning will be used in this research. Whereas the goal of an optimizer is to minimize an objective function, which means the difference between the predicted data and the expected values resulted from the experiment. The minimization consists of finding the set of parameters of the architecture that give best results in the targeted tasks such as classification, prediction or clustering.

Nadam is an modification of ADAM momentum that takes advantage of insights from NAG(Nesterov accelerated gradient) which are implementing Nesterov Momentum into Adam Optimizer Dozat[9]. Whereas Adam optimizer research was published back in 2014 Kingma and Ba[10]. This experiment will be conducted using python and using keras optimization library which including Nadam optimizer in the library.

2. Previous Works

Dynamic Keystroke has been researched since long time ago.In 2002 Bergadano et al.[11] conducted a research in dynamic keystroke for user authentication using self-collected dataset from the volunteer, data collected using same text for all individuals and resulting only 0.01% impostor pass the authentication. In 2003 Yu and Cho[8]conducted a preliminary experimental research for feature subset selection in keystroke dynamics identity verification and found that GA-SVM was resulting a good accuracy and learning speed. In 2007, Revett et al.[7] conducted a research about user authentication and starting to research for dynamic keystroke for authentication. The author proposed that biometric is robust, espescially fingerprint, but, biometric can be easily spoofed.

In 2009, Zahid et al.[12] conducted a research about dynamic keystroke in mobile phone to identify users. They used fuzzy classifier particle swarm optimization in the front-end and genetic algorithm for the back-end resulting in distinguishing 3 different features to be used in user identification: *Key hold time*, the time difference between pressing a key and releasing it. *Diagraph time*, the time difference between releasing one key and pressing the next one. *Error rate*, the number of times backspace key is released. Furthermore, those features are trained using 5 classifiers: Baive Bayes, Back Propagation Neural Network (BPNN), Radial Basis Function Network (RBFN), Kstar, J48. The aim of this research is to grant the access whether the user has right to enter the bank account based on PIN (Personal Identification Number) the user inputted.

Epp et al.[4] work on classifying user's emotional states through their keystrokes in the keyboard. Their work aim at classifying these following emotions: confidence, hesitance, nervousness, relaxation, sadness, and tiredness with a promising accuracies ranging from 77% to 88%. In addition, the result of anger and excitement emotional state is way higher, with accuracies of 84%. The dataset used were collected by themself, participants are asked to record their activities periodically in real time rather than emotions induced in the laboratory. There are 15 features introduced in this work, subsequently they are trained using decision trees algorithm.

Research for age-group classification through the typing pattern using machine learning has been conducted in 2018[3]. This research was conducted as one of the approach to distinguish the minor from Internet users using data benchmark from[6] work with Support Vector Machine and Fuzzy Rough Nearest Neighbour as their machine learning approach resulting in 91.2% accuracy.

From the literature reviewed corresponding with this work, most of them are using the conventional machine learning even though the proposed algorithms work excellent however the result can be improved. Therefore, this research proposed deep learning algorithm for competing with conventional machine learning. According to the research found, the result of using deep learning outperformed the conventional one.

3. Material and Methods

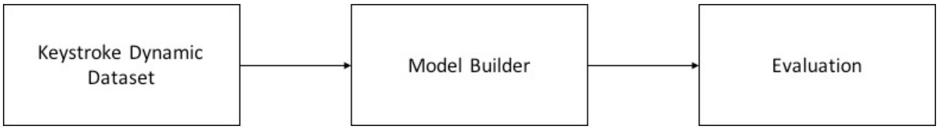

Fig. 1: Keystroke Dynamic Timing

3.1. Keystroke Dynamic Dataset

The biometric keystroke dynamic system is relatively unexplored compared to other behavioral authentications approach. Coupled with the limited number of studies that have been done compared with other biometric systems. Although generally lower than other behavioral authentication biometric system, dynamic keystroke has a number of advantages, such as low cost, transparent, non-invasive to its users, and offers the ability to continue monitoring system[6].

Keystroke dynamic is a field of science to learn a unique timing or pattern that exists in the style of typing from individuals. Some of the information that can be extracted from its characteristics is the time when keys on the keyboard are pressed, keys on the keyboard when lifted, and keystrokes timing from one keypad to another[4]

Dynamic Keystroke refers to one's habit of typing or rhythm when typing on a keyboard. The pattern and rhythm of an individual typing can characterize a person just as a writing style and signatures that could reflect his/her characteristics[13]. The character of a person's neurophysiology can distinguish people in many ways, one of which is from his typing style that makes a person have a unique behavior[14].

In this research, we used[15] dataset that consists of 51 subjects that contains 400 repetitions of a password. The data were acquired in 8 data-collection sessions, each session consists of 50 passwords. A session here refers at least 1 day interval in between session. This aim at capturing some of the day to day variation of each subject's typing. During the session, each subject accomplished the scenario between 1.25 and 11 minutes. Furthermore, 3 features used in this work are depicted in figure 2.

The labels denoted as T1,T2 and T3 from the timing resulted by user input. T1 is recorded as how long one key is pressed, when the user want to input "123" as the password. T1 is how long user pressed 1,2 and 3 in his/her keyboard resulting in 3 data for first label. T2 when the user again want to input "123" as the password, T2 is recorded as

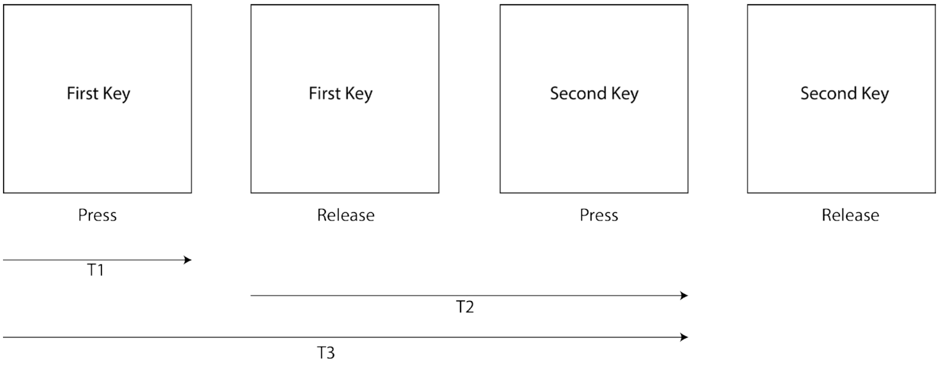

Fig. 2: Keystroke Dynamic Timing

the time between the user release 1 in his/her keyboard and press 2 in his/her keyboard, then release 2 and press 3 in his/her keyboard resulting in 2 data for second label. and for T3 is when user again want to input "123" as the password, T3 is how long the time between the user press 1 in his/her keyboard and press 2 in his/her keyboard, then, how long the time between user press 2 in his/her keyboard and press 3 in his/her keyboard resulting in 2 data for third label. Where the data for each label will be used as input for learning.

3.2. Model Builder

Dataset used are provided in csv format. Therefore some initialization steps are required to used the given features. There are 2 approaches used in this experiments:

- **Support Vectors Machine (SVM)**[16], a statistical method that as well known as supervised learning models that works by separating the hyperplane between the samples. SVM is a well known method that is usually applied into many disciplines that needs machine learning model.
- **Deep Learning**[17] is a computations models that contains of multiple processing layers that able to learn independently. This method has been well known and outperform other traditional machine learning algorithm in processing images, video, speech and audio.

4. Experiments and Results

The data we used in this experiment which are owned by[6] is a dataset of 20400 keystroke timing from typing '.tie5roanl'. The data provided is provided in csv. Coming with label and many variables that already explained in section 3.1. Keystroke Dynamic Dataset. The data for training the algorithm is treated as follows: first, label is separated from another variable used as training variable. second, the data is separated evenly, from 1 persons that typing password 400 times, the data separated to 320 data for training and 80 data for testing, this were applied to all subjects and resulting in 16320 data for training and 4080 data for testing. Lastly, 4080 data used for training will be tested using the model produced by learning phase. We counted how many the algorithm can predict the label perfectly using SVM and deep learning.

Nadam optimizer is used in this experiment since many research proposed nadam is increasing the accuracy of the conducted research. The best learning rate found for those optimizers was 0.002, β_1 was set to 0.9 and β_2 was set to 0.999 as conducted by Kingma and Ba[10].

After initialization of processing raw CSV data, subsequently statistical toolboxes from sklearn library were used to perform SVM with variations of kernel used and keras library to perform deep learning. SVM used in this experiment is divided into three kernels: Linear, RBF and Polynomial. SVM using linear kernel shows the best result with 71.15% accuracy. Deep learning outperform SVM, especially using nadam optimizer can reach around 92.60%. The training and evaluation dataset were divided into 80:20. The details are given in table 1.

Table 1: Result Table

Method	Accuracy
SVM (Linear)	71.15%
SVM (RBF)	30.07%
SVM (Poly)	2.35%
Deep Learning	88.61%
Deep Learning (Using Nadam)	**92.60**%

5. Conclusion

In this work, we proved that the method such as dynamic keystroke is able to identify user authentication in the system. Moreover, by only using 3 features shows a promising result. Later on, we would like to develop more features than can be applied as unique user authentication to increase the accuracy. One of the future works is tuning the variables existing in Nadam optimizer for resulting a better result. There are still a few researchs conducted regarding keystroke dynamic biometric systems compared to other biometrics. Keyboard Dynamics has a number of advantages, one of them is low cost because there is no additional tools and noninvasive to the user because users do not need to use and learn any additional tools to use keyboard dynamic. Keyboard dynamics also offers another features to be explored such as emotional state of the users from the typing behavior. Future research may add psychological features like emotions of users when typing the password for training and testing to boost the accuracy since emotional state of user could be identified from their typing style[4].

This research shows a high accuracy using optimizer in deep learning which are experimental research and resulting in 92.60% accuracy. Moreover, the implementation into real time password authenticating application will also become the future works. This research will continuing to compare the datasets if the users typing is a people who is not fluent typing using keyboard and whether the users is a teenager and an adult.

References

1. Giot, R., El-Abed, M., Hemery, B., Rosenberger, C.. Unconstrained keystroke dynamics authentication with shared secret. *Computers & security* 2011;**30**(6-7):427–445.
2. Idrus, S.Z.S., Cherrier, E., Rosenberger, C., Bours, P.. Soft biometrics for keystroke dynamics: Profiling individuals while typing passwords. *Computers & Security* 2014;**45**:147–155.
3. Roy, S., Roy, U., Sinha, D.. Protection of kids from internet threats: A machine learning approach for classification of age-group based on typing pattern. In: *Proceedings of the International MultiConference of Engineers and Computer Scientists*; vol. 1. 2018, .
4. Epp, C., Lippold, M., Mandryk, R.L.. Identifying emotional states using keystroke dynamics. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM; 2011, p. 715–724.
5. Hayes, J.. Identifying fatigue through keystroke dynamics. *The UNSW Canberra at ADFA Journal of Undergraduate Engineering Research* 2018;**9**(2).
6. Ali, M.L., Monaco, J.V., Tappert, C.C., Qiu, M.. Keystroke biometric systems for user authentication. *Journal of Signal Processing Systems* 2017;**86**(2-3):175–190.
7. Revett, K., Gorunescu, F., Gorunescu, M., Ene, M., Magalhaes, S., Santos, H.. A machine learning approach to keystroke dynamics based user authentication. *International Journal of Electronic Security and Digital Forensics* 2007;**1**(1):55–70.
8. Yu, E., Cho, S.. Ga-svm wrapper approach for feature subset selection in keystroke dynamics identity verification. In: *Neural Networks, 2003. Proceedings of the International Joint Conference on*; vol. 3. IEEE; 2003, p. 2253–2257.
9. Dozat, T.. Incorporating nesterov momentum into adam 2016;.
10. Kingma, D.P., Ba, J.. Adam: A method for stochastic optimization. *arXiv preprint arXiv:14126980* 2014;.
11. Bergadano, F., Gunetti, D., Picardi, C.. User authentication through keystroke dynamics. *ACM Transactions on Information and System Security (TISSEC)* 2002;**5**(4):367–397.
12. Zahid, S., Shahzad, M., Khayam, S.A., Farooq, M.. Keystroke-based user identification on smart phones. In: *International Workshop on Recent Advances in Intrusion Detection*. Springer; 2009, p. 224–243.
13. Zhong, Y., Deng, Y., Jain, A.K.. Keystroke dynamics for user authentication. In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE; 2012, p. 117–123.
14. Patil, R.A., Renke, A.L.. Keystroke dynamics for user authentication and identification by using typing rhythm. *International Journal of Computer Applications* 2016;**144**(9).

15. Killourhy, K.S., Maxion, R.A.. Comparing anomaly-detection algorithms for keystroke dynamics. *Proceedings of the International Conference on Dependable Systems and Networks* 2009;:125–134doi:10.1109/DSN.2009.5270346.
16. Vapnik, V.N.. Statistical Learning Theory. *Adaptive and learning Systems for Signal Processing, Communications and Control* 1998;**2**:1–740. doi:10.2307/1271368. URL `http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471030031.html`.
17. Lecun, Y., Bengio, Y., Hinton, G.. Deep learning. *Nature* 2015;**521**:436–444. doi:10.1038/nature14539.