Two-factor User Authentication Using Biometrics

Viacheslav Liskin [1[0000-0002-9418-0633]], Egor Serdobolskiy [1[0000-0002-6443-4954]] and Iryna Sopilko [2[0000-0002-9594-9280]], Tetiana Okhrimenko[2 [0000-0001-9036-6556]]

[1] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

[2] National Aviation University, Kyiv, Ukraine

liskinslava@gmail.com, serdobolskiy.vik@gmail.com, taniazhm@gmail.com

Abstract. The article is devoted to the biometrics algorithms of authentication in web application. The authors conduct an overview of biometrics algorithms. Described benefit of most popular biometrics' algorithms and two-factor user authentication. Selected one which most prefer for web application and doesn't require additional devices. Modeled this approach and tested it on real data. The modification of the keystroke dynamics algorithm, as the collection of input characteristics on the keyboard during a visit to the site had been proposed. The authors come to the conclusion that the use of a larger set of date for training will improve the algorithm and will be possible to increase the accuracy.

Keywords: authentication, biometrics, keystroke dynamics, machine learning, optimization.

1 Introduction

Biometric data [1] allows to identify and authenticate a person on the basis of a set of unique and specific for him identifiable and verifiable characteristics. Biometric authentication is data comparison on a person's characteristics with that person's biometric characteristics, that are taken to be true, to determine similarities. The peculiarity of this comparison is that these two datasets coincide should be almost identical, but not completely identical. This makes it possible use different methods to improve and compare biometric methods. The reason is that biometrics, even one person, almost cannot match by 100%.

The initial model is first stored in a database. Person who try to be authenticated is "visitor". Stored data is then compared with the biometric characteristics of the "visitor". Because the biometric characteristics are unique to everyone, they cannot be seen or stolen.

In order to be able to identify the individual, the first thing you need to do is to get the data from that user. The data obtained depend entirely on the method by which the user will be identified. For example, for the voice recognition it can be a record of their voices, for the face recognition it can be a photo of their face. These data are

 CybHyg-2019: International Workshop on Cyber Hygiene, Kyiv, Ukraine, November 30, 2019.

then compared with the biometric data which person provide to special device, which can read out the information.

In a fast moving world, biometrics has quickly established itself as the most appropriate means of identification and authentication of individuals, as it is a fast way to use unique biological characteristics.

Today, this technology is used in many areas, such as web applications, specially protected objects, mobile access and others.

Please note that the first paragraph of a section or subsection is not indented. The first paragraphs that follows a table, figure, equation etc. does not have an indent, either.

Subsequent paragraphs, however, are indented.

2 Types of biometric

Biometrics [2] is a science that analyzes the physical or behavioral characteristics inherent in each individual in order to be able to authenticate their identity. To describe biometrics in simple terms, we would describe it in such words as "measurement of the human body". Main characteristics on Fig 1.

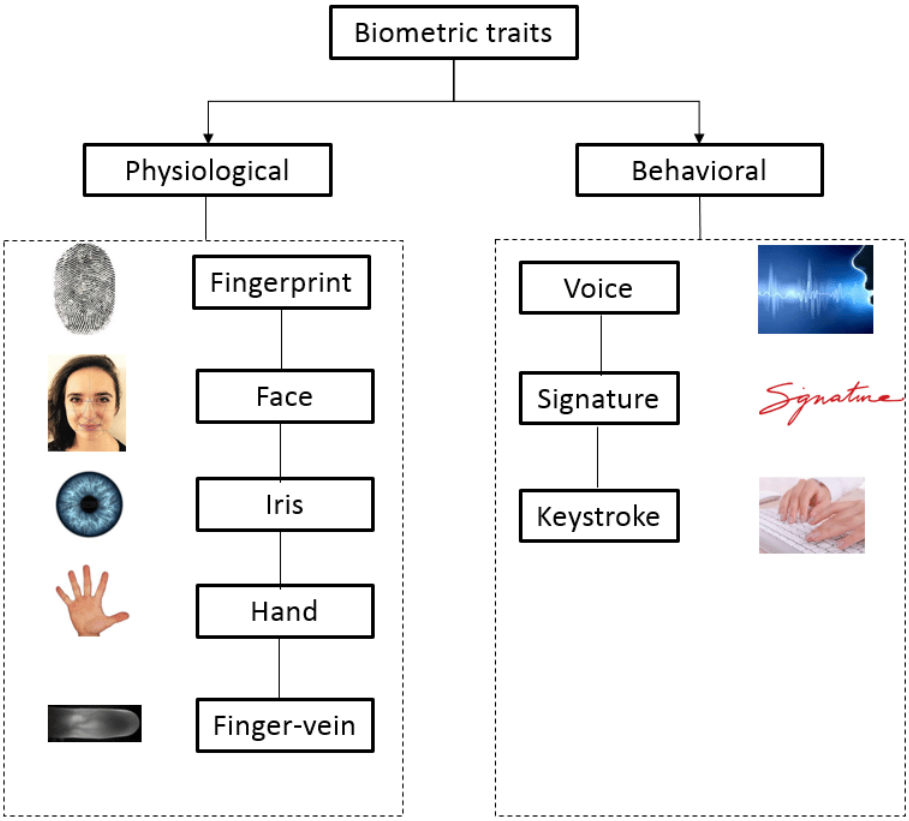

Fig. 1. Physical and behavioral characteristics

There are two categories of biometric technologies: physiological and behavior measurements.

The physiological measurements can be either morphological or biological [3]. They consist mainly of fingerprints, hand, finger, vein, iris and retina, and face shapes for morphological analysis.

The most common behavioral measurements [3] are voice recognition, signature dynamics (pen speed, acceleration, pressure, tilt), key dynamics, how to use objects, gait, step sound, gestures, etc.

The various methods used are the subject of constant research and development and are, of course, constantly being improved. However, not all measurements have the same level of reliability. Physiological measurements are generally considered to have the advantage of remaining more stable throughout a person's life.

3 Biometric authentication

Biometric authentication [4] devices use physical and behavioral characteristics such as fingerprints, facial patterns, iris, keyboard or retina handwriting to verify a user's identity. Biometric authentication is becoming increasingly popular for many purposes, including network logon. As it is hard to always carry with yourself special devices to confirm the authenticity of the identity.

A biometric template or identifier (a sample known to belong to an authorized user) should be stored in a database so that the device can compare it with the new sample received during the login process. Biometric data is often used in combination with smart cards in highly secure environments. The most popular types of biometric devices are:

- Fingerprint scanners
- Facial pattern recognition devices
- Hand geometry recognition devices
- Iris scan identification devices
- Retinal scan identification devices
- Keystroke Dynamics

Behavioral key biometrics [5] uses a manner and rhythm in which individual characters are printed on the keyboard. The rhythms of user key presses are measured to develop a unique biometric template of the user's text set for subsequent authentication. Vibration information can be used to create a template for future use in both identification and authentication tasks.

For web applications on of the best approach of two-step authentication is keystroke dynamics. Mainly, because this method requires only keyboard. Everyone has their own unique handwriting, which is very difficult to forge. The same can be said about keyboard handwriting. The main idea of this method is on Fig 2.

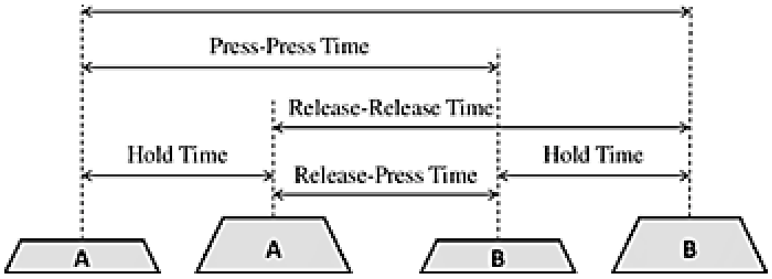

Fig. 2. The main idea of Keystore Dynamic method

4 Keystroke dynamics

The dynamics [5] of key presses usually includes analysis of characteristics such as the duration of key presses or groups of keys and the delay between successive keys, i.e. the time elapsed between one key and the next.

Typically, all keystroke evaluations include (1) a set of subjects to collect data and provide them with input assignments, (2) recording of keystroke times, (3) retrieval of elements suitable for training and testing the classifier, (4) training the classifier using one part of the printed data, and (5) performance testing of the classifier using another part of the printed data.

Before discussing the approaches used by researchers in the dynamics of keystrokes, the features that can be extracted from printed data are described here. When typing, the computer can record the time the key is pressed (waiting time), the time the key is pressed, and the time the key is delayed between the keys of the managers, i.e. the time elapsed between one key and the next. The time measured between key up and the key down is called Flight time. Thus, three timing functions can be derived from the source data: Pressure printing (PP), Release to release (RR) and Release to Press (RP). You can also retrieve other temporary information, such as the time it takes to write a word, a digital graph (two letters) or a three-letter graph (three letters). This item belongs to the category "press to press".

Digital charts contain two consecutive keystrokes, while trigrams contain three; this continues for any number of combinations, resulting in n-charts. Using this terminology, the word "renown" will have three digital charts (“re”,”no”,”wn”) and two trigrams (“ren”,”own”).

The recorded keystroke times are then processed to produce simple templates derived from performance statistics, such as average and standard deviation from the complex pattern recognition algorithm. All this information can be obtained during user input.

4.1 Traditional Benchmarks or Matrices for Keystroke Dynamics

Nowadays, there are many classifiers for keystroke dynamics, so these models are tested based on security parameters [6] such as false acceptance rate (FAR), false rejection rate (FRR) and equal error rate (EER). Graphic of FAR and FRR show on Fig. 3.

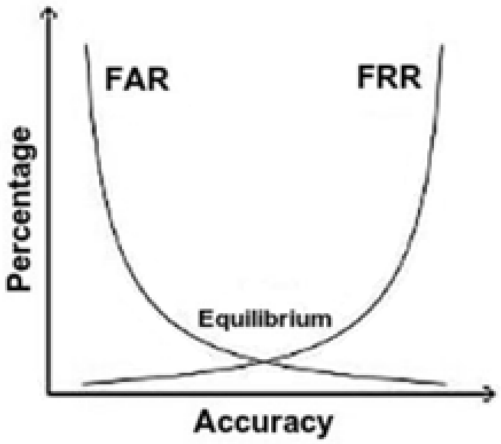

Fig. 3. False Acceptance Rate and False Recognition Rate

FAR – is ratio of the number of false matches divided by the total number of fraud attempts. Thus, FAR shows the number of scammers or impostors who are not exactly allowed to be genuine users.

FRR – is ratio of the number of false refusals divided by the total number of genuine attempts at a match. Thus, FRR gives the number of true users who are denied access to the system. A higher FRR is preferred in higher security systems.

EER is a FAR attitude divided by FRR. A lower EER value means a better system.

4.2 Feature Subset Selection

Nowadays, there are many classifiers for keystroke dynamics, so these models are tested based on security parameters such as false acceptance rate (FAR), false rejection rate (FRR) and equal error rate (EER).

The procedure of generation of subsets is a search procedure within which subsets of candidate characteristics are created on the basis of the evaluation criterion. The evaluation function is used to evaluate the subsets under study, the stop criterion is used to determine when to stop, and the validation procedure is used to check the validity of the subsets. FSS algorithms [7] are divided into three categories, and this classification is based on different evaluation criteria, namely: filter model, wrapper model and hybrid model. In all categories, the algorithms can be further differentiated

by how the space of a subset of features is studied and the exact nature of their evaluation function.

The filter model [8] is required to evaluate and select subsets of elements without using any learning algorithm. Sometimes the required set of functions is not selected or the filtering method does not allow selecting the required set of functions if the criterion used deviates from the criterion used for training the training machine. Another disadvantage of the filtering model is that the filtering approach may also not find a subset of parameters that could jointly maximize the criterion, since most filters evaluate the significance of each parameter only by evaluating one parameter at a time. Thus, the quality of training models deteriorates.

The wrapper model [9] is required to evaluate and select subsets of elements without using any learning algorithm. Sometimes the required set of functions is not selected or the filtering method does not allow selecting the required set of functions if the criterion used deviates from the criterion used for training the training machine.

The hybrid model [10] takes advantage of both the filter and the wrapper model, using their different evaluation criteria at different stages of the search. Hybrid methods are more effective because they combine the advantages of winding and filtering, because they do not allow you to retrain the predictor from scratch for every subsets of characteristics under study. However, they are very complex and limited to a specific training machine.

4.3 Classification

After the extraction of the feature and the selection of the feature, the next step is the classification step, which compares the saved template with the sample provided during the session. There are various methods used for classification. These classification algorithms are divided into four main categories [11]: Statistical Algorithms, Artificial Neural Networks, Pattern Recognition and learning based algorithms, Search heuristics and combination of algorithms.

Statistical approaches calculate the average, standard deviation of characteristics in the template. Distance measurement methods such as Euclidean distance, weighted Euclidean distance, Manhattan distance, etc. are used to compare the training data set with the test data set. There is no need for the data collected for authentication and verification at the push of a button to be linear, so sometimes these linear statistical approaches do not produce good results.

Thus, there is a need for some approaches that use probabilistic data rather than deterministic data. Other statistical methods can also be used for classification, such as the decision tree, Bayes classification (based on a hindsight probability), etc. In addition, the Montecarlo method [12] can be used for the dynamics of keystrokes and thus achieve an average false alarm rate of 9.62% and an average acceptance rate of 0.88%. Another approach used for classification purposes is the use of an artificial neural network.

Ting-Yi Chang [14] used ANN technology and keystroke capabilities to dynamically generate a long-lived private key, and found out that if the password was opened, the probability of hacking the private key would decrease. The advantage of this approach is that it can handle many of the parameters and thus gives good results. This method shown on Fig. 4.

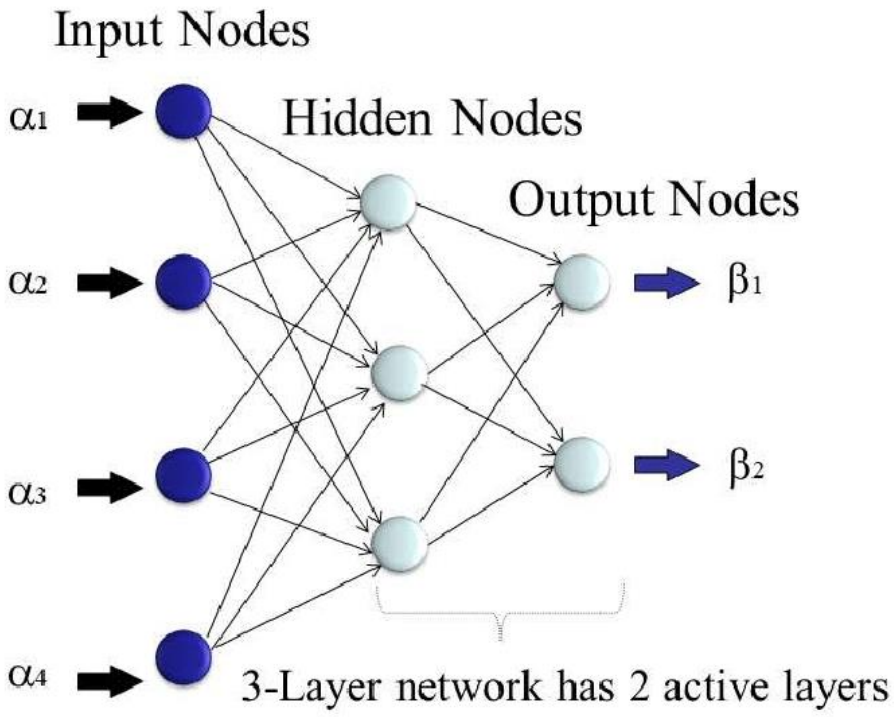

Fig. 4. Base ANN algorithm

Image recognition is defined as the act of obtaining initial data (samples, objects) and their classification into different categories based on algorithms. Pattern recognition includes machine learning algorithms, various classification methods, such as the closest neighbour, Bayesian classifier, and support vector machine, clustering methods such as K-means, etc.

Hyoung-joo Lee [13] used SVM, and it was noticed that retraining increases the efficiency of authentication and that quantization of vector learning to detect novelties surpasses other widely used novelty detectors. This method shown on Fig. 5.

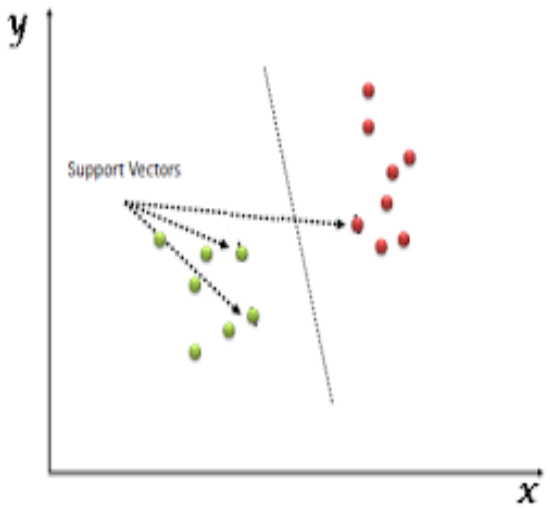

Fig. 5. Base SVM algorithm

The fourth approach usually includes evolutionary algorithms, such as genetic algorithm, Ant colony optimization, Particle swarm optimization etc. The advantage of using these evolutionary methods is that they can work with large databases. Ant colony optimization method shown on Fig. 6.

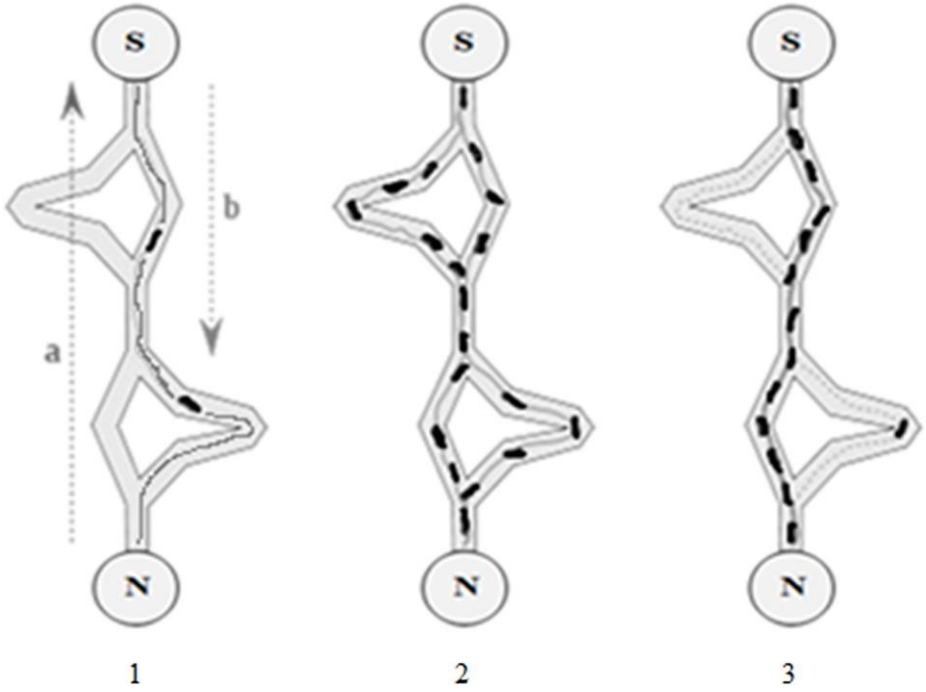

Fig. 6. Base ant colony optimization algorithm

5 Conclusions

In the course of this work, a two-factor authentication simulation was developed in a web application. The keystroke dynamics was chosen. This method allows to very accurately determining the identity of the person who enters the username and password without special devices. BeiHang Keystroke Dynamics Database" was selected as the data set for training. Further, the results were checked on the “Stonybrook Keystroke Patterns as Prosody in Digital Writings” of this data set. The algorithm has shown good results. On average, the accuracy was 99.5%.

Also, to improve this algorithm it will be possible to use a larger set date for training, which can increase the accuracy even more. Also, a modification of this algorithm is the collection of input characteristics on the keyboard during a visit to the site. This will increase the accuracy of data taken for the truth for the user.

References

1. “Biometrics: authentication & identification” – 2019, https://www.gemalto.com/govt/inspired/biometrics last accessed 2019/10/20
2. Paul Benjamin Lowry, Jackson Stephens, Aaron Moyes, Sean Wilson, and Mark Mitchell (2005). “Biometrics, a critical consideration in information security management”, in Margherita Pagani, ed. Encyclopedia of Multimedia Technology and Networks, Idea Group Inc.
3. Jain, Anil K.; Ross, Arun (2008). “Introduction to Biometrics”. In Jain, AK; Flynn; Ross, A. Handbook of Biometrics. Springer. pp. 1–22. ISBN 978-0-387-71040-2. 09.03.2011
4. “What is Biometrics?”. Biometrics Research Group. Michigan State University. 27.08.2017.
5. Kasem Wangsuk and Tanapat Anusas-amornkul, Trajectory Mining for Keystroke Dynamics Authentication. in Procedia Computer Science 24, 2013.
6. Ziyandinov A.I. Principles of building biometric authentication systems / Ziyandinov A.I. – M.: MIPT, 2005.
7. M. Dash and H. Liu, Feature Selection for Classification‖, in Intelligent Data Analysis 1, 1997.
8. Jinjie Huang, Yunze Cai and Xiaoming Xu, A hybrid genetic algorithm for feature selection wrapper based on mutual information In Pattern Recognition Letters 28, 2007
9. M. Hall, Correlation based feature selection for machine learning, Doctoral dissertation, University of Waikato, 1999.
10. M.E ElAlami, A filter model for feature subset selection based on genetic algorithm‖, in Knowledge-Based Systems 22, 2009.
11. Salil P. Banerjee, Damon L. Woodard, Biometric Authentication and Identification using Keystroke Dynamics: A Survey‖, Journal of Pattern Recognition Research 7, pp. 116-139, 2012.

12. Yong Sheng, Vir V. Phoha and Steven M. Rovnyak, A Parallel Decision Tree-Based Method for User Authentication Based on Keystroke Patterns‖, IEEE Transactions On Systems, Man, And Cybernetics – Part B: Cybernetics, vol. 35, No. 4, 2005, pp. 826-833
13. Alon Schclar, Lior Rokach, Adi Abramson, and Yuval Elovici, User Authentication Based on Representative Users‖, IEEE Transactions on systems, man, and cybernetics –part c: applications and reviews, vol. 42, №. 6, pp. 1669-1678, 2012.
14. Ting-Yi Chang, Dynamically generate a long-lived private key based on password keystroke features and neural network, Information Sciences 211, pp. 36-47, 2012.
15. Dychka, I., Tereikovskyi, I., Tereikovska, L., Pogorelov, V., Mussiraliyeva, S. Deobfuscation of computer virus malware code with value state dependence graph // Advances in Intelligent Systems and Computing. 2018. Vol. 754, pp 370-379.
16. Aitchanov, B., Korchenko, A., Tereykovskiy, I., Bapiyev, I. Perspectives for using classical neural network models and methods of counteracting attacks on network resources of information systems., News of the national academy of sciences of the republic of Kazakhstan series of geology and technical sciences. Vol. 5, № 425 (2017), pp. 202-212.