keystroke dynamics.
ensemble of classifiers,
biometrics

Rafal DOROZ[1], Piotr PORWIK[1], Hossein SAFAVERDI[1]

PERSON VERIFICATION BASED ON KEYSTROKE DYNAMICS

This paper presents a new multilayer ensemble of classifiers for users verification who use computer keyboard. The special keyboard extracts the key pressure and latency between keyboard keys pressed during password entered. When user is typing password the system creates a pattern based on time and key pressure. For users verification group of classifiers have been proposed. It allows to obtain the higher accuracy level compared to alternative techniques. The efficiency of the proposed method has been confirmed in the experiments carried out.

1. INTRODUCTION

Computers and internet have become a ubiquitous part of our lives. Since we depend so much on computers to store data and personal information, it has become more necessary to secure our information from intruders. Many networks are vulnerable for intruder attacks, because of low security. A password has been a solution for user authentication and identification in computer based applications [4],[17],[18]. Although password-based authentication and identification systems have many benefits they become unprotected when intruders and imposters try to log in to the system by means of valid password instead of valid user [6]. The system cannot recognize a valid user from an intruder who receives the log-in information illegally or even unintentionally. In order to overcome password-based authentication vulnerability, in our approach the two-stage authentication systems have been introduced. In practice many biometrics systems are exploited on the computer market, such as face recognition, fingerprints and signature analysis or gait dynamics [2],[5],[10],[12]. Unfortunately all of these behavioral biometric features require additional tools and special devices which increase in cost [14],[15],[16]. Instead of mentioned techniques we propose keystroke analysis. It is especially useful for users who work in the computer world environment.

In this work we propose a new structure of multilayer ensemble of classifiers for keystroke dynamic verifications. In order to obtain the best verification results, classifiers employ keystroke timing information, including the delay between and duration of each tap of a key as well as values of key pressure. Based on latency and pressure, a unique keystroke pattern for each person can be formed. These patterns comprise input data for ensemble of classifiers to verify whether a given user is genuine or not. The general structure of proposed classifiers depicted in Fig. 1.

[1] University of Silesia, Institute of Computer Science, Bedzinska 39, 41-200 Sosnowiec, Poland

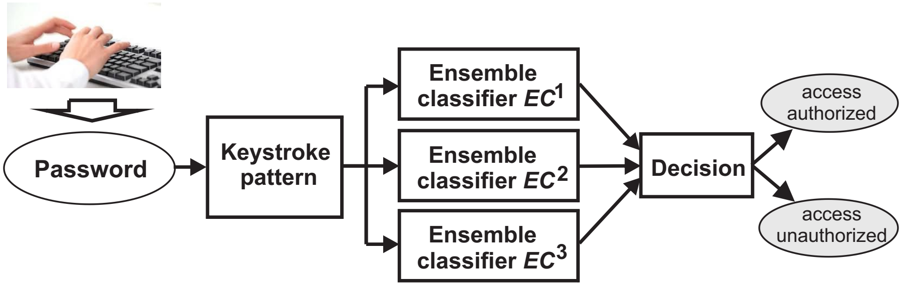
Password
Keystroke pattern
Ensemble classifier EC^1
Ensemble classifier EC^2
Ensemble classifier EC^3
Decision
access authorized
access unauthorized

Fig. 1. The general scheme of proposed classifiers ensemble.

2. Keystroke pattern extraction

In the first stage, length n of users password is established. Classifier inputs data $\mathbf{V}$ consists of the two sub-vectors $\mathbf{L}$ and $\mathbf{P}$, hence $\mathbf{V} = [\mathbf{L},\mathbf{P}, avg]$. The sub-vector $\mathbf{L}$ includes elements associated with delay between and duration of each tap of a key, whereas the vector $\mathbf{P}$ includes elements which reflect the values of key pressure. For such assumptions we formed both the sub-vectors:

$$\mathbf{L} = [l_1, l_2, ..., l_{n-1}], \qquad \mathbf{P} = [p_n, ..., p_{2n-1}] \quad \text{and} \qquad avg = \frac{1}{n}\sum_{i=n}^{2n-1} p_i. \tag{1}$$

In practice, key pressure values are sampled. In our case sampling rate is 100 samples/sec. Typing process is a behavioral feature, therefore each user can type a same password with different time period. During typing action a lot of pressure values are recorded. When user type slowly, then many pressure values are registered, while type fast then only a few of such values are gathered. From eq. (1) follows that length of the vector $\mathbf{P}$ should be a constant, therefore recorded data have to be adopted to the mentioned constant length of the vector $\mathbf{P}$. It can be explained by means of the following example. Let $y = f(t)$ be a function of the two variables, see Fig. 2. This function have some maximal values (peaks). We will search all maximal values which fulfill of the two conditions:

1) $|y_{i+1} - y_i| < y_{\min}$,
2) $|t_{i+1} - t_i| < t_{\min}$,

where t_{min} is the minimal horizontal time distance between two neighbor peaks, y_{min} is the threshold which determines how a given peak value should be greater than value of its neighbors from the left or right side, if these peaks are exist.
Idea of such measurements is presented in Fig. 2.

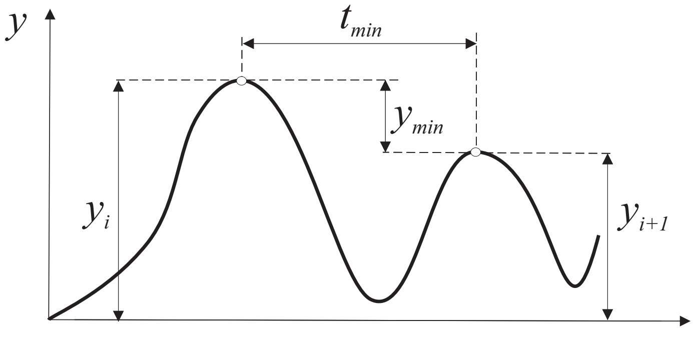

Fig. 2. Exemplary plot with marked peaks (by circle) and parameters.

In the next stage among all indicated peaks, the n highest are selected. If number of detected peaks is less than n, the missing value(s) is supplemented by zero value. Fig. 3 presents the examples of the two pressure functions where $n = 7$ peaks (marked with asterisk).

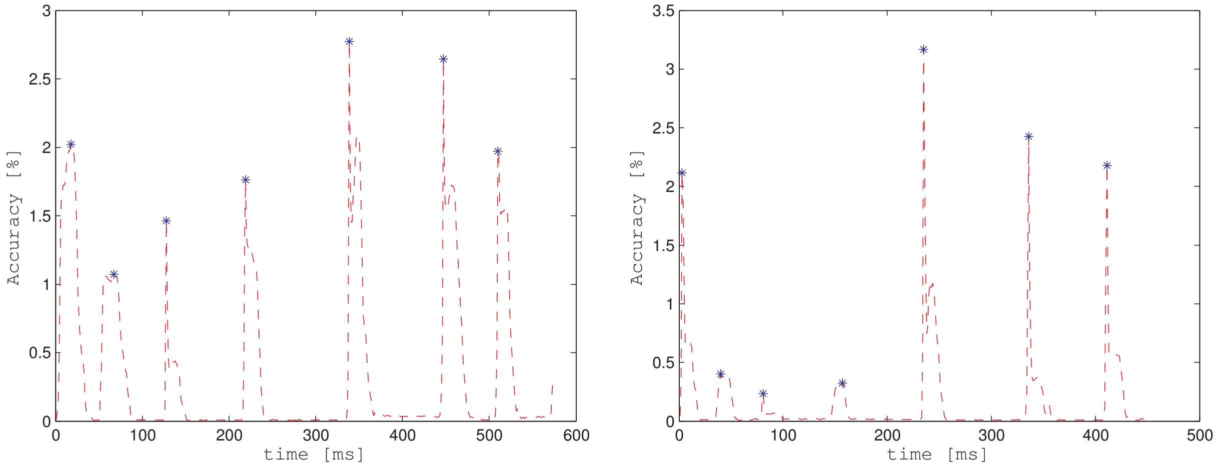

Fig. 3. Selected peaks from two exemplary time series of the pressure function.

3. Ensemble of classifiers

Keystroke pattern for each password is then subjected to classification. In our method, in order to classify the password three ensembles EC^1, EC^3 and EC^3 have been created. Each single ensemble consists of four sub-classifiers. It means that proposed group of ensembles consists of twelve classifiers which work in the parallel mode. Lets assume that $c_1^1, c_2^1, c_3^1, c_4^1$ belong to the ensemble EC^1. Similarly for other classifiers we have: $c_1^2, c_2^2, c_3^2, c_4^2$ belong to EC^2 and $c_1^3, c_2^3, c_3^3, c_4^3$, belong to EC^3. The general structure of the first ensemble of classifiers is shown in Fig. 4.

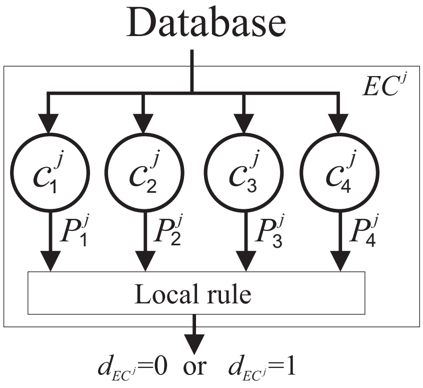

Fig. 4. Structure single ensemble of classifiers.

Each three ensemble of classifiers is learned based on different learning set DS_1^k, DS_2^k andDS_3^k, respectively. DS set contains the vectors $\mathbf{V}$.

$$\begin{gathered} DS_1^k = O^k \cup F^a, \quad DS_2^k = O^k \cup F^b, \quad DS_3^k = O^k \cup F^c, \\ a \neq b \neq c \neq k \quad and \quad k = 1, ..., N, \end{gathered} \tag{2}$$

where N is a total number of people inside database, set O^k consists of ten vectors $\mathbf{V}$, which were created based on ten passwords of k-th person. The sets F^a,F^b,F^c contain of ten vectors

$\mathbf{V}$ for the a-th, b-th and c-th persons respectively. These persons were randomly selected. It has to be mentioned that the person k must be different than persons a, b and c.

In verification mode every i-th classifier c_i^j of the ensemble EC^j calculates a probability $P_i^j \in [0, 1]$ that a given password belongs to the genuine or forged passwords. Each ensemble of classifiers consists of four classifiers and each of them returns its probability P_i^j. In the next step each ensemble of classifiers EC^j, where $j = 1, ..., 3$, makes a decision, based on probability P_i^j. Let S_g^j be a total probability P_i^j generated by j-th classifier, while S_f^j be a probability that a given user password is forged. For such assumptions, we have:

$$S_g^j = \sum_{i=1}^{4} P_i^j, \qquad S_f^j = \sum_{i=1}^{4} \left(1 - P_i^j\right), \tag{3}$$

where P_i^j is probability returned by i-th classifier of the ensemble EC^j. Finally, every ensemble EC^j, $j = 1, ..., 3$ creates a decision based on the following formula:

$$\begin{array}{llll} if & S_g^j > S_f^j & then & d_{EC^j} = 1 \\ otherwise & & & d_{EC^j} = 0 \end{array}, \tag{4}$$

where d_{EC^j} is an auxiliary variable which helps in decision creation. The ultimate decision of global ensemble classifier is made according to decision of all single of ensembles EC^j:

$$\begin{array}{llllll} if & \sum\limits_{j=1}^{3} d_{EC^j} \geq 2 & then & \text{user} & \text{is} & \text{legitimated} \\ otherwise & & & \text{user} & \text{is} & \text{illegitimated} \end{array}. \tag{5}$$

4. EXPERIMENTAL RESULTS

The database which was used in experiments is so-called ”try4-mbs” and it is publicly available [9],[8]. The database consists of 100 users who typed 10 times the same password ”try4-mbs”, so the database comprises of 1000 records. To collect the data in this database a special keyboard system with pressure sensors adhered underneath the keys was used. Thanks to this and a special program the database comprises of two features: latency between keys and pressure value of every key [9],[8]. Efficiency of the classifier was confirmed in the 10-fold cross validation test.

In practice password in the database was ”try4-mbs”, so number of password letters was $n = 8$. Therefore the vector $\mathbf{V}$ contains $2n = 16$ coordinates. The method proposed in this paper is continuation of the previous work [3] in which only latency feature was used to determine the vector $\mathbf{V}$. The results obtained by means of proposed approach is shown in Table 1.

Table 1. The best results obtained by use of only latency [3].

Ensemble classifier members		**Accuracy [%]**
c_1^j	Kstar [19]	98.4
c_2^j	BayesNet [19]	
c_3^j	LibSVM [1]	
c_4^j	HoeffdingTree [7]	

Whether apart from latency also pressure values have been included, the effectiveness of the method is higher. It presents Table 2. During our investigation the parameters t_{min} and y_{min} have been changed in wide range: $t_{\min} \in \{10, ..., 50\}$ with step 10, $y_{\min} \in \{0, ..., 40\}$ with step 10. From carried out experiments the following parameters $t_{\min} = 10$ and $y_{\min} = 30$ have been established. It provides the best classification level.

Table 2. The best results obtained by use of pressure and latency.

Ensemble classifier members		Accuracy [%]
c_1^j	Kstar [19]	99.8
c_2^j	BayesNet [19]	
c_3^j	LibSVM [1]	
c_4^j	HoeffdingTree [7]	

Comparison of the results gathered in Table 1 and Table 2 show that classifying accuracy is higher when input data are based on latency and pressure values.

It is difficult to compare the works of different researchers due to lack of standards for data preparation and collection [11]. Same standards could facilitate the exchange of information amongst researchers and provide a better way to compare different algorithms. Using databases which are publicly available allow us to compare different algorithms and classification techniques. Comparison of the various results based on the same database is depicted in Table 3.

Table 3. Classification accuracy for different verification techniques.

Method	Accuracy [%]
Our approach	99.8
Method: average fuzzy artmap [9]	98.78
Method: Average FMM [13]	98.32
Method: Voting FMM [9]	99.3
Method ARTMAP-FD [8]	89.3

5. Conclusions

In this work we proposed multilayer ensemble of classifiers to recognize the computer users by means of keystroke dynamics measurement. We conducted classification based on latency and pressure features. Adding pressure in classification mode led to increase the accuracy of classification in comparison to classification based only on latency. Obtained results suggest that described method could be used in professional biometric applications.

In future investigations we are going to check various databases and modify the structure of the ensembles and classifiers in order to obtain the higher accuracy of classification.

Acknowledgment

This work was supported by the Polish National Science Centre under the grant no. DEC-2013/09/B/ST6/02264.

Bibliography

[1] Chang C.-C., Lin C.-J. LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol., May 2011, Vol. 2. ACM, New York, NY, USA, pp. 27:1–27:27.

[2] Doroz R., Porwik P. Handwritten signature recognition with adaptive selection of behavioral features. Computer Information Systems Analysis and Technologies, 2011, Vol. 245 of Communications in Computer and Information Science. Springer Berlin Heidelberg, pp. 128–136.

[3] Doroz R., Porwik P., Safaverdi H. The new multilayer ensemble classifier for verifying users based on keystroke dynamics. Computational Collective Intelligence - 7th International Conference, ICCCI 2015, Madrid, Spain, September 21-23, 2015, Proceedings, Part II, 2015. pp. 598–605.

[4] GUVEN A., SOGUKPINAR I. Understanding users' keystroke patterns for computer access security. Computers & Security, 2003, Vol. 22. pp. 695–706.

[5] IDRUS S. Z. S., CHERRIER E., ROSENBERGER C., BOURS P. Soft biometrics for keystroke dynamics: Profiling individuals while typing passwords. Computers & Security, 2014, Vol. 45. pp. 147 – 155.

[6] KANG P., CHO S. Keystroke dynamics-based user authentication using long and free text strings from various input devices. Information Sciences, 2015, Vol. 308. pp. 72–93.

[7] KIRKBY R. Improving hoeffding trees. 2007.

[8] LOY C. C., LAI W. K., LIM C. P. Keystroke patterns classification using the artmap-fd neural network. Intelligent Information Hiding and Multimedia Signal Processing, 2007. IIHMSP 2007. Third International Conference on, 2007, Vol. 1. pp. 61–64.

[9] LOY C. C., PROF A., CHEE D., LIM P., LAI K., BERHAD M. 2005): ”pressure-based typing biometrics user authentication using the fuzzy ARTMAP neural network. Proc. of the 12th Int. Conf. on Neural Information Processing. pp. 647–652.

[10] MONROSE F., RUBIN A. D. Keystroke dynamics as a biometric for authentication. Future Generation Computer Systems, 2000, Vol. 16. pp. 351–359.

[11] PANASIUK P., SAEED K. Influence of database quality on the results of keystroke dynamics algorithms. Computer Information Systems Analysis and Technologies, 2011, Vol. 245. Springer Berlin Heidelberg, pp. 105–112.

[12] PORWIK P., DOROZ R., ORCZYK T. The k-NN classifier and self-adaptive hotelling data reduction technique in handwritten signatures recognition. Pattern Anal. Appl., 2015, Vol. 18. pp. 983–1001.

[13] QUTEISHAT A., LIM C. P., LOY C. C., LAI W. K. Authenticating the identity of computer users with typing biometrics and the fuzzy Min-Max neural network BIOMETRICS AND ITS APPLICATIONS). Biomedical fuzzy and human sciences : the official journal of the Biomedical Fuzzy Systems Association, jan 2009, Vol. 14. Biomedical Fuzzy Systems Association, pp. 47–53.

[14] RYBNIK M., PANASIUK P., SAEED K. User authentication with keystroke dynamics using fixed text. Biometrics and Kansei Engineering, 2009. ICBAKE 2009. International Conference on, Jun 2009. pp. 70–75.

[15] RYBNIK M., PANASIUK P., SAEED K., ROGOWSKI M. Advances in the keystroke dynamics: The practical impact of database quality. Computer Information Systems and Industrial Management, 2012, Vol. 7564 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 203–214.

[16] RYBNIK M., TABEDZKI M., SAEED K. A keystroke dynamics based system for user identification. Computer Information Systems and Industrial Management Applications, 2008. IEEE-CISIM, Jun 2008. pp. 225–230.

[17] SUNG K.-S., CHO S. GA SVM wrapper ensemble for keystroke dynamics authentication. Advances in Biometrics, 2005, Vol. 3832 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 654–660.

[18] TEH P. S., TEOH A. B. J., TEE C., ONG T. S. Keystroke dynamics in password authentication enhancement. Expert Systems with Applications, 2010, Vol. 37. pp. 8618 – 8627.

[19] VIJAYARANI S., MUTHULAKSHMI M. Comparative analysis of bayes and lazy classification algorithms. International Journal of Advanced Research in Computer and Communication Engineering, 2013, Vol. 2. pp. 3118–3124.