IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-3, Special Issue-1 E-ISSN: 2347-2693

Text-based Analysis of Keystroke Dynamics in User Authentication

Soumen Roy*
Dept. of Computer Sc. & Engg.
University of Calcutta
Calcutta -700 009, INDIA
soumen.roy_2007@yahoo.co.in

Utpal Roy
Dept. of Computer & Sys.Sciences,
Visva-Bharati, Santiniketan
Pin -731235, INDIA
roy.utpal@gmail.com

D. D. Sinha
Dept. of Computer Sc. & Engg.
University of Calcutta
Calcutta -700 009, INDIA
devadatta.sinha@gmail.com

***Abstract*—User Authentication process is the essential integral part of any secure or collaborative system. Where, Knowledge-based Authentication is convenient and low cost among currently used authentication processes. But, today, password or PIN is not limited in knowledge-based user authentication due to off-line guessing attacks; it demands higher level of security and performance together with low cost. Here, Keystroke Dynamics is the best possible solution, where the users are not only indentified by their password or PIN, their regular typing style is also accounted for. But this technique, as is now, suffers from accuracy level and performance. Thus, in order to realize this technique in practice a higher level of security and performance together with low cost version is demanded with an error to an accepted level. Hence, it is highly needed to identify the controlling parameters and optimize the accuracy and performance. In this paper we investigated typing style of 15 different individuals with 3 different texts and analyzed the collected data. Here, we introduced some effective factors which can optimize the accuracy and performance and at the end, we concluded by suggesting some future plans that also can be effectively implemented by this technique.**

Keywords- Keystroke Dynamics, Behavioral biometric, Computer Security, Manhattan Distance, Euclidean Distance, Mahanobolis Distance, Z Score, EER, FAR, FRR, Knowledge-based Authentication

I. INTRODUCTION

According to SplashData (who gather data from millions of stolen passwords posted online), the top three passwords in the year 2013 are "123456", "password" and "12345678". So we can say most of the people are uninspired while choosing a healthy password because, we, as people are still very lazy. It increases the probability of guessing attacks. To minimize these attacks, we pick up some word for password from relatively small dictionary and decorate it by adding extra text or combine capital, small letter with some symbols. It increases the complexity of password which is difficult to remember and generally, we forget to distinguish this type of healthy passwords for different access control systems. The use of password in Knowledge-based user authentication techniques is risky because of not only the possibility of off-line guessing attacks. The password is also risky as pressing the password

in public place. It is unsafe while all around the areas; class room, office, bank, college campus, railway station are covered by video or spy cameras. It is also unsafe if we pick up a word from a relatively small dictionary for password that may content our personal information. An attacker may collect our personal information and can check one by one until the actual result is obtained. So there is a probability of Brute force attack or shoulder surfing attack. Further it has been established that none of these technique is self-sufficient for the security purpose as per Hafiz, Z. U. K.

To solve these problem, here, in this paper we investigated a simple text-based, inexpensive biometric user authentication through keystroke dynamics. Keystroke Dynamics is a technology to segregate and distinguish people based on their typing rhythms. Here users are not only identified by their corresponding user ID and password, but their typing style is also accounted for. It has been observed that Keystroke Dynamics possesses with each human being is unique like same nature of our hand writing as per Gaines, Bleha and Killourhy.

Keystroke Dynamics as biometrics characteristics is not a new one. Keystroke Dynamics was first formally investigated by Bryan and Harter in 1897 as part of a study on skill gaining in telegraph operators. In 1975 Spillane suggested in an IBM technical bulletin that typing rhythms might be used for identifying the user at a computer keyboard. That bulletin described keystroke dynamics in concept. Forsen et al. in 1977 conducted preliminary tests of whether keystroke dynamics could be used to distinguish typists. Gaines et al. in 1980 produced an extensive report of their investigation with seven typists into keystroke dynamics. After then S. Bleha submitted his PhD thesis on Recognition system based on keystroke dynamics in 1988. R. Joyce and G. Gupta proposed an identity authentication based on keystroke latencies in 1990. F. Monrose et al. proposed keystroke dynamic as a biometric for authentication in 2000. Different online and offline applications already have been done by fixed text and free text keystroke dynamics. Keystroke dynamics research has been going on for the more than thirty three years. Many methods have been proposed during that time. Methods based on traditional statistics-such as mean times and their standard deviations-are common. Over the years, different pattern recognition methods have come into vogue and been applied to keystroke dynamics; neural networks, Fuzzy logic and support vector machines among others. We often used two features Dwell time and Flight time as biometrics features, Dwell time which refers to the amount of time between pressing and releasing a single key and Flight Time which refers to the amount of time between pressing and releasing two successive keys.

In this paper fixed-text typing style has been considered as a key issue of a security system. Recognizing typing style promises a parameter like biometric characteristics that may facilitate non-intrusive, cost-effective and continuous monitoring. But this technique, as is now, suffers from accuracy level, performance. Thus, in order to realize this technique in practice a higher level of security and performance together with low cost version is demanded with an error to an accepted level. Hence, it is highly needed to identify the controlling parameters and optimize the accuracy, performance as well as cost with new algorithms.

Here, System takes comprising of characters as well as the typing style of each subsequent character entered. It facilitates that no one can track the time or presses the character of password in same rhythm. It will prevent our system from off-line guessing attracts and also prevent to track by unauthorized people. In this paper our main objectives are finding out the optimizing issues in keystroke dynamics and try to optimize those issues, finding out the effective factors in keystroke dynamics and design new algorithms to compare with exiting algorithms and then performance based analysis with different factors for comparison and discussing different application areas where keystroke dynamics is suitable technique. The fixed-text of our proposed system uses the features of the password extracted from existing knowledge-based user authentication technique (comprising of characters) together by taking into account the duration of the depressed characters (dwell time) and pause duration between each subsequent characters (flight time) entered which is defined by Joyce, Monrose and Roy et al. A laboratory made sample password database has been used to train the system at the time of enrolment. Here system calculates the dwell times and flight times for each sample, then finds out the actual dwell and flight time by applying statistical methods on dwell and flight times of all samples of the fixed-text database. Features Mining mechanism such as Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Neural Networks explained by Harun, N. et al., any distance measurement mechanism such as Euclidean distance, Non-Weighted Probability, Weighted Probability Measure which is nicely explained by Monrose, F. and Rubin, A. D., Manhattan distance, Mahalanobis distance as per Killourhy and Shimaa, I. H. et al., Bhattacharyya Distance as per Janakiraman, R. & Sim, T., or Genetic algorithm,

particle swan optimization which is explained by Marcus, K. and Akila, M. and system decides whether the user is valid. Thus we can minimize the probability of any off-line guessing attacks since rhythm of the fixed-text is used as safeguard instead of original text for the password, which cannot be copied even after watching it several times. The rhythm of the fixed text as it is entered is used to validate the authenticity of the user rather than password. It updates itself continuously by Growing Window, Moving Window or Adaptive threshold mechanism, defined by Pin, S. T., which can help to recover the account and minimize Equal Error Rate (EER) in future.

II. KEYSTROKE DYNAMICS

A. *Basic Idea*

Keystroke dynamics is a behavioral biometrics which is the method of analysing the way a user types on a keyboard and classify him based on his regular typing rhythm. It is the study of whether people can be well-known by their typing rhythms, much like handwriting is used to recognize the author of a written text. A user's typing pattern may be unique because similar neuro-physiological factors that make written signatures unique.

B. *Science and Features Selection*

Our typing style can be easily calculated by simple program which can calculate key pressing and releasing time of each key and then generates key-hold time and key latency times. Let K_i represent entered character set and P_i and R_i represent the corresponding key press and key release time where 1<=i<=length of the entered word. The features of the keystroke dynamics as follows:

Key Duration (T_1) = R_i-P_i	(1)
Up Up Key Latency (T_2) = R_{i+1}-R_i	(2)
Down Down Key Latency (T_3) = P_{i+1}-P_i	(3)
Up Down Key Latency (T_4) = P_{i+1}-R_i	(4)
Down Up Key Latency (T_5) = R_{i-1}-P_i	(5)
Total Time Key Latency (T_6) = R_n-P_1	(6)
Di-graph Latency (T_7) = R_{i+1}-P_i	(7)
Tri-graph Latency (T_8) = R_{i+2}-P_i	(8)

Some new features also can be calculated like key pressure (Pressure sensitive keyboard is require), finger tips size (Touch screen keypad is needed), finger placement on keyboard (Camera is needed), keystroke sound (microphone is needed), error correcting mechanism, sequence of left-right control keys.

C. *Keystroke Dynamics as User Authentication*

There are different ways in which a user can be authenticated. However all of these ways can be categorized into one of three classes: "Something we know" e.g. password, "Something we have" e.g. token, "Something we are" e.g. biometric property. The keystroke Dynamics characteristics is the behavioural biometric characteristics what we have learned in our life not what the properties we are born with which is the good human characteristics that can be used to distinguish people.

D. *Security Issues*

Among various user authentication techniques knowledge-based, token-based and biometric-based authentication techniques, biometric authentication is most popular for their uniqueness characteristics and cannot be stolen or there is no chance to loss. Keystroke Dynamics is a behavioral characteristic which is unique and can be effectively implemented with the existing system with minimal alternation. It can be used as a safeguard of our password from different type of attacks.

E. Factors Affecting Performance

Some of the factors which affect the way of keystroke Dynamics as follows: Text length, sequences of character types, word choice, and number of training sample, statistical method to create template, mental state of the user, tiredness or level of comfort, keyboard type, keyboard position and height of the keyboard, hand injury, weakness of hand mussel, shoulder pain, education level, computer knowledge, and category of users.

F. Algorithms

Many classification methods have been applied in keystroke dynamics study over the last three decades. Following are the distance based algorithms were used to evaluate the system. But we can use any type of features mining algorithm.

1) Manhattan Distance

Manhattan distance is one distance based method which calculates the score and minimum score will be treated as perfect match and corresponding user will be treated as a genuine user.

Manhattan distance is formulated bellow:

$$M=\sum_{i=1}^{n}(|x_i - y_i|) \quad (9)$$

Where $x=(x_1, x_2, x_3, ...,x_n)$ represents stored vector and $y=(y_1, y_2, y_3, ..., y_n)$ represents the claim vector of the exercise sample.

2) Manhattan with Standard Deviation Distance

The standard deviation of each feature is calculated as in Equation 10. Here α_i represents standard deviation.

$$Ms=\sum_{i=1}^{n}(|x_i - y_i|)/\alpha_i \quad (10)$$

3) Euclidean Distance

The score is calculated as the squared Euclidean distance between the stored vector and claim vector as in Equation 11.

$$E = \sqrt[2]{\sum_{i}^{n}(|x_i - y_i|)^2} \quad (11)$$

4) Mahanabolis Distance

The standard deviation of each feature is calculated, where Mahanabolis distance is presented in Equation 12.

$$Eh = \sqrt[2]{\sum_{i}^{n}((|x_i - y_i|)/\alpha_i)^2} \quad (12)$$

5) Z Score Values

The z score is calculated in Equation 13:

$$Z=\sum_{i=1}^{n}(|x_i| - (|x_i|))/\alpha_i \quad (13)$$

Where μ (xi) are mean value and α_i is standard deviation.

G. Types

Generally the version of keystroke dynamics are free text keystroke dynamics (like paragraph of written text) and fixed text keystroke dynamics (like signature of written text).

H. Advantages

It is a low implementation cost, very simple, strong as biometric characteristics, continuous monitoring method.

I. Disadvantage

This technique can be effectively applied in application areas such as student or employee attendance system, distance based examination, password recovery mechanism, emotion recognition, private data encryption, continuous user verification, criminal investigation, identifying backdoor accounts, free-text user authentication etc.

J. Application Areas

This technique can be effectively applied in application areas such as student or employee attendance system, distance based examination, password recovery mechanism, emotion recognition, private data encryption, continuous user verification, criminal investigation, identifying backdoor accounts, free-text user authentication etc.

III. EXPERIMENTAL SETUP

We have implemented a program in JAVA for experimental purpose, which has the capability of capturing all key pressing and releasing events, which are used to create the database of different sample of passwords and timing templates. It can also calculate different score or distance between vectors. Fifteen users are invited to press three most common passwords six times each using same keyboard. First password which is the combination of only digits (i.e., "123456"), second password is the combination of alphabets (i.e., "password") and third one is combination of both (i.e., "kolkata123").

IV. EVALUATION AND ANALYSIS

Entered characters have been collected from common daily used words and the system calculates the key press and release time which is like table I. After the system calculates 8 timing factors in the table II.

Table 1: KEY PRESS & RELEASE TIME OF FIXED-TEXT "kolkata123"

Entered Key	Key press time	Key release time
K	0	109
O	172	281
L	375	484
K	609	733
A	749	889
T	1326	1451
A	1482	1623
1	1950	2059
2	2169	2278
3	2387	2496

Table 2: SAMPLE KEY HOLD AND KEY LATENCY TIMES WITH STANDARD DEVIATION FOR THE RECORDED FIXED-TEXT "kolkata123"

Key	Key hold time (T1)	Down Down key Latency (T2)	Up Up Key latency (T3)	Up Down Key latency (T4)	Down Up key Latency (T5)	Di-graph (T7)	tri-graph (T8)	Syllable graph	Total Time (T6)
K	78	187	213	292	109	182	293	293	3935
O	104	205	213	317	101	215	350	-	-
L	111	267	291	403	156	246	357	-	-
K	135	140	116	252	5	246	357	246	-
A	111	231	231	343	119	222	326	-	-
T	111	127	119	231	15	215	293	215	-
A	104	377	351	455	273	182	294	-	-
1	78	174	208	286	96	190	294	294	-
2	112	205	197	309	93	216	-	-	-

3	104	-	-	-	-	-	-	-	-

There is some performance measurement parameters that are used to evaluate performance of different biometric system define by Giot R.

A. False Acceptance Rate (FAR):

FAR is defined as the percentage ratio between falsely accepted illegal users against the total number of imposters accessing the system.
FAR defined by the following equation:

$$FAR = \frac{\text{Number of falsely accepted illegitimate users}}{\text{Total number of imposters}} \% \quad (14)$$

B. False Rejection Rate (FRR):

FRR refers to the percentage ratio between falsely denied genuine users against the total number of genuine users accessing the system.
FRR defined by the following equation:

$$FRR = \frac{\text{Number offalsely denied legitimate users}}{\text{Total number of genuine users}} \% \quad (15)$$

C. Equal Error Rate (EER):

EER is the rate at which both false acceptance and false rejection error are equal.
See the following figure.

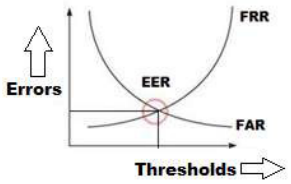

Figure 1: Equal Error Rate or Cross Error Rate

In our simulation program in JAVA, we have recoded each key press and release time for six sample of passwords (size of password <=10) and calculated key hold time and latency time between various down and up key sequence latencies, which are shown in the Table II.

Euclidean distances of the string “123456” of 15 users are very similar. Where “kolkata123” and “password” strings are strong to identify the user which is shown bellow in the figure 2.

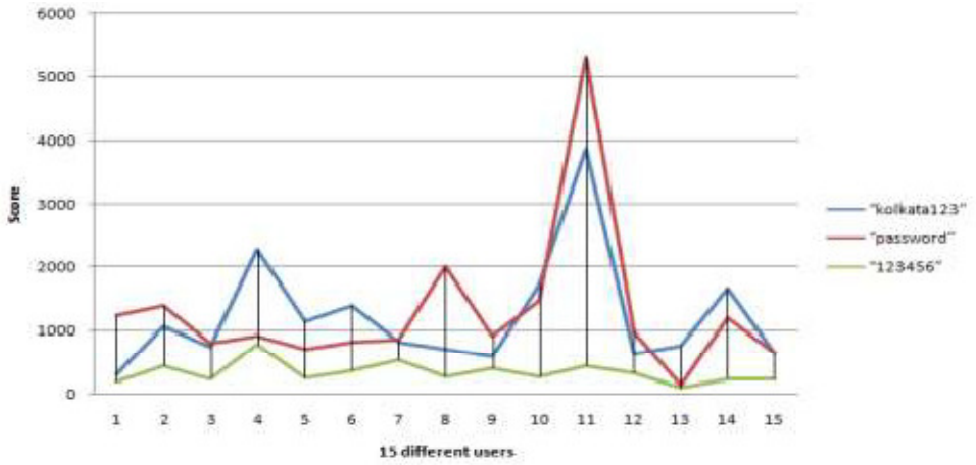

Figure 2: Euclidean distances of different set of data for the password "kolkata123", "password" and "123456"

As per the experimental result, we got 0.133% EER for the string "kolkata123" where 0.4% and 0.53% EER for the string "password" and "123456" separately. Size of the string and common used words are best choice to choose secrete question answer or hints where we are habituate to press this type of words daily. We have collected data from the users those are belonging to Kolkata, India, so they are habituated to press string *kolkata*. We got the better result for the string "kolkata123" because it is a longer size string than other tested strings.

In our experiment, we have collected rhythm of 270 predefined fixed-texts from 15 users from Kolkata and seen that the user's typing styles are dissimilar as per following line chart generated by the experimental database. Here we have considered key duration (kd) and 4 sequences of down and up key latencies (ddkl, uukl, dukl, udkl) as well as in different figure we considered total time and digraph and tri-graph times. Good suggestion is choose the password what we press daily such as user ID, name, place etc. Otherwise three factors will affect the system, finger movement time, key searching time and different keyboard.

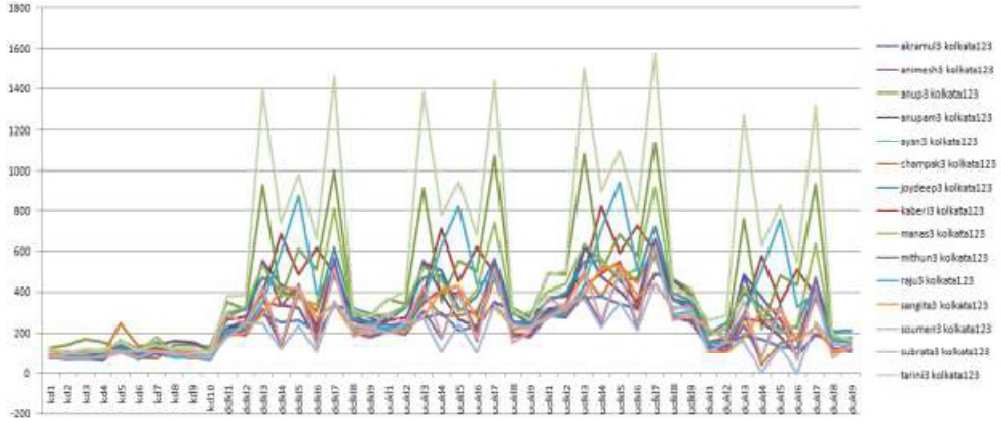

Figure 3: 15 same sample of passwords "kolkata123" for 15 different users considering kd, ddkl, uukl, dukl and udkl.

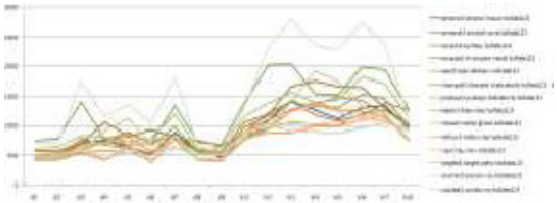

Figure 4: 15 same sample of passwords "kolkata123" for 15 different

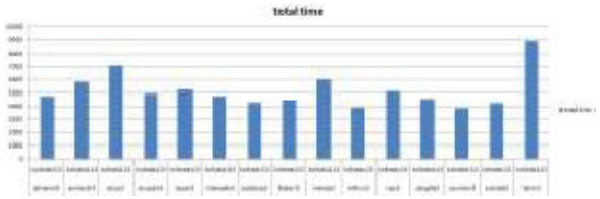

Figure 5: 15 same sample of passwords "kolkata123" for 15

In our experiment we considered all timing features. Some new features also can be calculated like finger tips size, key press sound, finger placement, key pressure which will make the system error less.

V. DISCUSSION AND FURTHER AREA OF RESEARCH

Keyboard is essential for a computer device, which can be used to recognize our typing style and very much unique as per our experimental result and cannot be copied or stolen but strong and low cost.
Here no extra device is need to recognize our gait. Keyboard is essential device which is enough to recognize our gait. It can be used as safe guard of our password in any access control system.

Different type of keyboard such as keypad, desktop keyboard or standard keyboard, and screen touch keypad may affect the way of keystroke dynamics. Basically keypad is not changing frequently in mobile phone. So this technique can be effective for mobile security as per Trojahn, M. and Ortmeier, F. otherwise artificial keystroke dynamics or keystroke sound implemented by Roth, J. et al. and Metaxas D. would be introduced.

Characteristics of human may change over time. So update mechanism is needed to update template after acceptable verification or identification. This technique can be effectively applied in application areas such as student or employee attendance system, distance based examination, password recovery mechanism, emotion recognition by Kolakowska, A., private data encryption, continuous user verification, gender identification, criminal investigation, identifying backdoor accounts, free-text user authentication etc.

Sometimes, score of different algorithms varies. It would be better if we combined all scores in a single equation like mean value calculation with given weights of all scores.

It would be better if we choose some daily used words as password for consistence typing style. Sequence of character is effective. Combination of left sided and right sided key is better. Length of the string must be high.

VI. CONCLUSION

Keystroke Dynamics is a behavioral biometric characteristic which can be used to segregate and distinguish people, this is the method where people can be well-known by their typing style much like hand-writing which is used to recognize the author of the written text. Here, the method classifies the people based on their typing style which is meshed up with password or PIN and peoples' typing styles are similar, because similar neuro-physiological factors like our written signature is unique all the time. This technique can be easily implemented in any computer system with small alternation. But it has some drawbacks; Person's typing speed may vary subsequently during a day or between two days depending on the mental state of a person, hand injury, emotional status of the person. Typing speed may also differ because of the position of the keyboard and nature of the keypad. To store all the information of different factors of typing rhythm for a person, huge memory is needed. So the accuracy of this method is not as much promising, performance level is quite less, needed memory space is not acceptable. But we can solve this problem by introducing all the effective factors or introducing artificial keystroke dynamics. Here we have to use the keyboard as piano or keystroke sound, pressure on key, placement of finger can be considered. The results from this study and others indicate that behavioral based biometrics generally and keystroke dynamics specifically provide a level of security and it can be implemented in any system.

This technique can be used in online criminal investigation, back door account identification, online typing examination, emotion recognition, lab attendant system many more. Our experimental result is near perfect but in real life it is yet to be improved which is our future work so that the system can cope with every situation.

REFERENCES

[1] Hafiz, Z. U. K. (2010). Comparative Study of Authentication Techniques. International Journal of Video & Image Processing and Network Security, IJVIPNSIJENS Vol: 10, No: 04.

[2] Gaines, R. et al. (1980). Authentication by keystroke timing: some preliminary results. Rand Rep. R-2560-NSF, Rand Corporation.

[3] Bleha, S. et al. (1990). Computer-access security systems using keystroke dynamics. IEEE Transactions on Pattern Analysis and Machine Intelligence 12, 1217–1222.

[4] Killourhy, K. S. (2012). A Scientific Understanding of Keystroke Dynamics. PhD thesis, Computer Science Department, Carnegie Mellon University, Pittsburgh, US.

[5] Joyce, R. & Gupta, G. (1990). Identity authorization based on keystroke latencies. Communication of ACM 33 (2) 168–176.

[6] Monrose, F. & Rubin, A. D. (2000). Keystroke dynamics as a biometric for authentication. Future Generation Computer Systems, Vol. 16, No. 4, pp. 351–359.

[7] Shimaa, I. H. et al. (2013). User Authentication with Adaptive Keystroke Dynamics. IJCSI, Vol. 10, Issue 4, July 2013.

[8] Pin, S. T. (2013). A Survey of Keystroke Dynamics Biometrics. The Scientific World Journal, Vol-2013, Article ID 408280.

[9] Giot, R. et al. (2011). Analysis of template updates strategies for keystroke dynamics. Computational Intelligence in Biometrics and Identity Management (CIBIM), 2011 IEEE Workshop, pp.21,28, 11-15.