Biometric Identification System Based in Keyboard Filtering*

Òscar Coltell[1], José M. Badía [2], Guillermo Torres [3]
Computer Science Department. Jaume I University
[1] "coltell@mail.uji.es", [2] "badia@mail.uji.es", [3] "al009757@anubis.uji.es"

Abstract

We have revised several authentication systems based on biometric technology to resume advantages and disadvantages. Because pure hardware biometric systems of user authentication have low rate on results over computational and economic cost, alternate biometric methods of low computational cost, based in software development, are also being evaluated. We have developed a first prototype of a software system to elicitate sets of 20 password stroke samples, named attacks, with a population of 10 different users totalling 200 attacks. The results obtained demonstrate that users follow generally certain patterns when they are writing their password, and is possible to reinforce the user's password authentication method by means of the analysis of user stroking patterns. In addition to, it is necessary to increase the population size and number of samples to establish standard and reliable rules. Finally, it is very difficult to find a general user pattern applied to every password.

1. Introduction

The main objective of our study is to assess the validity of keystroke analysis to identify an user who tries to access to a computer. To achieve this objective we made a fist approach to assess the validity of the method using short words as password patterns. It could verify the method utility as assistant subsystem in the user authentication process to access to computer systems. In a second approach, we have verified this method using long phrases or paragraphs. It could test the possibility to differentiate users during a computer session performing concrete tasks. These tasks currently evolve enough size texts.

In this way we have checked the authentication method validity based on the influence of the following: the selected password, its difficulty degree, the pattern associated to it and the way the user keystrokes it. In addition to, we have studied the influence of the learning or training process, the stability degree of the pattern ones it has been learned, and the influence of the tiredness or the pattern repetition.

Our purpose has been to design a simple method and to built a low cost software system that easily runs on every personal computer and operative system. By the other hand, this one is non-intrusive and troublesome for the user. It does not require special actions to be performed by the user. We have revised several authentication systems to resume advantages and disadvantages of them ones. A subset of those systems is applying the biometric technology in user authentication methods: face recognition, fingerprint recognition, iris recognition, hand and finger geometry methods, handwriting recognition, and so on. Because pure hardware biometric systems of user authentication have low rate on results over computational and economic cost, alternate biometric methods of low computational cost, based in software development, was also being evaluated.

In this point, we put a question: why there are few references to this recognition method including very specialized biometric sites as www.biometrics.org? We have found some old references [8] [11] [12] and [17]. This research was withdrawn long time ago. Why? The reasons could be the operative characteristics of the computer system, the great pattern variation, and the influence of environment and the user emotional state. Along this paper, we will try to response to these questions.

We begin discussing the security problem and biometric solutions based in keystroke analysis. We then describe the methodology applied in our study. We subsequently describe our technical solution to access control to computers based in the analysis of user's keystroke patterns. After that, we present results of our preliminary experiment with a reduced sample of the population. Finally, we give some conclusions and possible extensions to the method.

2. Security and biometric solutions

It is well known that one of the first aspects to consider in computer system security is the access control [18]. This try to guarantee that only authorized persons or computers can access to the system. To achieve this objective it is necessary to state any system that performs user's authentication when this one would to access it. The most extended mechanism that covers the authentication process it is the simple password. However, this isolated system is highly vulnerable. Therefore, it is very important to use any alternative mechanism o to reinforce the former one.

* This work has been founded by CICYT (grants IFD-0011 and TAP98-0465) and by the FUNDACIÓ BANCAIXA – CAIXA DE CASTELLÓ grant P1A98-14.

Beyond all doubt, the most secure mechanism for user's authentication is the biometric features exploitation, because these ones are intimately linked to every individual. Different environments that support biometric features are widely extended [3], [15], and [13]. It would be to note the ones based on fingerprints and iris recognition, and that is the most commercially implemented.

The main disadvantage of biometric methods is that they usually have associated the support of any kind of hardware device for its implementation. This fact increases highly their installation and it makes more difficult to use them. It is necessary to find a software system the simple as possible that guarantees user authentication of the computer system without to make the process more expensive. One of the mechanism that holds it is the keystroke control over a keyboard (keyboard filtering, keystroke identification). There are several previous papers that analyze the useless of this system as access control mechanism [1], [2], [5], [6], [8], [9], [11], [12], [14], [17], [16], and [19].

The mechanism feasibility is based in the hypothesis that every individual follows different guide lines when he is stroking a keyboard. In other words, when an individual is writing a character sequence he maintains certain rhythm that is different of the one of other individuals. Therefore, firstly it would be possible to identify a computer user based in this feature. The problem is to check if this hypothesis holds in the practice, under what circumstances, and if it is possible to use it as access control mechanism.

When an individual keystroke is characterized, its rhythm is usually measured. In other words, being a character sequence it has to measure the individual keystroke latencies that are equal to the time elapsed between two consecutive keystrokes of different characters. Knowing the sequence, the tuple formed by the latencies, may be used as pattern, which identified a computer user.

This biometric characteristic, not only can be used as authentication mechanism, also as verification mechanism of the identity (dynamic signature), and indeed as monitorization mechanism of user performances. Five generations or levels of the control of the access are differentiated (15): identification, authentication, verified authentication, reciprocal authentication, and re-authentication. The use of the easy password would be integrated in the generation of authentication, the control of the access through keystroke analysis would be framed in the verified authentication and the use of the same technique for dynamic verification would be found in the re-authentication generation.

The possibility to use the keystroke analysis for the dynamic verification or for the output control is due to the fact that a lengthily use of the system guarantees the possibility of differentiating much better the users. Regarding to the access control, the objective of the dynamic verification is to guarantee not only that the user who access is who says he is, but the authorized user does not change along all session. Some legal aspects over user electronic monitorization are developed [4]. And some studies over the influence of monitorization in the physiological activity of individuals are published [10], [17]. The last ones demonstrate that individuals modify certain physiological parameters themselves when they are advised that they are being monitorized. It may be important to use this mechanism as re-authentication method or access control process.

When we use keystroke analysis applied to access control we must to define some aspects. Firstly, it must define the character sequence that will be used for to state the identification pattern of each individual. In [8] and [11] four strings are used for this proposal: the login, the password, the personal identification number (PIN), and user first and last names. With them about 1% of false authentications and 7% of false rejects are obtained. In [2] it takes the user name obtaining a "reasonable" low ratio of false alarms. In [19] is used also the mean of individual latencies when he has previously stroked every successive key of the keyboard.

With respect to the verification system applied over the access control method based in keystroke analysis, the most extended technique is to define a variation margin over different latencies and to accept a pattern which match within this margin. Neuronal nets are applied to the pattern matching in the recognition process [2]. In [5], [6], [14] and [15] fuzzy logic is applied also with this aim. The recognition ratio is usually 70% with these techniques using the couple (user name, password).

3. Methodology

When we are starting the study of the feasibility of keystroke analysis as access control mechanism it is necessary to consider some questions: what to measure? How to measure it? Is influencing the difficulty degree? What is the recognition algorithm? And how to design and to perform the experiment? In the following, we would try to response to all ones.

3.1. Defining the Parameters to Measure

Firstly, it must to define the character sequence applied to identify the user. The sequence length must be balanced to avoid tiredness, when user writes a lot of characters, or insecurity, when the number of character is less than usually accepted computer passwords. Therefore, we have decided to work with classic models of computer passwords.

3.2. Defining the Measure System

Classic models of computer passwords allow us in our system to obtain the user name and password, and the user rhythm when he strokes his password. In order to authenticate the user we need to contrast the written password over the user identification pattern. With this aim, the user must to perform a creation process for every distinct password he uses or changes. The process will request to the user that he writes his password certain

number of times (nv) and it will saves the tuple composed by latencies measured between every successive keystrokes.

Our assumption is that the user keystrokes in a similar way his current password every time he tries to access to the computer. May be in first times he is writing a new password the pattern is significantly different or the variation degree goes beyond the variation margin. The solution is to discard the first ten to twenty password strokes. Thus, this is our training period and the pattern stabilization activity. In addition, we propose to discard minor and major latencies of each key couple in order to achieve the homogenization of our experiment results. For example, if 30 password keystrokes are made to define the pattern, only will be take as valid the last twenty ones. And the first 10 ones corresponds to the training period. In addition, we will discard minor and major latencies from the other 20 key couples, obtaining finally 16 latencies stored for each key couple.

The pattern associated to each user is defined by two arrays which contains the above 16 latencies. The first array contains mean values of latencies and it defines the mean keystroke rhythm associated to the user. The second one contains standard deviation values of each consecutive key, and it defines the homogeneity degree with which the user strokes his password. In other words, defines how the user strokes his password in the same way all times.

3.3. The Influence of the Difficulty Degree

We start with the hypothesis that the more difficult passwords to stroke generate more differentiate patterns that the more simple passwords. The former ones are more valid for authentication by means of keystroke analysis. The problem is to state the difficulty degree of a doing password. In [6] is applied the distance from key to key in a keyboard to define the above-mentioned factor. In our proposal, we take three different passwords as test set in order to find a more related factor.

The easiest level password is composed by characters that are near from the correct location of fingers which must maintain a good typist ("asdf – jklñ"). We have selected the string "nicanor88". The intermediate level password must to follow the model usually requested by modern UNIX systems. It is a password with a length great than 6 characters and which has at least 2 numeric or special characters. We have selected the string "16bonet16". The high level password is a non-significant character combination of numbers and letters. Concretely we are used a password composed by the acronym of any long phrase: "lqevsly39" (It corresponds to a famous film name which had its première in 1939, but partially translated to Spanish language: "*lo que el viento se llevó, year 1939*" – 'gone with the wind, year 1939').

3.4. Defining the Recognition Algorithm

Firstly, it is necessary to assure the password correctness previously to recognize an user. Then, only are considered valid for keyboard analysis keyed passwords that match completely with stored passwords. On the contrary, the former ones will be ignored and it will be replayed the password request, like UNIX systems are usually doing. If the keyed password is correct, then the time elapsed between every two consecutive keys stroked. The value obtained and stored will be compared against the stored user pattern.

We can suppose that the stored pattern for the user A is defined by two arrays: m = {m1, m2, m3, m4, m5, m6, m7, m8} and d = {d1, d2, d3, d4, d5, d6, d7, d8}, where the m array contains mean values of latencies and the d array contains standard deviation values of latencies. On the other hand, we can suppose that the measured pattern when the user strokes his password is the array s = {s1, s2, s3, s4, s5, s6, s7, s8}. Access will be allowed if the following rules holds:

$$|s_i - m_i| <= f * d_i$$

where n+1 is the keyed password length and f is an access constrain factor that defines the admissible error margin of the pattern to allow computer access. Preliminary experimental results show that this is a too restrictive criterion for all latencies measured. Therefore, we will allow a criterion looseness of one latency as maximum. In the Results section we will show the validity of this weaker criterion.

3.5. Describing the Experiment

The population sample is composed by 10 lecturers and students of computer science. Two experiments are been realized for each sample subject as we describe in the following.

Experiment number 1: Password introduction.

In the first experiment the access process to a UNIX computer is simulated. Firstly, the user's name (login) is requested, and subsequently the password is requested. If the password is not correct, the request sequence is repeated until three times. Then, the process is re-started. The user is enforced to use the mouse for starting the introduction of each new couple (user's name, password), in order to avoid the influence of bad habits when the user introduces speedy and repeatedly the same password.

Each user is allowed to perform series of training tests with the previously established schema. When the user thinks that he is ready, it starts the test phase when 20 samples of correct passwords are stored. All this process is applied using three distinct passwords corresponding to different difficulty levels.

Experiment number 2: Text test.

In the second experiment we pretend to check if introducing text of certain length it allows identifying all users of the sample. A text of 80 character length has been selected and users are requested to stroke it. This text contains letters, numbers and punctuation characters, and words of different difficulty levels are included. In this case, we have only considered results where the text has

been introduced correctly. In addition to, for each individual an auxiliary set of physiological and typing expertise data are solicited with the aim to study the possible influence of personal profiles over the keystroke pattern. This set contains, for example, information about computer expertise years, if the user types with all their fingers or with some fingers, if the user have took a typist course, and so on.

4. Technical Description of the System

Experimental tests are performed over a PC Intel Pentium platform running MS Windows 9x. The choice of this environment is justified by the low cost hardware and software features offered. The main reason of the implementation of an effective computer tool for access control based in keystroke analysis is focused on the possibility for adequately access to keystrokes and system clock ticks. This holds in our solution.

Windows is an event-oriented operative system where keystrokes, and other hardware events, are stored in a queue. This one can be reached by means of suitable tools. On the other hand, Windows provides functions, as "GetTickCount", that allow us to obtain clock ticks with one-second accuracy. This is necessary in our experiments. To implement the access control program Visual Basic 5.0 are used. Native code is used in the compiling stage to obtain a program with enough speed for measuring keystroke intervals.

The system portability to other environments is hardly dependent of their operative system capability for access in real-time to clock ticks and keystrokes. In this way, in a multitask and multi-user computer system running UNIX, where net and terminal connections are maintained, it should not possible to apply our technique directly. But if we use a PC computer as terminal and its own hardware, and if we can reach to their clock and keyboard functions without interference, the method can be also performed in this kind of environments.

5. Results

The Figure 1 shows patterns associated to different users when they are introducing the first of selected passwords. It can see that users are following distinct stroking rhythms, in spite of latencies are following a general pattern easily recognizable. This demonstrates that, in general, the same password is stroked by the majority of users following a pattern. Thus, in this Figure it can see clearly that the next to last latency is larger in near all cases. This is justified by the transition from a letter key to a number key in the upper side of the keyboard.

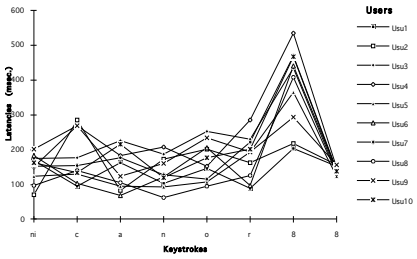

Figure 1. Patterns of 10 users in the sample introducing the first password ("nicanor88").

The Figure 2 shows standard deviations associated to different users when they are introducing the first of selected passwords. It can see that standard deviations are restricted to a clearly defined zone, with the exception of user no. 2 with the second latency. In the same way, it can see as different users have different deviation patterns.

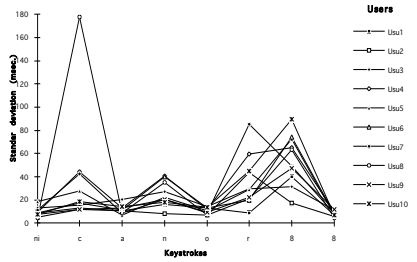

Figure 2. Standard Deviation of different patterns introducing the first password ("nicanor88").

The Figure 3 shows patterns associated to 20 password sequences carried out by the user 1 when he is introducing the first of selected passwords. It can see that, if we discard the cases which are more separated of mean values in each latency, we can obtain a very clear pattern associated to the user. In addition to, it is possible to see how keystroke patterns of the same user in different attempts are more similar between them that patterns associated to other users. This fact allows us to differentiate user's keystroke rhythms.

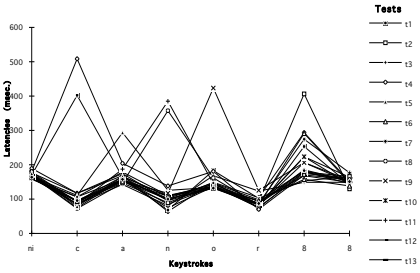

Figure 3. Keystroke pattern of user 1 introducing the first password ("nicanor88").

The Figure 4 shows mean patterns associated to different users when they are introducing the third of selected passwords. It can see that these patterns are

absolutely distinct from the ones in the first password. On one hand, measured latencies are higher that in the Figure 1. This proves that this password is very difficult to stroke that the first one. By the other hand, standard deviation are also higher that in the Figure 1. This reasserts this idea and this will have consequences related to the method validity with this kind of passwords.

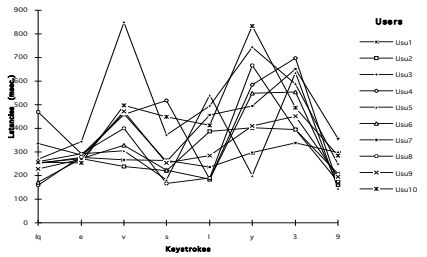

Figure 4. Patterns of 10 users in the sample introducing the third password ("lqevsly39").

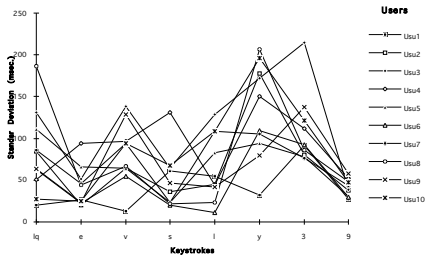

Figure 5. Standard Deviation of different patterns introducing the third password ("lqevsly39").

The Figure 5 shows standard deviations associated to different users when they are introducing the third of selected passwords. It can see that standard deviations are not restricted to a clearly bounded zone, because the difficulty level of password stroked.

If we study the criterion applied to accept a keystroke pattern as valid pattern, we can see that the filtering degree of the keystroke analysis system depends on standard deviations associated to different patterns. Therefore, if standard deviation is very high, the possibility to admit as valid attempts pertaining to other users increases notably (false impersonation).

		Password1		Password2		Password3	
User	f	Access Allow.	False Imp.	Access Allow.	False Imp.	Access Allow.	False Imp.
1	2	70	0,00	60	0,00	70	0,56
	2,5	85	0,00	75	0,00	80	1,11
	3	85	0,00	85	0,00	80	2,22
2	2	70	0,00	55	0,00	65	7,22
	2,5	90	0,00	80	0,00	80	17,22
	3	95	0,00	80	0,00	85	28,89
3	2	60	0,56	75	8,89	60	8,89
	2,5	70	0,56	80	11,11	70	24,44
	3	85	5,00	90	16,67	80	60,00
4	2	65	0,56	60	0,56	65	8,89
	2,5	80	2,78	80	5,00	80	25,00
	3	90	11,67	80	18,33	85	47,78
5	2	75	2,78	60	0,56	60	2,22
	2,5	85	5,56	80	1,67	80	6,11
	3	90	9,44	90	3,89	90	15,56
6	2	70	0,00	70	0,00	65	2,22
	2,5	80	0,00	50	2,78	80	7,78
	3	85	0,00	50	8,33	80	13,89
7	2	45	2,22	65	0,00	70	4,44
	2,5	70	4,44	75	2,22	95	15,00
	3	85	15,56	85	20,56	95	36,67
8	2	50	0,56	60	0,00	65	10,56
	2,5	70	2,78	80	0,00	75	19,44
	3	85	3,33	95	0,56	85	30,00
9	2	70	0,00	85	3,89	75	20,56
	2,5	85	0,00	100	10,56	90	33,33
	3	95	1,11	100	12,78	90	49,44
10	2	55	0,00	70	0,56	80	3,33
	2,5	95	1,67	75	0,56	85	16,67
	3	85	8,89	85	0,56	85	39,44

Table 1. Analysis of allowed access using different passwords depending on factor f

The Table 1 shows percentage results obtained of the access analysis with all selected passwords depending on the access constrain factor f. For every password, the percentage of access attempts allowed over the total of 20 for each of 10 users is shown. In addition to, the percentage of access attempts allowed to non authorized users over the total of 180 performed is also shown.

The Table shows that percentages of allowed access to authorized users are very high. Generally these ones surpass 70%. This demonstrates that different users follow a definite pattern. And if we use this pattern is possible to identify them in the majority of cases when they are stroking a determined password.

On the other hand, it can see that in the first and second passwords the number of allowed access to non-authorized users is very low. In the majority of users percentages are

raising 0%. In addition, the Table shows the effect of to modify the factor f value. If we decreases f, the constrain degree increases. Thus, attempts of non authorized users are rejected, but the possibility to reject attempts of authorized users also increases.

Finally, in the third password case, the number of success attempts of non-authorized users substantially grows. This is because special features of this password.

6. Conclusions

The analysis of keystrokes suitably applied can be a valid tool for reinforces the classic access control based on simple password methods. This fact is justified in that different users stroke the same password following a differentiating pattern. The validity of this method is very larger in the case of easily stroking passwords. In this case we consider that a password of this type is composed by an intelligible word and, optionally, by some number located before or after the word. However, when passwords are composed by letters and numbers without specific mean, the method is less effective. Possibly, in this case it would be necessary a longer training period to define adequately patterns associated to different users.

It is important to note that this kind of methods based on keystroke analysis offer several advantages over the other methods based on physiological biometric characteristics. The main advantage is the cost-effectiveness ratio. Because our method is supported by a software system, it does not need sophisticated and expensive devices to capture biometric features (iris, fingerprints, etc.). On the other hand, it is a simple, friendly and non-intrusive system. Finally, the use of this environment does not need to perform any additional process, because it could be integrated into the classic access schema of the computer based on the couple (user's name, password).

Acknowledgements

We would give thanks for their collaboration as sample members, to lecturers of the Computer Science Department at Jaume I University, and to computer science students at the same university.

7. References

[1] Bleha, S.; Slivinsky, C. and Hussein, B., (1990), "Computer-Access Security Systems Using Keystroke Dynamics.", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-12, no. 12. pp. 1217--1222.

[2] Brown, M. and Rogers, S.M., (1993), "User Identification via Keystroke Characteristics of Typed Names using Neural Networks.", International Journal of Man-Machine Studies, vol. 39, no. 6. pp. 999--1014.

[3] Consortium, B., "www.biometrics.org".

[4] CSL, (1993), "Guidance on the legality of keystroke monitoring.", CSL Bulletin,

[5] de Ru, W.G. and Eloff, J.H.P., "Information security - The next decade", IFIP TC11. Eleventh International Conference on Information Security 1995),

[6] de Ru, W.G. and Eloff, J.H.P., (1997), "Enhanced Password Authentication through Fuzzy Logic.", IEEE Expert/Intelligent Systems & Their Applications, vol. 12, no. 6.

[7] Deane, F.; Henderson, R., et al., (1995), "Theoretical Examination of the Effects of Anxiety and Electronic Performance Monitoring on Behavioural Biometric Security Systems.", Interacting with Computers, vol. 7, no. 4. pp. 395--411.

[8] Gupta, G.K. and Joyce, R., "User Authorisation Based on Keystroke Latencies", JCU-CS-89/5, Dpt. of Computer Science, James Cook University, (1989).

[9] Haunold, P. and Kuhn, W., (1993), "A Keystroke Level Analysis of Manual Map Digitizing.", Lecture Notes in Computer Science, vol. 716, pp. 406--.

[10] Henderson, R.; Mahar, D., et al., (1998), "Electronic Monitoring Systems: An Examination of Physiological Activity and Task Performance within a Simulated Keystroke Security and Electronic Performance Monitoring System.", International Journal of Human-Computer Studies, vol. 48, no. 2. pp. 143-157.

[11] Joyce, R. and Gupta, G., (1990), "Identity authentication based on keystroke latencies.", Communications of the ACM, vol. 33, no. 2. pp. 168--176.

[12] Joyce, R. and Gupta, G., (1990), "User Authorization Based on Keystroke Latencies.", Communications of the ACM, CACM, vol. 33, no. 2.

[13] Kim, H., (1995), "Biometrics, Is it a Viable Proposition for Identity Authentication on Access Control).", Computers and Security, vol. 14, no. 3. pp. 205--214.

[14] Labuschagne, L. and Eloff, J.H.P., (1997), "Improving system-access control using complementary technologies.", Computers & Security, vol. 15, no. 6. pp. 543--550.

[15] Labuschagne, L.; Eloff, J.H.P. and de Ru, W.G., "Practical Considerations for Access Control", Dpt. Computer Science. Rand Afrikaans Univ., (1996).

[16] Leggett, J.; William, G., et al., (1991), "Dynamic Identity Verification via Keystroke Characteristics.", International Journal of Man-Machine Studies, vol. 35, no. 6. pp. 859-870.

[17] Leggett, J. and Williams, G., (1988), "Verifying Identity via Keystroke Characteristics.", International Journal of Man-Machine Studies, vol. 28, no. 1. pp. 67--76.

[18] Pfleeger, C.P., "Security computing.", (IEEE Computer Society, 1996)

[19] Umphress, D. and Williams, G., (1985), "Identity Verification Through Keyboard Characteristics.", International Journal of Man-Machine Studies, vol. 23, no. 3. pp. 263--273.