Fuzzy and Markov Models for Keystroke Biometrics Authentication

DAT TRAN, WANLI MA, GIRIJA CHETTY and DHARMENDRA SHARMA
School of Information Sciences and Engineering
University of Canberra
ACT 2601
AUSTRALIA

Abstract: - Keystroke biometrics authentication system is based on a password and keystroke biometric features captured when a user is typing in the password. The system offers a higher level of security and convenience for computers. The system does not require additional hardware as it can be used with any existing keyboard, making it relatively inexpensive and fairly unobtrusive to the user. There have been existing research publications on keystroke biometrics authentication that have solved problems in selecting appropriate keystroke features and modeling users. However methods for calculating score to reduce authentication error are not taken into account. Therefore we propose to use Markov modeling and fuzzy set theory-based normalization methods for keystroke biometrics authentication that can reduce both false rejection and false acceptance rates. Experiments showed better performance for the proposed methods.

Key-Words: - Markov modeling, fuzzy normalization, user authentication, keystroke biometrics

1 Introduction

The process of verifying the identity of a user is known as user authentication. Authenticators can be passwords, biometric identification such as voiceprint and signature, and physical identification such as passport and credit card [1]. Passwords are excellent authenticators, but they can be stolen if recorded or guessed. Biometrics are useful to establish authenticity and for non-repudiation of a transaction, wherein a user cannot reject or disclaim having participated in a transaction [2]. However, biometrics can be counterfeited, so they cannot ensure authenticity or offer a guaranteed defense against repudiation. It is a good approach that different types of authenticators should be combined to enhance security and performance [3, 1]. A very good combination is a user authentication system which is based on a password and keystroke biometric features captured when the user is typing in the password. This user authentication system will operate on standard computers providing a high level of security. The system does not require additional hardware as it can be used with any existing keyboard, making it relatively inexpensive and fairly unobtrusive to the user [4]. There have been existing research publications on keystroke biometric-based user authentication, for example, Gaines et al. [5] in 1980, Leggett et al. [6] in 1991, Obaidat and Sadoun [7] in 1997, Yu and Cho [8] in 2003, Mandujano and Soto [9] in 2004, Hocquet et al. [10] in 2005, Villani et al. [11] and Chang [12] in 2006. These research works have solved problems in selecting appropriate keystroke features and modeling users based on those selected keystroke features.

However, the main requirements for a user authentication system are that the system should be fast for real-time processing, efficient, requires minimum storage, and robust against textual errors. Therefore we propose Markov modeling and fuzzy normalization methods to obtain these requirements. A password typed in using a computer keyboard is considered as a sequence of key characters consisting of letters, digits, common characters such as comma and semicolon, and invisible characters such as Shift key, and Ctrl key. The occurrences of key characters in a password can be regarded as a stochastic process and hence the password can be represented as a Markov chain where key characters are states. The occurrence of the first key character in the password is characterized by the initial probability of the Markov chain and the occurrence of the other key character given the occurrence of its previous key character is characterized by the transition probability. The initial and transition probabilities for the Markov chain representing the password are calculated and the set of those probabilities is regarded as a Markov model for that password.

We also propose a new fuzzy approach to normalization methods for keystroke biometrics authentication. For an input keystroke feature sequence and a claimed identity, a claimed user's score is calculated and compared with a given threshold to accept or reject the claimed user. Considering the user authentication problem based on fuzzy set theory, the claimed user's score is viewed as a fuzzy membership function. Fuzzy entropy, fuzzy c-means and noise clustering membership functions are proposed as fuzzy membership scores, which can overcome some of the problems of ratio-type scores and reduce the false acceptance rate. Experiments were performed to evaluate proposed normalization methods for keystroke biometrics authentication and showed better results for the proposed methods.

2 Observable Markov Model

2.1 Keystrokes Biometric Features

For each key on a computer keyboard which is pressed then released, key features extracted are as follows: key character, key code, time at which the key is pressed, time duration when the key is pressed until it is released, and time duration between the previous key is pressed and the current one is pressed.

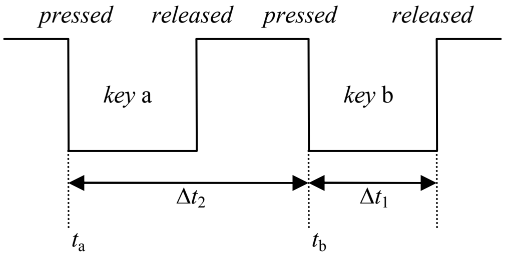

Fig. 1. Keystroke biometric feature extraction.

For example, Table 1 shows keystroke features for the word "University" typed by a user.

Table 1. Keystroke biometric features extracted for the word "University". Times are in seconds.

Key character	*Key Code*	*t*	Δt_1	Δt_2
(Shift key)	16	0.00000	0.51563	0.00000
U	85	0.31250	0.10938	0.31250
n	78	0.71875	0.07813	0.40625
i	73	1.03125	0.09375	0.31250
v	86	1.40625	0.10938	0.37500
e	69	1.68750	0.17188	0.28125
r	82	1.96875	0.10938	0.28125
s	83	2.35938	0.14063	0.39063
i	73	2.70313	0.09375	0.34375
t	84	2.98438	0.10938	0.28125
y	89	3.21875	0.07813	0.23437

A keystroke feature vector is of the form (*key character, key code*, *x*), where *x* = (*t*, Δt_1, Δt_2). A keystroke feature vector sequence is generated after a word is typed. We collected two separate sets of keystroke feature vector sequences. The first set was used as a training set which contained 400 sequences typed by 40 users, 20 female and 20 male. Each user repeated the same text "University of Canberra" 10 times. The second set was the authentication set which contained 200 sequences typed by the same 40 users using the same text repeated 5 times.

2.2 Observable Markov Model

The key code sequence was considered as the Markov state sequence. Since this sequence was not hidden so the model was called observable Markov model.

Let $S = (s_1 s_2 ... s_T)$, $X = \{x_1 x_2 ... x_T\}$, $V = \{v_1, v_2, ..., v_K\}$, and $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_M\}$ be a state sequence, an observation sequence, a set of symbols and a set of user models, respectively. The compact notation $\lambda = \{\pi, A, B\}$ indicated the complete parameter set of a user model where

1. $\pi = \{\pi_i\}$, $\pi_i = P(s_1 = i)$: the initial state distribution, *i* = 1, …, *N*, where *N* was the number of states;
2. $A = \{a_{ij}\}$, $a_{ij} = P(s_t = j \mid s_{t-1} = i, \lambda)$, *i*, *j* = 1, …, *N*, *t* = 2, …, *T*: the state transition probability distribution, and
3. $B = \{b_j(k)\}$, $b_j(k) = P(x_t = v_k \mid s_t = j, \lambda)$, *j* = 1, …, *N*, *k* = 1, …, *K*, *t* = 1, …, *T*: the observation probability distribution, denoting the probability that an observation x_t was generated in state *j*.

The state probabilities were calculated as follows

$$\pi_i = 1 \quad \text{if} \quad i = i^* \text{ and } \pi_i = 0 \quad \text{if} \quad i \neq i^* \tag{1}$$

$$a_{ij} = n_{ij} \Big/ \sum_{s=1}^{N} n_{is} \tag{2}$$

where n_{ij} denoted the number of pairs of state *i* followed by state *j* observed in the keystroke sequences. Note that the following properties are still held

$$\sum_{i=1}^{N} \pi_i = 1 \qquad \sum_{j=1}^{N} a_{ij} = 1 \tag{3}$$

The observation probability was calculated as follows

$$b_j(k) = \sum_{t \in x_t = v_k} \sum_{i=1}^{K} \xi_t(i,j) \Big/ \sum_{t=1}^{T} \sum_{i=1}^{K} \xi_t(i,j) \qquad (4)$$

where

$$\xi_t(i,j) = P(s_t = i, s_{t+1} = j \mid X, \lambda) \qquad (5)$$

The set of symbols $V = \{v_1, v_2, ..., v_K\}$ was the set of codewords in a codebook obtained by applying a vector quantization method to the set of keystroke sequences $X = \{x_1 x_2 ... x_T\}$, where x = (t, Δt_1, Δt_2). In our experiments, the number of symbols K was set to the number of states N and there was a codeword per state.

The probability $P(X \mid \lambda)$ was then calculated as follows

$$P(X \mid \lambda) = \sum_{all\ S} P(X \mid S, \lambda) P(S \mid \lambda)$$

$$= \sum_{all\ S} \prod_{t=1}^{T} a_{s_{t-1} s_t} b_{s_t}(x_t) \qquad (6)$$

3 Keystrokes Biometric Authentication

For a given input keystroke sequence X and a claimed identity, let λ_0 be the claimed user model and θ be a predefined decision threshold. The simplest decision making method is to use the absolute likelihood score as follows

$$S(X) = P(X \mid \lambda_0) \begin{cases} > \theta & accept\ \lambda_0 \\ \leq \theta & reject\ \ \lambda_0 \end{cases} \qquad (7)$$

where $S(X)$ is referred to as the similarity score of the given keystroke sequence X. This score is strongly influenced by variations in the input keystroke sequence such as the typing speed, keyboard type (laptop or desktop) and keyboard quality. It is very difficult to set a common decision threshold to be used over different users. This drawback is overcome to some extent by using normalisation. Figure 2 presents a typical user authentication system.

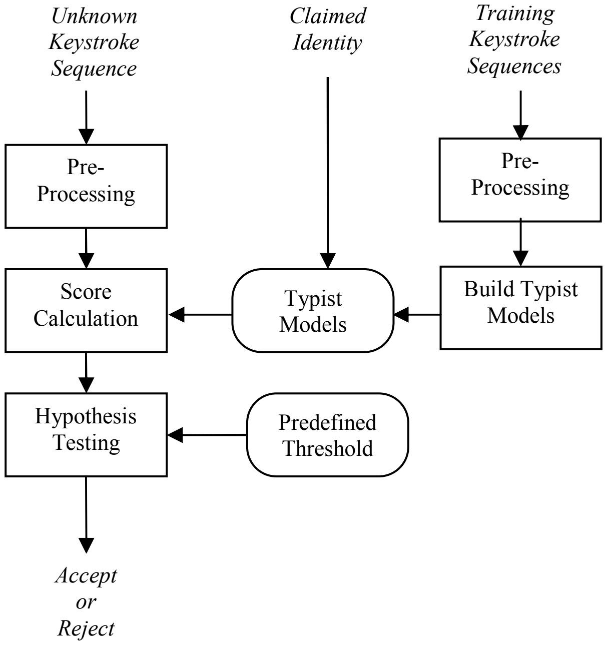

Fig. 2. A typical keystroke biometrics-based user authentication system.

In statistical approach, the authentication problem is usually formulated as a problem of statistical hypothesis testing [13]. Let λ be a model representing all other possible users, i.e. impostors. The problem formulation is to test the null hypothesis H_0: X is from the claimed user λ_0, against the alternative hypothesis H: X is from the impostors λ. If the probabilities of both the hypotheses are known exactly, according to Neyman-Pearson's Lemma, the optimum test to decide between these two hypotheses is a likelihood ratio test given by

$$S(X) = \frac{P(X \mid H_0)}{P(X \mid H)} \begin{cases} > \theta & accept\ H_0 \\ \leq \theta & reject\ \ H_0 \end{cases} \qquad (8)$$

However, in any practical authentication problem, it is impossible to obtain the exact probability density functions for either the null hypothesis or the alternative hypothesis. A parametric form of the distribution under each hypothesis is assumed to estimate these probability density functions. Let $P(X \mid \lambda_0)$ and $P(X \mid \lambda)$ be the likelihood functions of the claimed user and impostors, respectively. The similarity score is calculated as follows

$$S(X) = \frac{P(X \mid \lambda_0)}{P(X \mid \lambda)} \begin{cases} > \theta & accept\ H_0 \\ \leq \theta & reject\ \ H_0 \end{cases} \qquad (9)$$

The denominator $P(X \mid \lambda)$ is called the normalization term and requires calculation of all impostors' likelihood functions. However when the size of the population increases, a subset of the impostor models consisting of B "background" user models λ_i, $i = 1,...,B$ is used [14] and is representative of the population close to the claimed user.

Depending on the approximation of $P(X\,|\,\lambda)$ in (9) by the likelihood functions of the background model set $P(X\,|\,\lambda_i)$, $i = 1, \ldots, B$, we obtain different normalization methods. An approximation used in speaker authentication [15] is the arithmetic mean (average) of the likelihood functions of B background user models. The corresponding score for this approximation is as follows

$$S_1(X) = \frac{P(X\,|\,\lambda_0)}{\frac{1}{B}\sum_{i=1}^{B} P(X\,|\,\lambda_i)} \qquad (10)$$

If the geometric mean is used instead of the arithmetic mean to approximate $P(X\,|\,\lambda)$, we obtain the normalization method [16] as follows

$$S_2(X) = \frac{P(X\,|\,\lambda_0)}{\left[\prod_{i=1}^{B} P(X\,|\,\lambda_i)\right]^{1/B}} \qquad (11)$$

4 Fuzzy Normalization Method

4.1 Fuzzy Membership Scores

Consider the user authentication problem in fuzzy set theory [17]. To accept or reject the claimed user, the task is to make a decision whether the input keystroke feature sequence X is either from the claimed user λ_0 or from the set of impostors λ, based on comparing the score for X and a decision threshold θ. The space of input keystroke feature sequences can be considered as consisting of two fuzzy subsets for the claimed user and impostors. The similarity score means the fuzzy membership function, which denotes the degree of belonging of an input keystroke feature sequence to the claimed user. Accepting (or rejecting) the claimed user is viewed as a defuzzification process, where the input keystroke feature sequence is (or is not) in the claimed user's fuzzy subset if the fuzzy membership value is (or is not) greater than the given threshold θ. According to this fuzzy set theory-based viewpoint, currently used scores might be viewed as fuzzy membership scores and inversely, other fuzzy memberships can be used as the claimed user's scores.

In theory, there are many ways to define the fuzzy membership function, therefore it can be said that this fuzzy approach proposes more general scores than the current likelihood ratio scores for user authentication. These are termed fuzzy membership scores, which can denote the belonging of X to the claimed user. Based on this discussion, both the above-mentioned likelihood-based scores in (4) and (5) can also be viewed as fuzzy membership scores if their values are scaled into the interval [0, 1]. The next task is to find more effective fuzzy membership scores which can reduce both false rejection and false acceptance errors.

4.2 False Rejection Problem

Consider the false rejections of the claimed user and the false acceptances of impostors caused in the current likelihood ratio-based scores. A false rejection of the claimed user can arise because of the use of the background user set. The likelihood values of the background users are assumed to be equally weighted. However, this assumption is often not true as the similarity measures between each background user and the claimed user might be different. This drawback can be overcome by considering the user authentication in fuzzy set theory framework. From fuzzy clustering methods, the fuzzy c-means (FCM) membership score [18] is proposed as follows

$$S_3(X) = \frac{\left[-\log P(X\,|\,\lambda_0)\right]^{1/(1-m)}}{\sum_{i=0}^{B}\left[-\log P(X\,|\,\lambda_i)\right]^{1/(1-m)}} \qquad (12)$$

and the fuzzy entropy (FE) membership score [19] is proposed as follows

$$S_4(X) = \frac{\left[P(X\,|\,\lambda_0)\right]^{1/n}}{\sum_{i=0}^{B}\left[P(X\,|\,\lambda_i)\right]^{1/n}} \qquad (13)$$

where $m > 1$ and $n > 0$ control degree of fuzziness and degree of fuzzy entropy, respectively.

As an extension, a transformation is established to relate these fuzzy membership scores to currently used likelihood ratio scores. The proposed transformation is of the form

$$S_f(X) = \frac{f\left[P(X\,|\,\lambda_0)\right]}{f\left[P(X\,|\,\lambda)\right]} \qquad (14)$$

where $f[P]$ is a function of P. With $f[P] = (-\log P)^{1/(1-m)}$ and $f[P] = (P)^{1/n}$, we obtain the scores in (6) and (7), respectively.

4.3 False Acceptance Problem

The use of the normalization term can cause false acceptances of impostors because of the relativity of the ratio-based values. For example, the two ratios of (0.06 / 0.03) and (0.000006 / 0.000003) have the same value of 2. The first ratio can lead to a correct decision whereas the second one is unlikely since both likelihood values are very low. This problem can be overcome by applying the idea of the well-known noise clustering method proposed by Davé [20] in fuzzy clustering, where impostors' keystroke sequences are considered as noisy data and thus should have arbitrarily small fuzzy membership scores in the claimed user's fuzzy subset. This is implemented by simply adding to the normalization term a constant membership value $\varepsilon > 0$, which denotes the belonging of all input keystroke sequences to impostors' fuzzy subset.

The general form of the proposed scores after considering the false acceptances and the false rejections is proposed as follows

$$S_{f\varepsilon}(X) = \frac{f[P(X \mid \lambda_0)]}{f[P(X \mid \lambda) + \varepsilon]} \tag{15}$$

Applying the general form to all proposed scores, we obtain the following scores

$$S_{3\varepsilon}(X) = \frac{[-\log P(X \mid \lambda_0)]^{\frac{1}{1-m}}}{\sum_{i=0}^{B} [-\log P(X \mid \lambda_i)]^{\frac{1}{1-m}} + (-\log \varepsilon)^{\frac{1}{1-m}}} \tag{16}$$

$$S_{4\varepsilon}(X) = \frac{[P(X \mid \lambda_0)]^{1/n}}{\sum_{i=0}^{B} [P(X \mid \lambda_i)]^{1/n} + \varepsilon^{1/n}} \tag{17}$$

4 Experimental Results

The Markov modeling method was applied to build 40 user models. Experiments were performed on 40 users using each user as a claimed user with 3 closest background users and rotating through all users. The total number of claimed test keystroke sequences and impostor test keystroke sequences are 400 (40 claimed users x 10 test sequences) and 7800 ((40 x 39) impostors x 5 test sequences), respectively. Equal error rate (false acceptance rate = false rejection rate) results are shown in Table 2.

Table 2. Equal error rate (EER) results (%) for verifying 40 users, where $m = 2.0$, $n = 0.5$ and $\log \varepsilon = -31.0$ were applied

Similarity Score	User Authentication Error Rate (%)		
	Female	Male	Average
$S_1(X)$	17.8	18.8	18.3
$S_2(X)$	29.5	7.4	18.5
$S_3(X)$	**16.4**	**4.9**	**10.7**
$S_4(X)$	**20.2**	**4.6**	**12.4**
$S_{3\varepsilon}(X)$	**16.1**	**1.1**	**8.6**
$S_{4\varepsilon}(X)$	**20.0**	**2.8**	**11.4**

The current normalization method $S_2(X)$ produced the highest equal error rate (EER) of 18.5% and the proposed method $S_{3\varepsilon}(X)$ produced the lowest EER of 8.6%. The table shows that all the proposed methods perform better than the current methods.

5 Conclusion

Markov modeling and fuzzy normalization methods based on fuzzy set theory for user authentication have been presented and evaluated in this paper. Markov modeling method provides a simple yet efficient model for password modeling. Fuzzy normalization methods can provide better similarity scores, which can reduce both false acceptance and false rejection errors. The normalization method based on fuzzy c-means clustering and noise clustering has produced the better results than the fuzzy entropy-based method and current normalization methods.

References:

[1] L. O'Gorman, "Comparing Passwords, Tokens, and Biometrics for User Authentication", in Proceedings of the IEEE, 2003, vol. 91, no. 12, pp. 2021-2040.

[2] R.M. Bolle, J. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior, "Biometrics 101", IBM Research Report, IBM T. J. Hawthorne, New York, 2002.

[3] A. M. Namboodiri and A. K. Jain, "On-line Script Recognition", in Proceedings of the Sixteenth International Conference on Pattern Recognition, Canada, 2002, pp. 736-739.

[4] A. Jain, L. Hong, and S. Pankanti, "Biometric Identification", 2000, Communications of the ACM, vol.43, no.2, pp. 90-98.

[5] R. Gaines, W. Lisowski, S. Press, and N. Shapiro, "Authentication by keystroke timing: some preliminary results", Rand Report R-256-NSF. Rand Corporation, 1980.

[6] J. Leggett, G. Williams, M. Usnick, and M. Longnecker, "Dynamic identity verification via keystroke characteristics", International Journal of Man-Machine Studies, 1991, vol. 35, pp. 859–870.

[7] M. Obaidat and S. Sadoun, "Verification of computer users using keystroke dynamics", IEEE Transactions on Systems, Man and Cybernetics, Part B:P Cybernetics, 1997, vol. 27, no. 2, pp. 261–269.

[8] E. Yu and S. Cho, “Novelty Detection Approach for Keystroke Dynamics Identity Verification”, J. Liu et al. (Eds.): IDEAL 2003, LNCS 2690, pp. 1016–1023.

[9] S. Mandujano and R. Soto, “Deterring Password Sharing: User Authentication via Fuzzy c-Means Clustering Applied to Keystroke Biometric Data”, Proceedings of the Fifth Mexican International Conference in Computer Science (ENC’04), 2004, pp. 181-187.

[10] S. Hocquet, J-Y. Ramel, H. Cardot, “Fusion of Methods for Keystroke Dynamic Authentication”, in Proceedings of the Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID’05), 2005, pp. 224-229.

[11] M. Villani, C. Tappert, G. Ngo, J. Simone, H. St. Fort, S.-H. Cha, “Keystroke Biometric Recognition Studies on Long-Text Input under Ideal and Application-Oriented Conditions”, in Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW’06), 2006, pp. 39-46.

[12] W. Chang, “Reliable Keystroke Biometric System Based on a Small Number of Keystroke Samples”, G. M¨uller (Ed.): ETRICS 2006, LNCS 3995, 2006, pp. 312–320.

[13] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, New York: Academic, 1972.

[14] A. L. Higgins, L. Bahler and J. Porter: “Speaker Verification using Randomized Phrase Prompting”, Digital Signal Processing, 1991, vol. 1, pp. 89-106.

[15] D. A. Reynolds: “Speaker identification and verification using Gaussian mixture speaker models”, Speech Communication, 1995, vol. 17, pp. 91-108.

[16] C. S. Liu, H. C. Wang and C.-H. Lee: “Speaker Verification using Normalization Log-Likelihood Score”, IEEE Trans. Speech and Audio Processing, 1980, vol. 4, pp. 56-60.

[17] L. A. Zadeh, “Fuzzy sets and their application to pattern classification and clustering analysis”, Classification and Clustering, edited by J. Van Ryzin, Academic Press Inc, 1977, pp. 251-282 & 292-299.

[18] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York and London, 1981.

[19] D. Tran and M. Wagner: “A Fuzzy Approach to Speaker Verification”, International Journal of Pattern Recognition and Artificial Intelligence, 2002, vol. 16, no. 7, pp. 913-925, World Scientific Publishing.

[20] R. N. Davé: “Characterization and detection of noise in clustering”, Pattern Recognition Letters, 1991, vol. 12, no. 11, pp. 657-664.