Biometric Access Control Through Numerical Keyboards Based on Keystroke Dynamics

Ricardo N. Rodrigues, Glauco F.G. Yared, Carlos R. do N. Costa, João B.T. Yabu-Uti, Fábio Violaro, and Lee Luan Ling

Laboratory of Pattern Recognition and Computer Networks,
Department of Communications, School of Electrical and Computer Engineering,
State University of Campinas, Albert Einstein Av., 400,
PO Box 6101, Postal Code 13083-852, Campinas, SP, Brazil
ricardonagel@gmail.com,
{glauco, ccosta, yabuuti, fabio, lee}@decom.fee.unicamp.br

Abstract. This paper presents a new approach for biometric authentication based on keystroke dynamics through numerical keyboards. The input signal is generated in real time when the user enters with target string. Five features were extracted from this input signal (ASCII key code and four keystroke latencies) and four experiments using samples for genuine and impostor users were performed using two pattern classification technics. The best results were achieved by the *HMM* (EER=3.6%). This new approach brings security improvements to the process of user authentication, as well as it allows to include biometric authentication in mobile devices, such as cell phones.

1 Introduction

Access control to computational systems has been becoming more important nowadays, and the most well known and usual mechanism to guarantee the security of the information systems is through user authentication by a password. However, this type of security mechanism is fragile. A neglegent user compromise the security mechanism when one uses fragile passwords like the birth date, phone numbers, etc. On the other hand, the cost and the simplicity of this classic security mechanism justify its use, and in some situations it remains as the principal mechanism, supplemented with other security strategies. The intention of this work is to improve the process of password authentication using biometric features.

The biometric technology employed in this work is typing biometrics, also known as keystroke dynamics. The typing biometrics is an authentication process of analyzing the user's typing rhythm in a terminal at a keyboard during the identification. Authentication based on keystroke dynamics is a continuous or static process. The static manner analyzes the input keyboard at a particular moment, for example when the user types his password, while the continuous approach analyzes all inputs in the keyboard during a user session [1].

D. Zhang and A.K. Jain (Eds.): ICB 2006, LNCS 3832, pp. 640–646, 2005.

The personal authenticational system proposed in this paper only captures the keystroke dynamics from numerical keyboards (numerical passwords). This numerical password based approach was firstly introduced in [11]. The use of only a numerical keyboard causes more complex problems than using a full computer keyboard since only one hand is used to enter the password; as consequences, less information is avaible for authentication.

The Numerical Typing Dynamic biometrics can be incorporated into mobile phones, Automated Teller Machine (ATM) systems and to control the access to restricted areas. The methodology adopted in this work has low processing cost, is non-intrusive and statically authenticate users (only consider the input when the user types his password).

2 Related Works

Biometric authentication by keystroke dynamics has been an active research area since 1990 [1]-[7]. In this section, we briefly provide an review of the previous.

- **Target string.** It is the string that will be provided by the user and monitored by the system. In [2] four target strings were used during the authentication (username, password, first name and last name). However, in some works the password is the only target. Another important aspect about the target string is the string size. In [3] the authors concluded that the numbers of errors in classification process increases as the length of target string decreases.
- **Amount of samples for obtaining the template.** Samples are collected during the user enrollment phase. Some or total of these form the system training set classifier. The number of samples varies largely in literature, varying from three to thirty samples per user [4]. In [1] the authors have observed that the minimum number of samples that does not compromise the system performance is around six samples per user.
- **Feature extraction.** Two of more often observed features during the typing are the period of time in which the key remains pressed and the keystroke latency that corresponds to the time interval to move between successive keys [4].
- **Adaptation mechanism.** Biometric features are subject to small changes over the time. Most of previous works in the literature seldomly mention this important aspect. To solve this problem, a suitable adaptation mechanism or a re-enrollment procedure can be implemented to keep the users templates updated. In [6] whenever a new positive authentication occurs, the users template are updated. The database updating consist of including the new sample, discarding the oldest one and re-training the user's reference feature model.
- **Classification.** In [1]-[3] the authors used statistical Classifiers in their experiments, such as k-means classifiers, Bayes decision rules, etc. In [4], [5] and [7], artificial neural networks have been used to identify the user.

3 Methodology

In this work, we limit our investigation over numerical passwords collected from numerical keyboards. This keyboard type can be found in cell phones, ATM machines and most other access control systems. Each password is composed of a sequence of eight numerical characters, which is robust and easily memorized. For classifier a Hidden Markov Models (*HMM*) is implemented and tested.

3.1 Feature Extraction and Test Database

Ten samples, each with eight numerical characters were collected from each user in each session of 4 sessions, totalizing in 40 samples per user. Twenty people have been invited to contribute with their password samples for the experiment.

Let n denote the password sample size (n characters) can contain some keystroke features. The keystroke feature vector of sample w of user account a can be expressed as $k_{a,w} = (k_1(a,w),\ k_2(a,w),\ \ldots,\ k_n(a,w))$. Each element $k_i(a,w)$, where $i \leq n$, represents one of the following features [4]:

- The time interval when a key remains pressed (Down-Up or DU). This is represented by the expression $DU_{a,w} = \{DU_1(a,w),\ldots,DU_n(a,w)\}$, where $DU_i(a,w) = T_{i.up}(a,w)$ - $T_{i.down}(a,w)$. $T_{i.up}(a,w)$ is the instant where the key i is released and $T_{i.down}(a,w)$ is the time instant when key i is pressed;
- The time interval until the next key is pressed (Up-Down or UD). This is represented by the expression $UD_{a,w} = \{UD_1(a,w),\ldots,UD_{n-1}(a,w)\}$, where $UD_i(a,w) = T_{i+1.down}(a,w)$ - $T_{i.up}(a,w)$;
- The time interval between two consecutive pressed keys (Down-Down or DD). This is represented by $DD_{a,w} = \{DD_1(a,w),\ldots,DD_{n-1}(a,w)\}$, where $DD_i(a,w) = T_{i+1.down}(a,w)$ - $T_{i.down}(a,w)$;
- The interval between two consecutive released keys (Up-Up or UU). This is represented by $UU_{a,w} = \{UU_1(a,w),\ldots,UU_{n-1}(a,w)\}$, where $UU_i(a,w) = T_{i+1.up}(a,w)$ - $T_{i.up}(a,w)$.

For illustration purpose, figure 1 shows the feature extraction for a user typing a given target string.

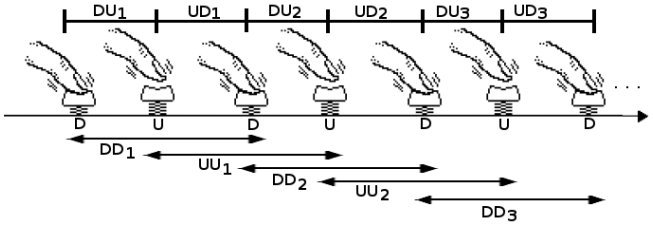

Fig. 1. Representation of the features observed during the typing of a given target string. DU is the time when a key remains pressed, UD is the time interval until the next key is be pressed, DD is the time interval between two consecutive pressed keys and UU is the time interval between two consecutive released keys.

3.2 Statistical Classifier

In [8] the authors suggest that the user template be composed of the mean and standard deviation of sample feature vector acquired during the enrollment. Every time a user needs to authenticate his password, the system calculates the distance of the target string to the template. If the distance is larger than the threshold, the user is authenticated. The template implemented is derived from sample feature vector K={DD, UD, DU, UU} according to equations (1) and (2)

$$\mu_{K_i(a)} = \frac{1}{N}\sum_{j=1}^{N} K_i(a,j), \tag{1}$$

$$\sigma_{K_i(a)} = \frac{1}{N-1}\sum_{j=1}^{N} |K_i(a,j) - \mu_{K_i(a)}|. \tag{2}$$

In authentication, through the ASCII code provided by the system determines the intending user and the retrieves the template from the corresponding account for authentication. The 1 to 1 comparison consists of computing the distance between the template and the input sample feature vector through the equation (3)

$$D_K(a,w) = \frac{1}{n}\sum_{i=1}^{n} \frac{K_i(a,w) - \mu_{K_i(a)}}{\sigma_{K_i(a)}}, \tag{3}$$

where n is the number of latency feature in K, K={DD, UD, DU, UU}. If $D_K(a,w) \leq \tau_k(a)$, for all K, the user is considered authentic. $\tau_k(a)$ is a empirically defined threshold for user account a. In [8] an data updating mechanism for the model is considered similarly with [6]. The authors reported error rates are below 2%.

3.3 Classifier Using *HMM*

HMM based systems have been widely used in pattern recognition [9]. This is due to the necessity of constructing models that analyze pattern's temporal variability. Moreover, the use of Gaussian mixtures allows ones to model complex distributions, and consequently build complex decision boundaries for classification processes.

The probabilistic model used to represent each user's password is modeled by a continuous *HMM* with 15 states and *left-to-right* topology. Each state is associated with the time when the user presses the key (down-up or DU) or with the time interval between two consecutive pressed keys (up-down or UD).

Let $A_{i(i+1)}$ denote the transition probability from state i to state $i+1$. Each new input feature value $K_i(a,w)$, i varying from 1 to 15 state is modeled by 6 Gaussian distribution. The input feature value $K_i(a,w)$ makes the *HMM* system advance from state i to state $i+1$.

The basic idea of *HMM* model as to associate each latency measure (observed feature) to a state of the model. Therefore, a 8 digits password will resulting in 15 latencies. For each typed password, we estimate the likelihood probability $P(O_{a,w}|\lambda_a)$ of the corresponding *HMM* model through the Viterbi algorithm [9], where

$$O_{a,w} = \{DU_1(a,w), UD_1(a,w), DU_2(a,w), UD_2(a,w), \ldots,$$

$$DU_{n-1}(a,w), UD_{n-1}(a,w), DU_n(a,w)\},$$

and λ_a is the set of feature value associated with user a, used for the model estimation.

If the likelihood probability estimate is superior to a threshold value, say a $\tau(a)$, the user is declared authentic; otherwise, he is considered an impostor. In this work, the threshold is obtained from the training data as follows:

$$\tau(a) = \mu_{P_a} - 3\sigma_{P_a} \tag{4}$$

where

$$\mu_{P_{(a)}} = \sum_{w=1}^{N} \frac{P(O_{a,w}|\lambda_a)}{N} \tag{5}$$

and

$$\sigma_{P_{(a)}} = \frac{1}{N-1} \sum_{w=1}^{N} |\mu_{P_{(a)}} - P(O_{a,w}|\lambda_a)| \tag{6}$$

For the system training the Baum-Welch algorithm implemented in Hidden Markov Toolkit (*HTK*[10]) is used. Although the system model was initially trained by a set of fixed N samples for each user a, the system has been trained again by most recent N true samples every time a new sample is positive authenticated.

4 Experiments and Results

The experiments were performed on a *Pentium IV* microcomputer platform, using only numeric keyboard part. Twenty users of both sex aging from 20 to 60 years old with different levels of familiarity with the numerical keyboard have participated in the experiment. The target strings are composed of eight numbers freely chosen by the users. Two kinds of data were collected: *Authentic database* were each user was undergone 4 sessions. In each session 10 sample was collected. This results in 800 samples in total. *Faked database* were for each true password, 30 samples are collected from ones other than the true user. This results in 600 faked samples in total.

For the statistical classifier design and system evaluations, three sets of training samples were used for model dimensioning:

I. Ten samples (N=10) to construct the model with K={DD,UD,DU,UU};
II. Twenty samples (N=20) to construct the model with K={DD,UD,DU,UU};
III. Thirty samples (N=30) to construct the model with K={DD,UD,DU,UU}.

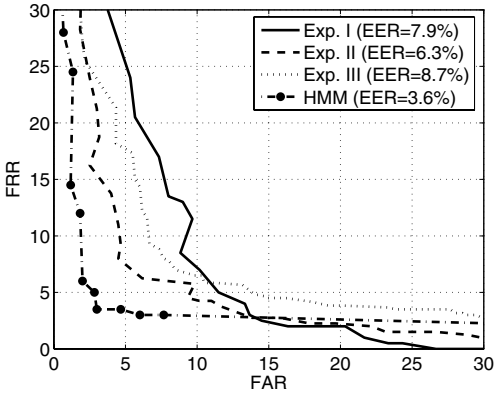

Fig. 2. The *ROC* performance of the experiments with the statistical classifiers (trained by N=10 samples, N=20 samples and N=30 samples) and HMM classifier trained by N=30 samples

In addition to the statistical classifiers, an *HMM* classifier was implemented, based on the using K={UD, DU} feature pattern. For model training, thirty samples (N=30) were used.

To represent False-acceptance and false-rejection rates (FAR and FRR), we used *ROC* (*Receiver Operating Characteristics*) curve. Each point of the *ROC* curve represent a specific operating condition of the biometric system, wich is a function of decision thresholds.

Figure 2 shows the *ROC* performance of the statistical classifiers design with different amounts of training samples and the HMM trained by 30 samples. Notice that the *EER* for the statistical classifier is considerably higher that supported in [8] (*EER* of 1.6%). However, this result is not surprising due to the fact that the experiments were made in an alphanumerical keyboard, in each user provided with his passwords with target string much bigger than 8 characters, what resulted in more representative latencies than in numerical keyboards. The performance of the *HMM* classifier (with N=30) wich outperforms all the 3 statistical classifiers in experiments. The method of updating the model allowed a more efficient form for modeling the typing dynamics of users.

5 Conclusions and Future Works

This work presented a novel methodology for biometric authentication based on the latency features extracted from the keystroke dynamics and *HMM* modelling approach. The biometric systems has the goal of improving the process of access control to restricted areas or to increase the security of banking transactions. The experimental results reveal the potentially of hidden Markov models, resulting in an *EER* of 3.6%. This rate is competitive when is compared to that obtained by the statistical classifier considered in [8].

We also observed that they include the influence of some practical aspects need to be considered in the analysis of the system performance. They include the familiarity of users with target strings, the updating mechanism, the precision of data acquisition and, mainly, the number of training samples.

For the future work, we intend to make our database even more robust in forms of population size and number of data acquisition sessions. Also other topologies of the *HMM* with different model structures should be investigated.

References

1. F. Monrose and A. D. Rubin, *Keystroke Dynamics as a Biometric for Authentication*, Future Generation Computer Systems, Vol. 16, no. 4, pp. 351-359, March 1999.
2. R. Joyce and G. Gupta, *Identity authentication based on keystroke latencies*, commun. ACM, vol. 33, no. 2, pp. 168-176, 1990.
3. d. Bleha and M. Obaidat, *Dimensionality reduction and feature extraction applications in identifying computer users*, IEEE Trans. Syst., Man, Cybern., Vol. 21, no. 2, pp. 452-456, Mar.-Apr. 1991.
4. D. T Lin, *Computer-access authentication with neural network based keystroke identity verification*, in Proc. Int. Conf. Neural Networks, vol. 1, 1997, pp. 174-178.
5. M. S. Obaidat and B. Sadoun, *Verification of computer user using keystroke dynamics*, IEEE Trans. Syst., Man, Cybern., vol. 27, no. 2, pp. 261-269, Mar.-Apr. 1997.
6. F. Monrose, M. K. Reiter, and S. Wetzel, *Password hardening based on keystroke dynamics*, in Proc. 6th ACM Conf. Computer Security, Singapore, Nov. 1999.
7. F. W. M. H. Wong, A. S. M. Supian, A. F. Ismail, L. W. Kin, and O. C. Soon, *Enhanced user authentication through typing biometrics with artificial neural network and k-nearest neighbor algorithm* in Conf. Rec. 35th Asilomar Conf. Signals, Syst., comput., Vol. 2, 2001, pp. 911-915.
8. L. C. F. Araújo, L. H. R. Sucupira Jr., M. G. Lizárraga, L. L. Ling, and J. B. T. Yabu-uti,*User authentication through typing biometrics features*, IEEE Trans. on Signal Processing, vol. 53, No. 2, Feb. 2005.
9. R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification* Wiley-Interscience Publication, 2nd Edition, Oct. 2000.
10. Cambridge University Engineering Departament, *The HTK Book*, Cambridge University, 2002.
11. T. Ord and S. M. Furnell, *User authentication for keypad-based devices using keystroke analysis*, Proc. Second International Network Conference (INC 2000), Plymouth, UK, pp. 263-272.