USER TUNED FEATURE SELECTION IN KEYSTROKE DYNAMICS

Jyothi Bhaskarr Amarnadh
Dept of Electronics and Communications
Indian Institute of Technology, North Guwahati-781039, Guwahati, India

Hugo Gamboa, Ana Fred
Instituto de Telecomunicações, Instituto Superior Técnico
IST - Torre Norte, Piso 10, Av. Rovisco Pais 1049-001 Lisboa, Portugal

Keywords: Pattern recognition, Biometrics, Keystroke Dynamics, Feature selection.

Abstract: In this paper, we present a new approach for user biometric verification based on keystroke dynamics. In our approach, the performance of simple classifiers (namely KNN and Bayes classifiers) is tested in a user tuned feature selection method, based on a open password approach. The impact of the training set size is studied, obtaining good results in a preliminary study on a population of 20 users.

1 INTRODUCTION

Keystroke dynamics is part of a class of biometrics known as behavioral biometrics. Behavioral biometrics is related to the dynamic characteristic traits of a person which can be used to determine his/her identity. Examples are Handwriting, Voice, Speech, Language, Gesture and typing patterns, etc.

In comparison to the other biometric techniques, probably keystroke dynamics is one of the easiest technique to implement. The reason is keystroke recognition is completely software-based solution and there is no need for any additional hardware. The already existent basic hardware in the context of a user at his computer, namely the keyboard, is sufficient for this technique.

1.1 Keystroke Dynamics

Keystroke dynamics (or typing rhythms) (Monrose and Rubin, 2000) has been shown to be a useful behavioral biometric technique. This method analyzes the way a user types on a terminal, by monitoring the keyboard input. Typing characteristics have been firstly identified in telegraphic communications where it has been coined as the "fist of the sender" (Bartlow and Cukic, 2006) given that the Morse code operators could be distinguished one from another from their typing rhythms.

The interest in Keystroke Dynamics as a new format for biometric identification has been recognized from the growing number of publications with increasing performance, along with new approaches for biometric security systems. As a maturity indication of this technique, we note that the International Committee for Information Technology Standards (INCITS) has already produced standardization guidelines for data format for Keystroke dynamics(Friant, 2006). The document defines the format for interchange of keystroke data, containing information related to the type of keyboard (standard, laptop, pda-keypad or pda-touchscreen among other) and the keyboard country/layout identification. It also establishes how the events information will be kept in the file specifying the format of input code (like if it is ASCII or UNICODE) and the time resolution used.

The techniques being proposed in the context of Keystroke Dynamics, can be divided into two modes: short code (log-in verification) and long code (continuous verification). The first case the user will type his user name and/or his password and the system uses this non-free text information to try to verify the user identity (Modi, 2005). The continuous approach the user is in an already authenticated environment and continues to be monitored in a free-text approach where, if the keystroke dynamics varies too much from the user model, the user can be asked for some other strong biometric information (Shepherd, 1995) to regain access to the system.

Bhaskarr Amarnadh J., Gamboa H. and Fred A. (2008).
USER TUNED FEATURE SELECTION IN KEYSTROKE DYNAMICS.
In *Proceedings of the First International Conference on Bio-inspired Systems and Signal Processing*, pages 546-550
DOI: 10.5220/0001069305460550

Any biometric technique is usually accompanied by some metrics which evaluate its performance (Jain et al., 2004). The rate at which the attempts of genuine user are falsely classified as impostor and rejected by the biometric system is called false rejection rate (FRR). The rate at which the attempts of an intruder are falsely classified as genuine and accepted by the biometric system is called false acceptance rate (FAR). The FAR and FRR indicate the errors that could possibly occur while making decision by the biometric system. The equal error rate (EER) is the error rate when FAR equals FRR. The receiving oprating curve (ROC) is the graph of FAR as a function of FRR. All these metrics (FAR,FRR,EER,ROC) depend on the collected data ie. the population of users and samples per each user. These are typical reported performance measures in the biometrics field that we will use in the rest of the paper.

We address some of the works conduced during the last years in the area. In (Bergadano et al., 2002) a continuous mode verification implementation provided a result of Equal error rate (EER) of 1.75% in a population of 44 users (extracting 4 samples per user) and 110 impostors. The users had to write 300 characters text of in approximately 2 minute period of acquisition. In (Hocquet et al., 2005) a fusion experiment in a short code mode with 15 users, three different classification algorithms were developed and preformed a fusion of the results obtaining a final value of 1.8% EER. Another study (de Magalhaes et al., 2005; Revett et al., 2006), reports the construction of a system based on short code login reporting a 5.8 EER in a population of 43 users. The report with more users to date has been conduced in the base of log-in short code mode (Jiang et al., 2007) with a population of 56 users presenting the result of 2.54 EER. We note that the population size is yet in a order of magnitude lower than the sizes used in other more conventional biometric techniques.

1.2 Our Proposal - User Tuned Feature Selection

In this paper, we present a new approach for keystroke dynamics based on a open password (all the users are aware of pass sentence). We test the performance of KNN and Bayes classifiers for user recognition with all the features obtained (namely press times of keystrokes and latency times of keystrokes) with all the users typing a common sentence of 23 characters.

We selected a set of features for each user by sequential backward feature selection, thereby improving the performance considerably (Silva, 2007; Siedlencki and Sklansky, 1993).

Table 1: WIDAM Data message.

Field	Bytes
message ID	1
event ID	1
object ID	1
relative position X	2
relative position Y	2
absolute position X	2
absolute position Y	2
other information	4
timestamp	4

In the following section, we describe the architecture of the data acquisition system. Section 3 presents the classifiers for user recognition and feature selection. Experimental results obtained from collected data are presented in section 4. Section 5 presents the conclusions and future work.

2 DATA ACQUISITION SYSTEM

The Keystroke Dynamics data was acquired using a web-based acquisition system of human computer interaction developed by the research group. The system called Web Interaction Display and Monitoring (WIDAM) (Gamboa and Ferreira, 2003). The system has the capabilities of monitoring the events that occur in a particular web page. Examples of the events are the mouse movements and clicks, and more relevant to this biometric technique, the keypress and keyup events generated while entering text in a web page form. The data recorded in the WIDAM system is listed in table 1, but the information required for this implementation is the following: (1) the type of event (if it is a keyup or keydown); (2) the keycode; (3) the timestamp of when the event occured; (3) keyboard modifiers flags, idicating if a shift, ctrl or alt key is being pressed.

For the Keystroke Dynamics experience, the user is presented with a Web Page instructing to insert several times a specific sentence. The sentence is common to all the users that we called the open-password. This sentence is used as the genuine and impostor data given that all the collected data for on user can be used as impostor data for all the other.

The WIDAM Architecture is composed by a client and server applications as depicted in figure 1.

The user accesses the WIDAM application via a web browser that connects to the server. Then the server sends back to the user a web page that is capable of monitoring and displaying the user interaction. This page creates a connection to the server and selects one of the services provided by WIDAM. Then

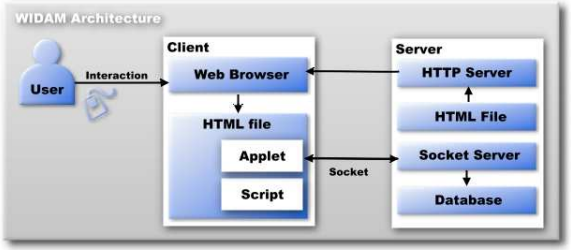

Figure 1: WIDAM architecture.

the client and the server exchange messages using a protocol defined by the authors.

In the server all the information is being recorded in a database in order to be accessed both on real time or off-line for the study of keystroke dynamics. In a previous work (Gamboa et al., 2007), the same structure was used to study the biometrics capabilities of the mouse movements dynamics.

3 RECOGNITION SYSTEM

The input to the recognition system is the press times of keystrokes and latency times of keystrokes from the users, recorded using the WIDAM module described previously.

3.1 Classifier System

We use the K nearest neighbor (KNN) (Duda et al., 2000) classifier at the first hand for user recognition. The nearest neighbors are determined based on the Euclidean distance of a testing sample from all the samples of the training data.

The Euclidean distance (d) between two samples $X = (x_1, x_2, x_3, ..., x_n)$ and $Y = (y_1, y_2, y_3, ..., y_n)$ is

$$d = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \tag{1}$$

We gather K nearest neighbors of a testing sample amongst all the samples from the training data, and depending on the majority of the nearest neighbors, we classify each attempt as either a genuine attempt or impostor attempt. If the majority of the nearest neighbors are from the training data of the genuine user, attempt is classified as genuine or else as an impostor attempt.

A Bayes classifier can be used if the distribution of training data is known. The technique of Bayes classifier is based on Bayesian theorem and is best suited when the size of training data is high. The posterior probabilities of each class for the measurement vector $\mathbf{X}$ and based on the maximum of all the posterior probabilities, we classify that $\mathbf{X}$ belongs to the class which has the maximum posterior probability, i.e. using the MAP rule.

$$\hat{w} = \arg\max_i p(w_i|\mathbf{X}) \tag{2}$$

The posterior probability that measurement vector $\mathbf{X}$ belongs to the class w_i is determined by

$$p(w_i|\mathbf{X}) = \frac{p(\mathbf{X}|w_i)p(w_i)}{p(\mathbf{X})} \tag{3}$$

where $p(\mathbf{X}|w_i)$ is the prior probability of the class membership. Since $p(\mathbf{X})$ is a scaling factor, the classification is based on maximum of the product $p(\mathbf{X}|w_i)p(w_i)$.

$$\hat{w} = \arg\max_i \{p(\mathbf{X}|w_i)p(w_i)\} \tag{4}$$

Since the measurement vector $\mathbf{X}$ is the set of all features, for simplicity of modeling we assume independence among the features conducing to the prior probability $p(\mathbf{X}|w_i)$ as follows:

$$p(\mathbf{X}|w_i) = \prod_{j=1}^{n} p(x_j|w_i) \tag{5}$$

where n is the total number of features.

3.2 Feature Selection

The classifiers described previously are implemented with the set of all the features and tested for performance. We select a subset of features from the set of all the features (all the press times of keystrokes and all the latency times of keystrokes) to best discriminate one user from the other (Jain and Ross, 2002).

For feature selection, we chose a set of features for each user which minimizes his/her performance metric M (we will detail latter the metrics used). To minimize M, sequential backward feature selection is used (Fukunaga, 1990). The algorithm for sequential backward feature selection is as follows:

1. Consider the set of all features $f = \{f_1, f_2, f_3, ..., f_n\}$ and M of f ie M_f.
2. Create a subset F_i by excluding the feature f_i from f and calculate M of F_i ie M_{Fi} for i=1,2,3,...,$n_{features}$ where $n_{features}$ is the number of features in f.
3. Choose F to be the set of features among all the sets F_i which has minimum M_{Fi} ; M_F is the minimum value of M_{Fi} for i=1,2,3,...,$n_{features}$.

4. If $M_F <= M_f$, then f=F and M_f=M_F and go to 2 ; else, go to 5.
5. The desired feature vector is $f_{desired}$=f and $M_{fdesired}$=M_f.

The performance metric M is FAR+αFRR in which α can be varied. We start feature selection by choosing $\alpha = 0$, so that M is FAR (ie. minimizing FAR alone) and we can increase α to infinity so that M is FRR (ie. minimizing FRR alone).We can achieve trade-off between FAR and FRR for intermediate values of α .

4 RESULTS

We implemented KNN and Bayes classifiers for a system which contains 20 users and each of the user typing the keyword for 17 times. We tested with 9 of the 17 samples as training data and rest 8 samples as testing data. When testing the system for one user, an impostor is considered to be the one of the other users.

When KNN classifier applied for the above system with the set of all the features for various values of K, the optimum value of K for the system is found to be K=1 and for this value of K, the FAR obtained is 1.48% and FRR obtained is 28.125%. Similarly, the system when being tested with Bayes classifier with the set of all features assuming the distribution of training data to be gaussian, the FAR obtained is 2.13% and FRR obtained is 40.625%.

Since the performance of KNN classifier is better compared to Bayes classifier, we used the KNN classifier to select the set of features as described in the previous section, to improve the performance of the system. Feature selection has been done for each user separately and the KNN classifier is implemented for the set of features obtained from feature selection for the respective user. For the overall system, on implementing the KNN classifier with feature selection using M=FAR ($\alpha = 0$)and 9 training samples, the FAR is minimized to 0% and FRR obtained is 21.25% for the optimum value of K=1. More importance is placed on FAR because it is vital to minimize the rate at which an intruder successfully bypasses the authentication system as a genuine user for any biometric system.

Since it is not practical to have 9 samples as the training data (asking the user to type for 9 times to gather the training data), we reapplied the feature selection by considering only 4 samples per user as the training data and 13 samples per user as the testing data.On implementing the KNN classifier with feature selection and using M=FAR ($\alpha = 0$), ignoring

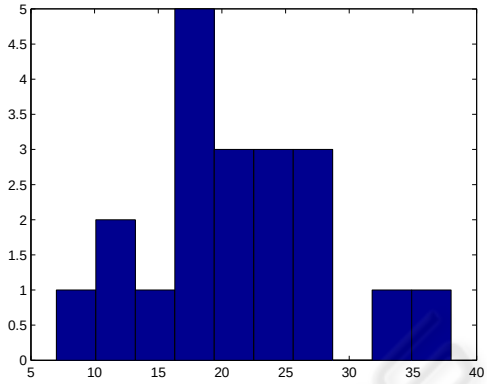

Figure 2: Histogram of length of feature vectors computed over 20 users.

FRR, the FAR obtained is 0.546% and FRR obtained is 56.59%. Since we cannot afford to have such a high FRR, we chose to have a trade-off between FAR and FRR ie. we decrease FRR at the expense of increasing FAR. This is achieved by applying sequential backward feature selection to minimize M=FAR+αFRR for a value of $\alpha > 0$.

Setting the limit for FRR to be 15%, on implementing feature selection for each user with the KNN classifier, the FAR obtained for the system is 0.8502% and FRR of the system is 15.00% for the value of $\alpha = 1$.

The average of length of features per user after feature selection is 21.35. Figure 1 represents the histogram of length of features computed over 20 users.

5 CONCLUSIONS

A novel approach has been presented based on behavioral biometric information obtained from Keystroke Dynamics. The technique identifies relevant keystroke typing patterns of a user by an user tuned feature selection. For the implementation of this technique, we used a system that includes a data acquisition module for the collection of keystroke data (time instances of key-up and key-down); the recognition module which includes classification system that uses the nearest K neighbors and decision rule for user recognition, and feature selection to identify user specific features to improve the performance of classification system.

The user authentication is based on the nearest neighbor method and initially all the features obtained from the data acquisition are used for recognition. A feature selection algorithm was applied which re-

duces the initial set of features to a subset of features which is unique for each user. Application of KNN classifier to the selected set of features for the respective users decides the authenticity of the user identity claim.

The results show that the proposed technique could be a competitive biometric technique which minimizes the rate at which an imposter bypasses the authentication system. Apart from that, as mentioned earlier, this technique does not require any additional hardware. The existent hardware namely keyboard is sufficient for this technique which makes it inexpensive.

The open password approach followed enabled a more complete study given that we had access to more impostor data, and validated the possibility of using a known sentence.

REFERENCES

Bartlow, N. and Cukic, B. (2006). Evaluating the reliability of credential hardening through keystroke dynamics. In *ISSRE '06: Proceedings of the 17th International Symposium on Software Reliability Engineering*, pages 117–126, Washington DC.

Bergadano, F., Gunetti, D., and Picardi, C. (2002). User authentication through keystroke dynamics. *ACM Trans. Inf. Syst. Secur.*, 5(4):367–397.

de Magalhaes, S. T., Revett, K., and Santos, H. M. D. (2005). Password secured sites: Stepping forward with keystroke dynamics. In *NWESP '05: Proceedings of the International Conference on Next Generation Web Services Practices*, page 293, Washington, DC, USA. IEEE Computer Society.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification*. Wiley-Interscience Publication.

Friant, D. (2006). Keystroke dynamics format for data interchange. Technical Report INCITS M1/06-0268, InterNational Committee for Information Technology Standards.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. Academic Press Professional, Inc., San Diego, CA, USA.

Gamboa, H. and Ferreira, V. (2003). Widam - web interaction display and monitoring. In *5th International Conference on Enterprise Information Systems, ICEIS'2003*, pages 21–27, Anger, France. INSTICC Press.

Gamboa, H., Fred, A., and Jain, A. (2007). Webbiometrics: User verification via web interaction. In *Biometrics Symposium, BCC, Baltimore, USA*.

Hocquet, S., Ramel, J.-Y., and Cardot, H. (2005). Fusion of methods for keystroke dynamic authentication. *autoid*, 0:224–229.

Jain, A., Ross, A., and Prabhakar, S. (2004). An introduction to biometric recognition. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(1):4– 20.

Jain, A. K. and Ross, A. (2002). Learning user-specific parameters in a multibiometric system. In *Proc. of International Conference on Image Processing (ICIP)*.

Jiang, C.-H., Shieh, S., and Liu, J.-C. (2007). Keystroke statistical learning model for web authentication. In *ASIACCS '07: Proceedings of the 2nd ACM symposium on Information, computer and communications security*, pages 359–361, New York, NY, USA. ACM Press.

Modi, S. K. (2005). Keystroke dynamics verfication using spontaneous password. Master's thesis, Purdue University.

Monrose, F. and Rubin, A. D. (2000). Keystroke dynamics as a biometric for authentication. *Future Generation Computer Systems*, 16(4).

Revett, K., de Magalhães, S. T., and Santos, H. M. D. (2006). Enhancing login security through the use of keystroke input dynamics. In *Advances in Biometrics, International Conference, ICB 2006, Hong Kong, China, January 5-7, 2006, Proceedings*, pages 661–667.

Shepherd, S. J. (1995). Continuous authentication by analysis of keyboard typing characteristics. In *European Convention on Security and Detection (ECOS 95)*, pages 111–114.

Siedlencki, W. and Sklansky, J. (1993). On automatic feature selection. In Chen, C. H., Pau, L. F., and Wang, P. S. P., editors, *Handbook of Pattern Recognition and Computer Vision*. World Scientific.

Silva, H. (2007). Feature selection in pattern recognition systems. Master's thesis, UNIVERSIDADE Técnica de Lisboa, Instituto Superior Técnico.