Capturing Cognitive Fingerprints from Keystroke Dynamics for Active Authentication

Journal:	*IT Professional*
Manuscript ID:	ITProSI-2013-02-0016
Manuscript Type:	SI: Jul/Aug 2013 - Security
Date Submitted by the Author:	02-Feb-2013
Complete List of Authors:	Chang, Morris J.; Iowa State University, Electrical & Computer Eng. Fang, Chi-Chen; Iowa State Universirt, Electrical & Computer Eng. Ho, Kuan-Hsing; Iowa State University, Electrical & Computer Eng. Kelly, Norene; Iowa State University, Human Development & Family Studies Wu, Pei-Yuan; Princeton University, Electrical Eng. Ding, Yixiao; Iowa State University, Electrical & Computer Eng. Chu, Chris; Iowa State University, Department of Electrical and Computer Engineering Gilbert, Stephen; Iowa State University, Human Computer Interaction Kamal, Ahmed; Iowa State University, Electrical and Computer Engineering Kung, Sun-Yuan; Princeton,
Keywords:	D.4.6.b Authentication < D.4.6 Security and Privacy Protection < D.4 Operating Systems < D Software/Software Engineering, I.5.2.c Pattern analysis < I.5.2 Design Methodology < I.5 Pattern Recognition < I Computing Methodologies, K.6.5.a Authentication < K.6.5 Security and Protection < K.6 Management of Computing and Information Systems < K Computing Milieux

Capturing Cognitive Fingerprints from Keystroke Dynamics for Active Authentication

J. Morris Chang, Chi-Chen Fang, Kuan-Hsing Ho, Norene Kelly, Pei-Yuan Wu, Yixiao Ding, Chris Chu, Stephen Gilbert, Amed E. Kamal, Sun-Yuan Kung

Introduction

Conventional authentication systems verify a user only during initial login. Active authentication performs verification continuously as long as the session remains active. This work focuses on using behavioral biometrics, extracted from keystroke dynamics, as "something a user is" for active authentication. This scheme performs continual verification in the background, requires no additional hardware devices and is invisible to users.

Keystroke dynamics, the detailed timing information of keystrokes when using a keyboard, has been studied for the past three decades. The typical keystroke interval time is expressed as the time between typing two characters, which is also known as a digraph. The keystroke rhythms of a user are distinct enough from person to person such that they can be used as biometrics to identify people. However, it has been generally considered much less reliable than physical biometrics such as fingerprints. The main challenge is the presence of within-user variability.

Due to within-user variability of interval times among identical keystrokes, most past efforts have focused on verification techniques that can manage such variability. For example, a method called Degree of Disorder (DoD) [1, 2] was proposed to cope with the time variation issues. It argued that while the keystroke typing durations usually vary between each sample, the order of the timing tends to be consistent. It suggested that the distance of the order between two keystroke patterns can be used to measure the similarity.

A recent paper [3] provided a comprehensive survey on biometric authentication using keystroke dynamics. This survey paper classified research papers based on their features extraction methods, feature subset selection methods and classification methods. Most of the systems described in this survey were based on typing rhythm of short sample texts, which is dominated by the physical characteristics of users and too brief to capture a "cognitive fingerprint." In the current keystroke authentication commercial market, some products combine the timing information of the password with password-based access control to generate the hardened password [4, 5, 6].

In this paper, we present a biometric-based active authentication system. This system continuously monitors and analyzes various keyboard behavior performed by the user. We extract the features from keystroke dynamics that contain cognitive factors, resulting in cognitive fingerprints. Each feature is a sequence of digraphs from a specific word. This method is driven by our hypothesis that a cognitive factor can affect the typing rhythm of a specific word. Cognitive factors have been largely ignored in the keystroke dynamics studies of the past three decades. The rest of this paper will detail our project's: (1) search for cognitive fingerprints; (2) building of an authentication system with machine learning techniques; and (3) results from a large scale experiment at Iowa State University.

Searching for cognitive fingerprints

Physical biometrics rely on physical characteristics such as fingerprints or retinal patterns. The behavioral biometric of keystroke dynamics must incorporate cognitive fingerprints to advance the field, but the cognitive fingerprint does not have a specific definition. We hypothesize that natural pauses (delays between typing characters in words) are caused by cognitive factors (e.g., spelling an unfamiliar word or after certain syllables) [7, 8, 9, 10, 11], which are unique among individuals. Thus, a cognitive factor can affect the typing rhythm of a specific word. In this research, each feature is represented by a unique cognitive typing rhythm (CPR) which contains the sequence of digraphs from a specific word. Such features include natural pauses among its timing information (e.g., digraphs) and could be used as a cognitive fingerprint. Conventional keystroke dynamics does not distinguish timing information between different words and only considers a collection of digraphs (e.g., tri-graphs or N-graphs). Cognitive factors, thus, have been ignored.

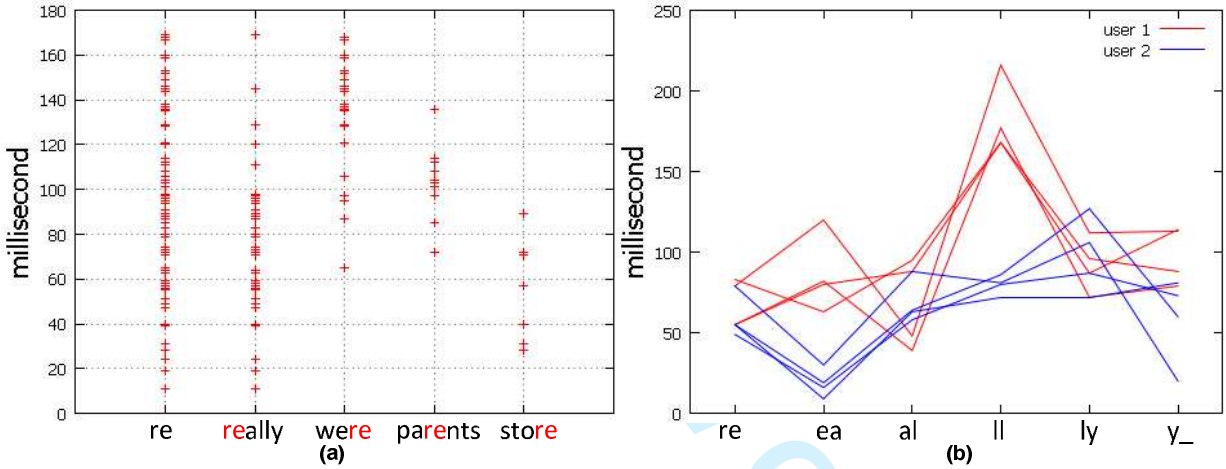


Figure 1. (a) Digraph "re" from the same user (b) Two users typed the same word "really"

As we can see from Figure 1(a), there is a collection of digraphs ("re") observed from the same user. One might think the collection of digraphs represent part of a keystroke rhythm. However, as we more closely examine each collection of digraphs, these digraphs are clustered around different words that contain the digraphs. For example, for the collection of digraphs "re", we can separate these digraphs according to four different words (i.e., *really*, *were*, *parents*, and *store*). This shows that examining digraphs in isolation might result in missing some important information related to specific words. This observation confirms our hypothesis: a cognitive factor can affect the typing rhythm of a specific word. Thus, we extract CPR from keystroke dynamics and use them as features (cognitive fingerprints) for active authentication. Each feature is a sequence of digraphs of a specific word (instead of a collection of digraphs). For each legitimate user, we collect samples of each feature and, then, build a classifier for that feature during the training phase of machine learning.

Building authentication system with machine learning techniques

We have developed two authentication systems based on two different machine learning techniques. The first one uses off-the-shelf SVM (support vector machine) library [12] while the second one employs an in-house developed library based on KRR (Kernel Ridge Regression) [13]. These libraries are used to build each classifier during the training phase. While it is not possible to know the patterns of all imposters, we use patterns from the legitimate user and some known imposters to build each classifier and expect that it can detect any potential imposter within a reasonable probability. This is a two-class (legitimate user vs. imposters) classification approach in machine learning. We build a trained profile with multiple classifiers for each legitimate user. During the testing phase (i.e., authentication), a set of testing data is given to the trained profile for verification. Each classifier under testing yields a matching score between the testing dataset and trained file. The final decision (accept or reject) is based on a sum of scores fusion method.

Other than differing basic machine learning libraries, the two systems share the same feature selection and fusion method. In the fusion method, we evaluate each classifier to determine the confidence level of its decision. Such evaluation is conducted during the training phase with datasets from each legitimate user and imposters. The basic idea is illustrated in Figure 2. A subset of the dataset is used to train a temporary classifier. The remaining dataset is used to test the classifier. Such testing will be repeated multiple times to ensure a good estimation. This technique is called cross-validation (a.k.a. rotation estimation).

From results of these tests, we can estimate the probabilities of true acceptance (P_{ta}) and false acceptance (P_{fa}) of the classifier. For example, after the testing with dataset from legitimate user, there are N acceptances out of M samples, P_{ta} is N/M. The confidence of decision (W_a) on acceptance is expressed as the ratio of P_{ta} to P_{fa}. The confidence of decision on rejection (W_r) is expressed as the ratio of the probability of true rejection (1-P_{fa}) to the probability of false rejection (1-P_{ta}).

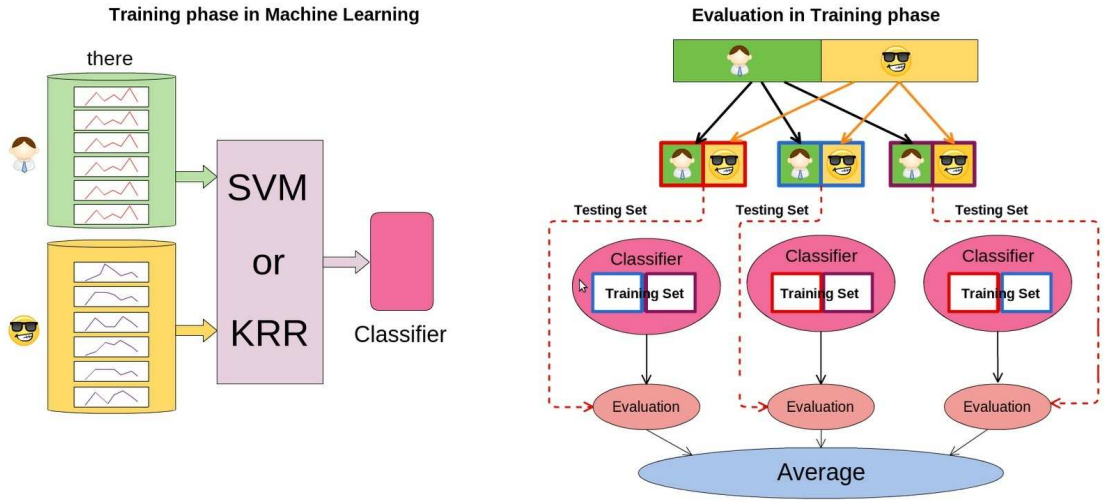


Figure 2. Training and cross-validation in machine learning

After the training, in the trained profile, there are W_a and W_r for each classifier. During the testing phase, each classifier generates a decision (acceptance or rejection). Either W_a or W_r will be applied to this decision. The final decision is based on the sum of scores of all involved classifiers.

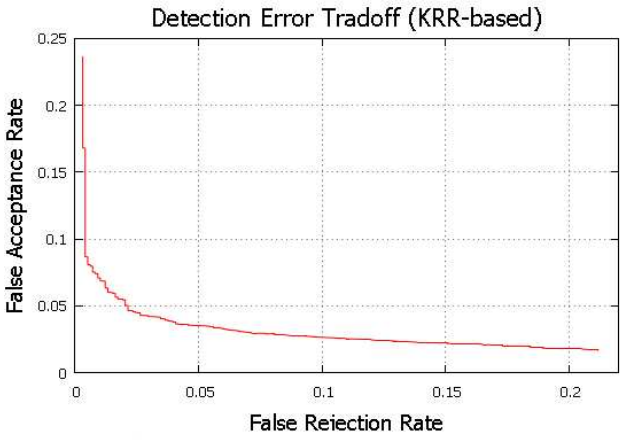
A large scale experiment at Iowa State University

For this project, we developed a web-based software system to collect the keystroke dynamics of individuals in large scale testing at Iowa State University. This web-based system provided three simulated user environments: typing short sentences, writing short essays, and browsing web pages. The users' cognitive fingerprints were stored in a database for further analyses. Machine learning techniques were used to perform pattern recognition to authenticate users.

During November and December of 2012, email invitations were sent to 36,000 members of the ISU community. There were 1,977 participants completed two segments that each lasted about 30-minutes, and resulted in about 900 words for each participant for each segment. In addition, 983 participants (out of the 1,977) completed another segment of approximately 30-minutes in length, in which about 1,200 words were collected for each participant. We then developed 983 individual profiles (trained files). Each profile was trained under two-class classification in which one legitimate user had 2,100 collected words and the imposter training set was based on collected words from other 982 known participants. Each profile was tested with the data of the 1,977 participants (testing dataset of 900 words per participant).

	SVM	KRR
FAR	0.055	0.055
FRR	0.007	0.0177
training time	15 m/user	29 s/user
testing time	0.6 s/user	3.5 ms/user
size of training file	20 MB/user	1 MB/user

(a)



(b)

Figure 3. Experiment results

The experiment results are presented in Figure 3 where the performance comparison of two verification systems is summarized in 3 (a), and the DET (Detection Error Tradeoff) chart from KRR-based system is given in 3 (b). In summary, the proposed scheme is effective for authentication and has been verified through a large-scale dataset.

References

[1] F. Bergadano *et al.*, "User authentication through keystroke dynamics". *ACM Trans. Inf. Syst. Secur.*, vol. 5, pp. 367–397, Nov. 2002.

[2] D. Gunetti and C. Picardi, "Keystroke analysis of free text," *ACM Trans. Inf. Syst. Security*, vol. 8, no. 3, pp. 312–347, Aug. 2005.

[3] M. Karnan *et al.*, "Biometric personal authentication using keystroke dynamics: A review," *Appl. Soft Computing*, vol. 11, no. 2, pp. 1565–1573, Mar. 2011.

[4] F. Monrose *et. al.*, "Password hardening based on keystroke dynamics," in *Proceedings of the 6th ACM Conference on Computer and Communications Security*, Singapore, Nov. 1999, pp. 73–82.

[5] AdmitOne Security, http://www.biopassword.com/index.asp

[6] ID Control, http://www.idcontrol.com/

[7] C.M. Levy and S. Ransdell, "Writing signatures," in *The Science of Writing: Theories, Methods, Individual Differences, and Applications*, C.M. Levy and S. Ransdell, Eds. Mahwah, NJ: Lawrence Erlbaum, 1996, pp. 149–162.

[8] D. McCutchen, "A capacity theory of writing: Working memory in composition," *Educational Psychology Review*, vol. 8, no. 3, pp. 299-325, Sept. 1996.

[9] D. McCutchen, "Knowledge, processing, and working memory: Implications for a theory of writing," *Educational Psychologist*, vol. 35, no. 1, pp. 13-23, 2000.

[10] T. Olive, "Working memory in writing: Empirical evidence from the dual-task technique," *European Psychologist*, vol. 9, no. 1, pp. 32-42, Dec. 2004.

[11] T. Olive *et al.*, "Verbal, visual, and spatial working memory demands during text composition," *Applied Psycholinguistics*, vol. 29, no. 4, pp. 669–687, Oct. 2008.

[12] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Syst. and Technology,* vol. 2, no. 3, article no. 27, Apr. 2011.

[13] S.Y. Kung," Kernel Methods and Machine Learning," Cambridge University Press, 2013