Chapter 28
A Keystroke Biometric System for Long-Text Input

Charles C. Tappert
Pace University, USA

Mary Villani
Pace University, USA

Sung-Hyuk Cha
Pace University, USA

Robert S. Zack
Pace University, USA

ABSTRACT

A novel keystroke biometric system for long-text input was developed and evaluated for user identification and authentication applications. The system consists of a Java applet to collect raw keystroke data over the Internet, a feature extractor, and pattern classifiers to make identification or authentication decisions. Experiments on more than 100 participants investigated two input modes—copy and free-text—and two keyboard types—desktop and laptop. The system can accurately identify or authenticate individuals if the same type of keyboard is used to produce the enrollment and questioned input samples. Longitudinal experiments quantified performance degradation over intervals of several weeks and two years. Additional experiments investigated the system's hierarchical model, parameter settings, assumptions, and sufficiency of enrollment samples and input-text length. Although evaluated on input texts up to 650 keystrokes, the authors found that input of 300 keystrokes, roughly four lines of text, is sufficient for the important applications described.

INTRODUCTION

This paper describes the development and evaluation of a keystroke biometric system for long-text input. The system has user-identification and user-authentication Internet applications that are of increasing importance as the population of application participants continues to grow. An example user-authentication application is verifying the identity of students taking online quizzes or tests, an application becoming more important with the student enrollment in online classes increasing and instructors becoming concerned about evaluation security and academic integrity. Similarly, in a business setting employees can be required to take online examinations in their training/orientation programs where the companies would like the exam-takers authenticated.

DOI: 10.4018/978-1-4666-2919-6.ch028

An example user-identification application in a small company environment is a closed system of known employees where there has been a problem with the circulation of inappropriate (unprofessional, offensive, or obscene) e-mail, and it is desirable to identify the perpetrator. Because the inappropriate email is being sent from computers provided by the company for employees to send email and surf the Internet during lunch and coffee breaks, there are no ethical issues in capturing users' keystrokes. In addition, as more businesses moving to e-commerce, the keystroke biometric in Internet applications can provide an effective balance between high security and customer ease-of-use (Yu & Cho, 2004).

Keystroke biometric systems measure typing characteristics believed to be unique to an individual and difficult to duplicate (Bolle, Connell, Pankanti, Ratha, & Senior, 2004; Jin, Ke, Manuel, & Wilkerson, 2004). The keystroke biometric is one of the less-studied behavioral biometrics. Most of the systems developed previously have been experimental in nature. However, several companies such as AdmitOne (2008) and BioChec (2008) have recently developed commercial products for hardening passwords (short input) in computer security schemes.

The keystroke biometric is appealing for several reasons. First, it is not intrusive and computer users type frequently for both work and pleasure. Second, it is inexpensive since the only hardware required is a computer with keyboard. Third, keystrokes continue to be entered for potential repeated checking after an authentication phase has verified a user's identity (or possibly been fooled) since keystrokes exist as a mere consequence of users using computers (Gunetti & Picardi, 2005). This continuing verification throughout a computer session is sometimes referred to as dynamic verification (Leggett & Williams, 2005; Leggett, Williams, Usnick, & Longnecker, 1991).

Most of the previous work on the keystroke biometric has dealt with user authentication, and while some studies used long-text input (Bergadano, Gunetti, & Picardi, 2002; Gunetti & Picardi, 2005; Leggett & Williams, 2005), most used passwords or short name strings (Bender & Postley, 2007; Bolle et al., 2004; Brown & Rogers, 1993; Giot, El-Abed, & Rosenberger, 2009a; Monrose, Reiter, & Wetzel, 2002; Monrose & Rubin, 2000; Obaidat & Sadoun, 1999; Revett, 2008; Rodrigues et al., 2006). Fewer studies have dealt with user identification (Gunetti & Picardi, 2005; Peacock, Ke, & Wilkerson, 2004; Song, Venable, & Perrig, 1997). Gunetti and Picardi (2005) focused on long free-text passages, similar to this research, and also attempted the detection of uncharacteristic patterns due to fatigue, distraction, stress, or other factors. Song et al. (1997) touched on the idea of detecting a change in identity through continuous monitoring.

Researchers tend to collect their own data and no known studies have compared techniques on a common database, although a recent study made a password database available to the scientific community (Giot, El-Abed, & Rosenberger, 2009b). Nevertheless, the published literature is optimistic about the potential of keystroke dynamics to benefit computer system security and usability (Woodward, Orlans, & Higgins, 2002). Gunetti and Picardi (2005) suggest that if short inputs do not provide sufficient timing information, and if long predefined texts entered repeatedly are unacceptable, we are left with only one possible solution, using users' normal keyed text-input interactions with computers, *free text*, as we do in this research.

Generally, a number of measurements or features are used to characterize a user's typing pattern. These measurements are typically derived from the raw data of key press times, key release times, and the identity of the keys pressed. From key-press and key-release times a feature vector, often consisting of keystroke duration times and keystroke transition times, can be created (Woodward et al., 2002). Such measurements can be collected from all users of a system, such as a computer network or web-based system, where