Applied Soft Computing 11 (2011) 1565–1573

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Review Article

Biometric personal authentication using keystroke dynamics: A review

M. Karnan [a,*], M. Akila [b], N. Krishnaraj [c]

[a] Tamilnadu College of Engineering, Coimbatore, India
[b] Vivekanandha College of Engineering for Women, Tiruchengode, India
[c] M.S. University, Tirunelveli, India

ARTICLE INFO

Article history:
Received 29 January 2009
Received in revised form 7 September 2009
Accepted 1 August 2010
Available online 17 August 2010

Keywords:
Biometrics
Keystroke dynamics
Pattern recognition
Neural networks
False alarm rate
Imposter pass rate

ABSTRACT

Authentication is the process of determining whether someone or something is, in fact, who or what it is declared to be. As the dependence upon computers and computer networks grows, the need for authentication has increased. Biometrics is the science and technology of authentication by identifying the living individual's physiological or behavioral attributes. Keystroke dynamics is a behavioral measurement and it utilizes the manner and rhythm in which each individual types. The approaches in keystroke dynamics can be categorized by the selection of features and the classification methods employed. The objective of this review paper is to summarize the well-known approaches used in keystroke dynamics in the last two decades.

Contents

1. Introduction

Authentication [1,2] is the process of verifying whether the digital identities of computers and the physical identities of people are authentic. User authentication is the process of verifying the identity of a person. There are multiple authentication technologies that verify the identity of a user before granting access to system resources. However, these technologies provide different levels of security, and none can be said to secure a system completely. User's claimed identity [3] can be verified by one of the following methods [4]:

- The user presents a secret (something they know), such as a password or Personal Identification Number (PIN). Passwords are the commonly used method but individual user generally chooses very easy passwords for a machine to guess. Even the best password can be stolen by dictionary and brute force attacks. A digital thief who is armed with the password would appear to be the legitimate system user.

* Corresponding author. Tel.: +91 4545244385.
E-mail address: karnanme@yahoo.com (M. Karnan).

1568-4946/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2010.08.003

- Often, a token (in the user's possession) such as a smart card or key is presented for identification. However, in a multifactor environment stronger user identification techniques such as combining password/PIN with the tokens are needed. In spite of enhanced security users find smart cards and tokens inconvenient because they are susceptible to loss or theft, and they must be kept close at hand. Authentication methods that require user to remember assigned PIN numbers can cause some problems for those who forget their PIN numbers or make typographical errors, especially if the smart card gets locked after a certain number of attempts.
- The user presents a personal physical attribute (something they are), such as fingerprint (Biometric authentication). Utilizing biometrics for personal authentication is becoming convenient and considerably more accurate than the previous methods. This is because biometrics links the event to the claimed identity while password or token may be used by someone else. This method is becoming socially acceptable because it is convenient (nothing to carry or remember), accurate (provides for positive authentication), and can provide an audit trail.

With an aim to give a comprehensive and critical survey of methods employed in keystroke dynamics, this paper is organized as follows: Section 2 discusses the overview of Biometrics. Various approaches towards research work on keystroke dynamics in the last two decades are reported in Section 3; and finally the conclusions are given in Section 4.

2. Biometrics

Biometric authentication [5] is an automatic method that identifies a user or verifies the identity based upon the measurement of his or her unique physiological traits (face [6], palm [7], iris [8], etc.) or behavioral characteristics (voice [9], handwriting [10], signature [11], keystroke dynamics [12], etc.). Physiological biometrics are biological/chemical traits that are innate or naturally grown to; and behavioral biometrics are mannerisms or traits that are learned or acquired. Biometrics can be broadly grouped into following categories [13]. They are Access control and attendance, Computer and enterprise network control, Financial and health services, Government, Law enforcement and Telecommunications.

Both physiological and behavioral systems can be logically divided into two, namely, enrollment phase and authentication/verification phase. During the enrollment phase as shown in Fig. 1 user biometric data is acquired, processed and stored as reference file in a database. This is treated as a template for future use by the system in subsequent authentication operations. During the authentication/verification phase user biometric data is acquired, and processed. The authentication decision shall be based on the outcome of a matching process of the newly presented biometric to the pre-stored reference templates.

There are two basic types of recognition errors: the False Alarm Rate (FAR) and the Imposter Pass Rate (IPR) [14]. FAR is the percentage of genuine users incorrectly categorized as imposters and IPR is the percentage of imposters incorrectly matched to a genuine user's reference template. Equal Error Rate (EER) is the rate of setting at which both false alarm and imposter pass errors are equal. EER is also known as the cross-over error rate (CER). The lower the ERR (or CER), more accurate is the system. The overall performance of a biometric system is assessed in terms of its accuracy, speed, storage, cost and ease-of-use.

3. Keystroke dynamics

Keystroke dynamics is a behavioral measurement and it aims to identify users based on the typing of the individuals or attributes such as duration of a keystroke or key hold time, latency of keystrokes (inter-keystroke times), typing error, force of keystrokes etc. The analogy is made to the days of telegraphy when operators identify each other by recognizing "the fist of the sender" [15,16]. Both the National Science Foundation (NSF) and National Institute of Standards and Technology (NIST), United States of America have conducted studies establishing that typing patterns are unique to the typist. The advantages of keystroke dynamics are obvious in computer environment as it provides a simple natural method for increased computer security. Static keystroke analysis is performed on typing samples produced using predetermined text for all the individuals under observation. Dynamic analysis implies a continuous or periodic monitoring of issued keystrokes. It is performed during the log-in session and continues after the session.

Over the years, researchers have evaluated different features/attributes, feature extraction, feature subset selection and classification methods in an effort to improve the recognition capabilities of keystroke biometrics.

3.1. Features and feature extraction methods

The feature extraction is used to characterize attributes common to all patterns belonging to a class. A complete set of discriminatory features for each pattern class can be found using feature extraction. Major feature extraction methods and features/attributes are detailed in this section. Gaines et al. [17] use statistical significance tests between 87 lowercase letter interkey latencies to check if the means of the keystroke interval times are the same. Young and Hammon [18] experimented with the time periods between keystrokes, total time to type a predetermined number of characters, or the pressure applied to the various keys and used it to form the feature template. Garcia [19] creates a template using the mean and covariance of keystroke interval times. If the verification vector is statistically similar to the template vector, then the attempt will be classified as an authentic attempt. Joyce and Gupta [3] propose two additional login sequences: the user's first name and the last name as the feature subset. This improved the performance of the method considerably. Obaidat and Sadoun [20] suggest inter-key and key hold times to be recorded using Terminate and Stay resident (TSR) program in MS-DOS based environment.

The standard keyboard interrupt handler was replaced by a special scan codes to record the time stamp. Lin [12] suggests a modified latency measurement to overcome the limitation of negative time measure, i.e. when the second key is pressed before the release of the first key. William and Jan [21] propose typing difficulty feature in the feature subset to increase categorization. Robinson and Liang [22] use user's login string to provide a characteristic pattern that is used for identity verification. In [23], the authors examine the use of keystroke duration and latency between keystrokes and combine it with the user's password. Monrose and Rubin [24] propose that users can be clustered into groups comprising disjoint feature sets in which the features in each set are pair wise correlated. In [25], box plot algorithm was used as a graphical display with many features and the features extracted was normalized using normal bell curve algorithm. Francesco et al. [26] use timing information to obtain the relative order of trigraphs. It is used to compare two different sets of sorted trigraphs and to measure the difference in the ordering between them. In [27], the users were asked to type the usual pass-

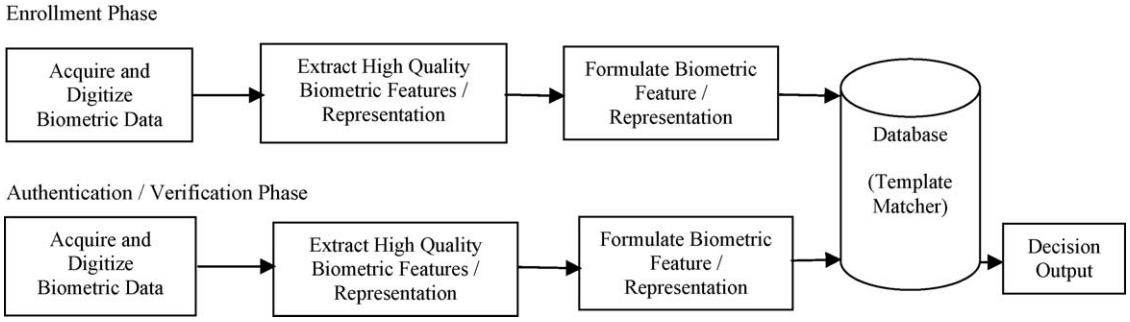

Fig. 1. Biometric system.

word, or passphrase twelve times to get the digraph for further processing.

Chen et al. [28] extracted features from the frequency domain signal which include mean, root mean square, peak value, signal in noise and distortion, total harmonic distortion, fundamental frequency, energy, kurtosis, and skew ness. Nick and Bojan [29] incorporated shift-key pattern along with keystroke inter-key and hold times. Kenneth [30] proposes motif signature which is used for classification.

In order to improve the quality of data Pilsung et al. [31] and Sungzoon Cho and Seongseob Hwang [32] proposed artificial rhythms which increase uniqueness and cues to increase consistency. Hu et al. [33] use trigraphs (three consecutively typed keys) also known as Degree of Disorder as features and normalize the distance to find the feature subset. Mariusz et al. [34] propose an approach to select the most interesting features and combine them to obtain viable indicator of user's identity. Christopher et al. [35] used three stage software design process along with data capture device hardware together with pattern recognition techniques like Bayesian and Discrimination function for classification. They used keystroke pressure and duration as features. Woojin [36] applies discrete wavelet transformation (DWT) to a user's keystroke timing vector (KTV) sample in the time domain, and then produces the corresponding keystroke wavelet coefficient vector (KWV) in the wavelet domain. In [37], a comparative study of many techniques considering the operational constraints of use for collaborative systems was made.

Table 1 summarizes the abovementioned feature extraction methods and features used. Majority of the studies have identified three fundamental attributes: duration, latency and digraph. Many statistical properties of the attributes such as mean, standard deviation and Euclidean distance are measured and are used to construct the user reference profile. Each method has its own pros and cons as the number of test subjects differs.

3.2. Feature subset selection methods

Feature subset selection [38,39] is applied to high dimensional data prior to classification. Feature subset selection is essentially an optimization problem, which involves searching the space of possible features to identify one that is optimum or near-optimal with respect to certain performance measures, since the aim is to obtain any subset that minimizes a particular measure (classification error, for instance) [40,41].

Major feature subset selection methods are detailed in this section. Yu and Cho [42,43] also proposed a Genetic Algorithm – Support Vector Machine (GA-SVM) based wrapper approach for feature subset selection in which GA is employed to implement a randomized search. SVM, an excellent novelty detector with fast learning speed, is employed as a base learner. Ki-seok Sung and Sungzoon Cho [44] proposed a one step approach similar to that of Genetic Ensemble Feature Selection (GEFS). They used SVM as base classifier for classification similar to that of Yu and Cho [43]. In particular, so-called "uniqueness" term is used in a fitness function, measuring how unique each classifier is from others in terms of the features used. To adapt SVM authors use Gaussian kernel. GA was used to filter the data and to carry out a selection of characteristics. Gabriel et al. [45,46] designed a hybrid system based on Support Vector Machines (SVM) and Stochastic Optimization Techniques. Standard Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) variation were used and produced a good result for the tasks of feature selection. Glaucya et al. [47] used weighted probability measure by selecting N features of the features vector with the minors of standard deviation, eliminating the less significant features. Very few studies have been done using evolutionary techniques and swarm intelligence for feature subset selection methods. Table 2 summarizes the abovementioned feature subset selection methods and their error rates.

3.3. Classification methods

Classification aims to find the best class that is closest to the classified pattern. A table of some sort is maintained that contains a user's details along with associated reference signature collected during the enrollment process. When those access details are entered, the system looks up the respective details and performs a similarity measure of some sort. The user's keystroke dynamics extracted during log-in session is compared and classified with the stored reference signature in the database. If they are within a prescribed tolerance limit – the user is authenticated. If not – then the system can decide whether to lock up the workstation – or take some other suitable action. The following sections categorize into seven classes the major publications in classification methods in identifying legitimate user's features. They are statistical methods, neural networks, pattern recognition techniques, hybrid techniques and other approaches.

3.3.1. Statistical methods

A reference signature is mostly calculated using the distance measure of keystroke latency and duration. The standard statistical measures like mean, standard deviation, etc. of the reference signature are used for generating template and classification. Gaines et al. [17] suggest recording of successive keystrokes and use of probability distributions of the times to distinguish subjects. Umphress and Williams [48] compared the mean, and standard deviation of keystroke latencies and digraph between reference profile and test data. Garcia [19] utilizes the covariance matrix of the vectors of reference latencies as a measure of the consistency of the individual's signature.

In [18], keystroke duration, total time to type a string, and the pressure applied to type are measured and used for classification. Joyce and Gupta [3] captured the norm building a mean reference

Table 1
Features and feature extraction methods.

Sl.No	Feature	Method	Remarks
1	Digraph latency [17]	*T*-test	*T*-tests were carried out to check if the means and standard deviation of the inter-key latency are the same
2	Keystroke time and/or pressure [18]	Mathematical model	Time periods and other characteristics are analyzed in a mathematical model to create features which make up a template
3	Keystroke interval [19]	Mean and Covariance matrix	When the individuals sought verification, they are required to type their name and a verification vector is created
4	Inter-key timings using a modified login sequence [3]	Mean and Standard deviation	Latency timing was captured during the user's login process
5	Combining key hold and inter-key times [20]	TSR program in MS-DOS	Best identification performance was achieved by using both measurements
6	Modified Keystroke Latency [12]	Mean	Keystroke latency measurement procedure was modified by counting the time duration between two successive keys pressed
7	Keystroke latency and typing difficulty [21]	Center of gravity (fuzzy logic)	Fuzzy rules were framed using the keystroke latency and typing difficulty measures
8	Keystroke latency and duration with mean userID length [22]	Programmable interrupt timer	Keystroke duration gave more accurate characterization of typing style
9	Hardened password [23]	Mean and Standard deviation	Typing patterns are combined with the user's password
10	Keystroke latency and duration [24]	Factor Analysis and k-nearest neighbor algorithm	Covariance matrix for a different user over the same set of features is used
11	Keystroke latency [25]	Box plot algorithm and normal bell curve algorithm	Intel Time Stamp Counter was used to capture the timings and Visual C++ acted as an interface
12	Trigraph duration allowing typing errors [26]	Normalization and mean	The distance between two samples is computed in the basis of the relative positions of the trigraphs
13	Digraph [27]	Average, Median and Standard deviation	The data is stored and the average, the median and the standard deviation of the times for each digraph is calculated and stored along with the statistical measures for the total time spent on each password/passphrase
14	Keystroke pressure [28]	Fast Fourier Transform	The pressure discrete time signals are transformed into the frequency domain by using Fast Fourier Transform
15	Key hold and inter key times of Long password with shift key behavior [29]	Java event handler	The feature subset aids in the classification process
16	Digraph, duration time and trigraph with amino acids [30]	Position specific scoring matrices (motif)	The frequency of each amino acid residue at each position and the number of elements within the motif was experimented
17	Artificial rhythm and cues [31,32]	Hypotheses test	Improve the quality of the data which produced improved feature subsets
18	Trigraphs [33]	Distance normalization	It refers to the elapsed time between the first key pressed and the third key pressed.
19	Keystroke latency and duration, average keystrokes per minute, overlapping of specific keys combinations, amount of errors, method of error correction [34]	Distribution function	In order to quantify the data representing time, the data's were split into multiple bins for easier perception
20	Keystroke duration and force [35]	Euclidean distance	Three feature points like amplitude, 2nd derivative and area under each peak was used as features along with duration
21	Keystroke duration and latency [36]	Discrete wavelet transform	DWT separates keystroke timing vector (KTV) into multi-resolution components so that the latent features in KTV can be well observed and extracted in keystroke wavelet coefficient vector (KWV)
22	Time between two key press	Mean vector, average and standard deviation, Euclidean distance, square of the norm of the average, Euclidean distance between units vector	In off line mode they gave an ERR value of 9.78% when using a global threshold and 4.28% when using an individual threshold. In on line mode 64% of users were correctly identified in one try and 50% during all the tests
	Time between two-key release Time between one release and one press Time between one press and one release [37]		

signature for eight sets of the users' keystroke patterns consisting of username, password, first name, and last name. Francesco et al. [26] use timing information to obtain the relative order of trigraphs. It is used to compare between two different sets of sorted trigraphs and measure the difference in the ordering between them. Gunetti and Picardi [49] present a method to compare typing samples of free text that is used to verify personal identity. K-nearest neighbor classification based authentication is proposed by Hu et al. [33] which is a simple classifier that can easily apply any distance measurement into the classification mechanism. Mariusz et al. [34] classified the verification vector to have the same class as the closest of referring vectors. A simple software based keyboard mon-

Table 2
Feature subset selection methods.

Sl.No	Method	Remarks
1	GA – SVM with Gaussian Kernal [42,43]	The degree of diversity and quality are guaranteed, and thus they gave result in an improved model performance and stability
2	GA – SVM wrapper ensemble [44]	It reports an average FAR of 15.78% with minimum FAR of 5.3% and maximum FAR of 20.38% for raw data with noise
3	GA – PSO [45,46]	Standard GA and PSO variation was used and produced a good result for the tasks of feature selection and personal identification with an FAR of 0.81% and IPR of 0.76%
4	Weighted probability measure [47]	They obtained optimum result using 90% of the features with 3.83% FRR and 0% FAR

itoring system has been suggested by Shepherd [50]. The software uses the rollover capability to correctly determine the intervals. In [51], Hidden Markov Model is used as a classifier to classify the feature subsets. Table 3 summarizes the abovementioned classification methods using statistical methods.

3.3.2. Neural networks

Rather than performing a sequential set of instructions, neural networks are capable of exploring many competing hypotheses in parallel. Because of this quality, neural networks are considered to have the greatest potential in the area of biometrics.

Angela and Sharon [52] propose a neural network system which can be located at each processor, thus parallelizing the learning and testing. In [53], perceptron algorithm has been experimented using keystroke interval as features for classification.

Bleha and Obaidat [54] use Linear Perceptron as their classifier to verify the identity of users. Half of the sample data collected was used as training data as the remaining half were used for testing. Brown and Rogers [55] suggest two separate orthogonal digraph components which added significant predictive power by Back Propagation Neural Network (BPNN). Multi layer neural network systems [56] use the time intervals between the keystroke as features. Three types of networks are studied, the feed forward BPNN, Sum of Product (SOP) trained with the modification of Back Propagation and the combination of two networks. Brown and Rogers [57] utilize Adaline and BPNN to identify the typing pattern characteristic of a particular user. Also, a simple measure of geometric distance is used for comparison. A three-layered BPNN [12] with flexible number of input nodes is used to discriminate valid users and impostors according to each individual's password keystroke pattern. Cho et al. used keystroke duration and inter-key time as their feature data in their web based java applet experiment [58]. Shereen [59] used a common 64-input node DARN network to authenticate the users through two different validation approaches. Table 4 summarizes the abovementioned classification methods using neural networks.

3.3.3. Pattern recognition techniques

Pattern recognition [60,61] is the scientific discipline whose goal is the classification of objects or patterns into a number of categories or classes. Various pattern recognition techniques used for keystroke dynamics feature selection and classification are dis-

Table 3
Statistical methods.

Sl. No.	Method	Remarks
1	Mean and standard deviation [17]	Successive keystrokes is recorded and used for authentication with the result of 4% FAR and 0% IPR for seven users
2	Mean, standard deviation and digraph [48]	The mean, standard deviation of keystroke latencies and digraph between reference profile and test data are compared. A result of 17% FAR and 30% FRR was obtained
3	Geometric distance [19]	The Mahalanobis distance function was used to determine similarity between reference and verification profiles and achieved good performance
4	Euclidean distance [18]	Euclidean distance measure between two vectors of typing of characters, total time periods and the pressure is measured and stored as template
5	Mean reference signature (mean and standard deviation) [3]	Built a mean reference signature for eight sets of the users' keystroke patterns consisting of username, password, first name, and last name and computed norm and achieved 16.7% FAR and 0.25% IPR
6	Degree of disorder [26]	Used timing information to obtain the relative order of trigraphs. It reduces the effect of variations in the absolute timing data on the authentication mechanism with 4% FAR and 0.01% IPR
7	N-graphys [49]	Method to compare typing samples of free text with less than 5% IPR and less than 0.005% FAR
8	*k*-Nearest neighbor approach [33]	Input needs only to be verified against limited user profiles within a cluster which effectively reduces the verification load significantly with the similar performance as [12]
9	Manhattan distance [34]	Manhattan distance was used to find the distance between referring keystroke feature vector and the feature vector to be classified with overall accuracy of 75.68%
10	Mean and variance [50]	Rollover capability was proposed to correctly determine the intervals and used mean and variance to determine the feature subset and for identification
11	Hidden Markov model [51]	HMM has ability of handling stochastic process and achieved an EER of 3.6%

Table 4
Neural networks.

Sl.No	Method	Remarks
1	Variation in Hebbian rule [52]	One neural network is located at each processor, thus parallelizing the learning and testing. Their tests produced a overall result of 22% FAR
2	Perceptron Algorithm [53]	Perceptron algorithm to provide linear decision functions to classify users and achieved an overall misclassification error of 2%
3	Linear Perceptron [54]	Used linear perceptron as their classifier to verify the identity of users with IPR of 8% and FAR of 9%
4	Back Propagation model [55]	Two separate orthogonal digraph components added significant predictive power. They carried out an improved preprocessing step to achieve IPR of 0% and FAR of at best 11.5%
5	Back propagation, sum-of-products and hybrid sum-of-products [56]	Used the time intervals between the keystroke as features and used for further classification with the accuracy of BPNN with 97.5%, SOP with 93.7% and Hybrid SOP with 96.2%
6	Adaline and BPNN [57]	Adaline technique gave an FAR of 17.4% in the smallest imposter group, and FAR of 18% for the larger imposter group. The partially connected BPNN gave an FAR of 12% for the smaller group, and an FAR of 14% for the larger imposter group
7	BPNN and RMSE [12]	BPNN was used to discriminate valid users and impostors according to each individual's password keystroke pattern. RMSE improved system verification performance. The resulting system gave 1.1% FAR and 0% IPR
8	Auto associative neural network [58]	Web based Java applet was used to get the keystroke timing with 1% of FAR and 0% of IPR
9	Deterministic RAM network (DARN) [59]	DARN performs quite well in distinguishing one user from several others, but the performance generally deteriorates when there are more users to authenticate

cussed. Bleha et al. [62,63] performed real time measurements of keystroke duration and used algorithms like Bayes classifier, Fisher's Linear discriminate (FLD) followed by a minimum distance classifier which provided good results. Obaidat [64] used techniques like Potential Function, Bayes decision rule, K-means algorithm, Minimum distance algorithm to classify the data.

In [22], three different classifiers Minimum Intra-class Distance Classifier (MICD), nonlinear classifier and inductive learning were applied. Descriptive statistics were generated for the mean and standard deviation. Zhang and Sun [65] modeled keystroke times as AR model according to the Wold Decomposition Theorem. Using the nearest neighborhood classifier, they classified the samples. Nick and Bojan [29] suggest shift key patterns for feature matching process and also claim that their approach offers adequate improvement when taken as an unobtrusive holistic approach merging password-based authentication with a behavioral biometric. Table 5 summarizes the abovementioned classification methods using pattern recognition techniques.

3.3.4. Hybrid techniques

Many researchers have proposed methods of combination of various neural networks, pattern recognition, statistical measures, etc. Obaidat et al. [20] describe neural network paradigms like Fuzzy ARTMAP, Radial Basis Function Network (RBFN), and Learning Vector Quantization (LVQ), Back Propagation Neural Network (BPNN), Counter Propagation Neural Network (CPNN) and classical pattern algorithms such as, Hybrid Sum-Of-Products (HSOP), Sum-Of-Products (SOP), K-Means Algorithm, Cosine Measure Algorithm, Minimum distance Algorithm, Potential function and Bayes rule algorithms for classification and gave moderate performance. In [66], a suite of techniques for password authentication using neural networks (3-layer feed-forward network implementing with the back propagation algorithm), fuzzy logic (center of gravity), statistical methods (average and standard deviation), and several hybrid combinations of these approaches are described. Pin et al. [67] propose a fusion technique by using weighted sum rule and Direction Similarity measure (DSM) with the scores that is transformed by a Gaussian probability density function. Azweeda et al. [68] propose a first attempt to realize the application of password authentication through neuro-fuzzy pressure-based typing biometric system. Table 6 summarizes the abovementioned classification methods using hybrid techniques.

3.3.5. Other approaches

Three important contributions are by Fabian et al. [23,24,69]. The idea presented in the work of Joyce and Gupta [3] is used as a foundation for their research in [69]. Clustered criterion was used to determine the feature subset and Euclidean measure, weighted and non-weighted probability measures were used for classification. Monrose et al. [23], propose an algorithm that combines the password with the keystroke latency and duration to generate a hardened password. In 2000, they [24] used [69] as the base and used k-nearest neighbor to find the feature subsets and Bayesian Classifier and other probability functions for classification. A methodology employing fuzzy logic to measure he user's typing biometrics has been suggested by William and Jan [20]. Jugurta et al. [70] have claimed that a single memory less nonlinear mapping of time intervals can significantly improve the performance of verification/identification algorithms. Kenneth [30] proposes that user's login is discretised into the amino acid alphabet based on the normalization time associated with the login details (ID/password combination). The authentication sequence is then compared with the stored motif for the login ID. It gives a score based on the algorithm specified using a simple global alignment technique. In [71],

Table 5
Pattern recognition techniques.

Sl. No	Method	Remarks
1	Bayesian, minimum distance classifier, Fisher Linear Discriminate (FLD) [62,63]	Bayes classifier gives the lowest probability of committing a classification error. FLD was used to reduce the dimensionality of the patterns
2	Potential function, Bayes decision rule, *K*-means algorithm, minimum distance algorithm [64]	Potential Function and Bayes decision rule gave FAR of 0.7% and 0.8% and FRR of 1.9% and 2.1% respectively for the combination of inter key times and key hold times. LVQ, RBFN and ART-2 gave 0% for both FAR and FRR for the combined approach
3	MICD, nonlinear classifier and inductive learning [22]	Timing vectors were collected and classification analysis is applied to discriminate between them with average FAR of 10% and IPR of 9%
4	AR model [65]	World classification accuracy using AR model as feature for the order of AR model of 30 was 41.67% and using AR model coefficients by Burg method as feature for the order of AR model of 30 was 37.96%
5	Decision tree, probabilistic, on-line linear separation, and meta learning, One R, Naive Bayes, Voted Perceptron, and Logit Boost and Breiman and Cutler's Random Forests algorithm [29]	Approaches were conducted by scripting runs to the command line interface of Weka machine learning software. Training and test sets need not be explicitly separated. A 14% FAR and 1% IPR was achieved

Table 6
Hybrid techniques.

Sl. No	Method	Remarks
1	Pattern recognition and neural network [20]	Fuzzy ARTMAP, RBFN, BPNN, CPNN and LVQ neural network paradigms were used. HSOP, SOP, Potential function and Bayes' rule algorithms gave moderate performance
2	Neural networks, Fuzzy logic, statistical methods and hybrid combinations [66]	Fuzzy classifier with lower and upper bound, BPNN, average and standard deviation and combination of these approaches are used. Calculation of FAR and IPR of these combination is discussed in detail in [44]
3	Direction similarity measure and Gaussian probability density function [67]	A weighted sum rule is applied by fusing the Gaussian scores and the DSM to enhance the final result with an EER of 9.96%
4	Adaptive neural Fuzzy inference system [68]	Combining fuzzy logic with neural network could increase the system's ability to learn the user's keystrokes patterns

Table 7
Other approaches.

Sl. No.	Method	Remarks
1	Euclidean distance, weighted and non-weighted probability [69]	Clustered profiles reduce the search time. A highest level of recognition of authentic users is 90%
2	Hardened Password [23]	Their heuristic approach effectively hides information about the user's features with 40% of false negative rates approximately
3	Euclidean distance, weighted and non-weighted probability, Bayesian Classifier [24]	The performance of the classifiers on a dataset of 63 users ranges from 83.22 to 92.14% accuracy depending on the approach being used
4	Reinforced Password using Fuzzy logic [21]	Fuzzy logic to measure the keystroke features has been suggested with 5 fuzzy rules and used the Center of gravity method
5	Time interval histogram [70]	Single memory less nonlinear mapping of time intervals can significantly improve the performance
6	Global alignment algorithm [30]	No prior knowledge is required. Efficient and can be used in on-line manner with FAR of 0.4% and IPR of 0.6%
7	Fuzzy c-Means clustering [71]	Provides additional flexibility regarding membership and removes ambiguity like whether a point belongs to the cluster or not
8	Biopassword [72]	The patent do not reveal the classification method used. But biopassword is one of the leading product in keystroke commercial market

the user profile is ciphered using DES as an extra-level of security that makes password authentication stronger and classified using fuzzy c-means clustering technique. The patent of Biopassword [72] does not reveal the classification method used. But biopassword is one of the leading products in keystroke commercial market. Table 7 summarizes the classification methods done by other approaches.

There are two factors that complicate the assessment of these classification methods. First, the reported results are based on different training sets and different tuning parameters. The number of samples used has a direct effect on the classification performance. However, this factor is often ignored in performance evaluation, which is an appropriate criteria if the goal is to evaluate the systems rather than the learning methods. The second factor is the training time and execution time. Although the training time is usually ignored by most systems, it may be important for real-time applications that require online training on different feature sets.

4. Conclusion

This paper attempts to provide a comprehensive survey of research on keystroke dynamics described in the last two decades. When appropriate, the relative performance of methods is reported. But, in so doing, it is cognizant that there is a lack of uniformity in how methods are evaluated and, so, it is imprudent to explicitly declare which methods indeed have the lowest error rates. The techniques were categorized based on the features, feature extraction methods and classification methods employed and their performance has been discussed. Though the error rates are reported for each method when available, tests are often done on different test subjects, feature set and classification methods, so, comparisons are often difficult. It has been observed that most of the work employed keystroke duration, latency or digraph as features, and the combined use of these features leads to low FAR and IPR. As a common feature, many of the systems described in this survey strive to work with very short sample texts. Hence, one may well note that it is unfair to compare the outcomes of such systems. Further research is also needed to reduce both False Alarm Rates (FAR) and Imposter Pass Rates (IPR) to levels which will be acceptable to the user.

Keystroke biometrics has an advantage over most of other biometric authentication schemes, namely, user acceptance. Since users are already accustomed to authenticating themselves through usernames and passwords, most proposed keystroke biometric methods are completely transparent. There are numerous applications which can benefit from its success, and additional studies will further validate its use as an identity verifier. This is especially relevant to the popularity of keyboards as a primary input device in data processing systems. The concept of keystroke dynamics is not limited to the traditional keyboard but any interface in which keys must be pressed can benefit from similar techniques. Keystroke biometrics has also shown great potential as the features can be collected without the need for special hardware.

References

[1] S.M. Matyas, J. Stapleton, A biometric standard for information management and security, Computers & Security 19 (n. 2) (2000) 428–441.
[2] Duane Blackburn, Chris Miles, Brad Wing, Kim Shepard, Biometrics Overview, National Science and Technology Council (NSTC) Committee on Technology Committee on Homeland and National Security, 2007.
[3] R. Joyce, G. Gupta, Identity authentication based on keystroke latencies, Communications of the ACM 33 (2) (1990) 168–176.
[4] H.M. Wood, The use of Passwords for Controlled Access to Computer Resources, NBS Special Publication, US Department of Commerce, pp. 500–509, 1977.
[5] Samir Nanavati, Michael Thieme, Raj Nanavati, Biometric's Identity Verification in a Networked World, John Wiley and Sons Inc./Wiley Computer Publication, 2003.
[6] D. Voth, Face recognition technology, IEEE Intelligent Systems 18 (3) (2003) 4–7.
[7] W. Shu, D. Zhang, Automated personal identification by Palmprint, Optical Engineering 37 (8) (1998) 2659–2662.
[8] Li Ma, Tieniu Tan, Yunhong Wang, Dexin Zhang, Personal identification based on iris texture analysis, IEEE Transactions On Pattern Analysis And Machine Intelligence 25 (12) (2003) 1519–1533.
[9] D. O' Shaughnessy, Speaker recognition, IEEE ASSP Magazine 3 (4) (1986) 4–17.
[10] C. Tappert, Adaptive on-line handwriting recognition, in: Proceedings of Seventh International Conference on Pattern Recognition, Montreal, PQ, Canada, 1984, pp. 1004–1007.
[11] N. Herbst, C. Liu, Automatic signature verification based on accelerometry, IBM Journal of Research and Development 21 (1977) 245–253.
[12] D.-T. Lin, Computer-access authentication with neural network based keystroke identity verification, in: Proceedings of the International Conference on Neural Networks, Houston, TX, USA, 1997, pp. 174–178.
[13] K. Strandburg, Daniela Stan Raicu, Privacy and Technologies of Identity: A Cross Disciplinary Conversation, Springer, 2006.
[14] A. Jain, L. Hong, Sharath Pankanti, Biometric identification systems, signal processing, ACM 83 (12) (2003) 2539–2557.
[15] Benjamin Miller, Identification news: vital signs of identity, IEEE Spectrum 31 (2) (1994) 22–30.
[16] Alen Peacock, Xian Ke, Matthew Wilkerson, Typing patterns: a key to user identification, IEEE Security and Privacy 2 (2) (2004) 40–47.
[17] R. Gaines, W. Lisowski, S. Press, N. Shpiro, Authentication by Keystroke Timing: Some preliminary results, Rand Report R-256-NSF. Rand Corporation, 1980.
[18] J.R. Young, R.W. Hammon, Method and apparatus for verifying an individual's identity, Patent No. 4,805,222, U.S. Patent and Trade, Mark Office, 1989.
[19] J. Garcia, Method and apparatus for verifying an individual's identity, Patent No. 4,805,222, U.S. Patent and Trade, Mark Office, 1989.
[20] M.S. Obaidat, Balqies Sadoun, Verification of computer users using keystroke dynamics, IEEE Transactions on Systems Man, and Cybernetics-Part B 27 (1997) 261–269.
[21] Willem G. de Ru, Jan H.P. Eloff, Enhanced password authentication through fuzzy logic, IEEE Expert: Intelligent Systems and Their Applications 12 (6) (1997) 38–45.
[22] A. John, Robinson, M. Vicky, Liang, Computer user verification using login string keystroke dynamics, IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 28 (2) (1998) 236–242.
[23] F. Monrose, M.K. Reiter, S. Wetzel, Password hardening based on keystroke dynamics, in: Proceedings of the 6th ACM Conference on Computer and Communications Security, Kent Ridge Digital Labs, Singapore, 1999, pp. 73–82, ISBN:1-58113-148-8.
[24] F. Monrose, A.D. Rubin, Keystroke: dynamics as a biometric for authentication, Future Generation Computer Systems 16 (2000) 351–359.
[25] Fadhli Wong, Mohd Hasan Wong, A.S.M. Supian, A.F. Ismail, Lai Weng Kin, Ong Cheng Soon, Enhanced user authentication through typing biometrics with artificial neural networks and k-nearest neighbor algorithm, Proceedings of the Conference on Systems and Computers 2 (2001) 911–915, ISBN: 0-7803-7147-X.
[26] B. Francesco, D. Gunetti, C. Picardi, User authentication through keystroke dynamics, ACM Transactions on Information and System Security 5 (4) (2002) 367–397.
[27] Magalhaes, Paulo Sergio Santos, Henrique Dinis dos, An improved statistical keystroke dynamics algorithm, in: Proceedings of the IADIS – Virtual Multi Conference on Computer Science, Lisboa, 2005, ISBN:972-8939-00-0.
[28] Chen Change Loy, Dr. Weng Kin Lai, Dr. Chee Peng Lim, Development of a Pressure-based Typing Biometrics User Authentication System, ASEAN Virtual Instrumentation Applications Contest Submission, 2005.
[29] B. Nick, C. Bojan, Evaluating the reliability of credential hardening through keystroke dynamics, in: Proceedings of the IEEE Computer Society International Symposium on Software Reliability Engineering, Raleigh, NC, USA, 2006, pp. 117–126, ISBN: 0-7695-2684-5.
[30] Kenneth Revett, A Bioinformatics Based Approach to Behavioral Biometrics, Frontiers in the Convergence of Bioscience and Information Technologies, Jeju City, 2007, pp. 665–670.
[31] Pilsung Kang, Sunghoon Park, Seong-seob Hwang, Hyoung-joo Lee, Sungzoon Cho, Improvement of keystroke data quality through artificial rhythms and cues, Computers and Security 27 (2008) 3–11.
[32] Sungzoon Cho, Seongseob Hwang, Artificial Rhythms and Cues for Keystroke Dynamics Based Authentication, in: Proceedings of the International Conference on Biometrics, Hong Kong, China, vol. 3832, 2006, pp. 626–632, ISBN: 3-540-31111-4.
[33] J. Hu, D. Gingrich, A. Sentosa, A k-nearest neighbor approach for user authentication through biometric keystroke dynamics, in: Proceedings of IEEE International Conference on Communications, 2008, pp. 1556–1560, ISBN: 978-1-4244-2075-9.
[34] Mariusz Rybnik, Marek Tabedzki, Khalid Saeed, A keystroke dynamics based system for user identification, in: Proceedings of the 7th Computer Information Systems and Industrial Management Applications, 2008, pp. 225–230, ISBN: 978-0-7695-3184-7.
[35] Christopher S. Leberknight, George R. Widmeyer, Michael L. Recce, An investigation into the efficacy of keystroke analysis for perimeter defense and facility access, in: Proceedings of the IEEE International Conference on Technologies for Homeland Security, 2008, 345–350, ISBN: 978-1-4244-1977-7.

[36] W. Chang, Reliable Keystroke Biometric System Based on a Small Number of Keystroke Samples, Emerging Trends in Information and Communication Security, Springer-Verlag, 2006, pp. 312–320.
[37] Romain Giot, Mohammad EI-Abed, Christopher Rosenberger, Keystroke dynamics authentication for collaborative systems, IEEE Computer Society (2009) 172–179.
[38] J. Yang, V. Honavar, Feature subset selection using a genetic algorithm, IEEE Intelligent Systems and their Applications 13 (1998) 44–49.
[39] G.H. John, R. Kohavi, K. Pfleger, Irrelevant features and the subset selection problem, in: W.W. Cohen, H. Hirsh (Eds.), Machine Learning: Proceedings of the Eleventh International Conference, Morgan Kaufmann, San Francisco, 1994, pp. 121–129.
[40] K.N. Shiv Subramaniam, S. Raj Bharath, S. Ravinder, Improved authentication mechanism using keystroke analysis, Proceedings of the International Conference on Information and Communication Technology 7 (9) (2007) 258–261.
[41] S.K. Singhi, H. Liu, Feature subset selection bias for classification learning, in: Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, 2006, pp. 849–856, ISBN:1-59593-383-2.
[42] E. Yu, S. Cho, Keystroke dynamics identity verification: its problems and practical solutions, Computers and Security 23 (2004) 428–440.
[43] E. Yu, S. Cho, GA-SVM Wrapper approach for feature subset selection in keystroke dynamics identity verification, Proceedings of the International Joint Conference on Neural Networks 3 (2003) 2253–2257, ISBN:0-7803-7898-9.
[44] K.-S. Sung, S. Cho, GA SVM Wrapper ensemble for keystroke dynamics authentication, in: Proceedings of the International Conference on Biometrics, vol. 3832, Hong Kong, China, 2006, pp. 654–660, ISBN: 978-3-540-31111-9.
[45] L.F.B.G. Azevedo Gabriel, D.C. George, E.C.B. Cavalcanti, Carvalho Filho, An approach to feature extraction for keystroke dynamics systems based on PSO and feature weighting, IEEE Congress on Evolutionary Computation (2007) 3577–3584.
[46] Gabriel L.F.B.G. Azevedo, George D.C. Cavalcanti, E.C.B. Carvalho Filho, Hybrid solution for the feature selection in personal identification problems through keystroke dynamics, in: Proceedings of the IEEE International Joint Conference on Neural Networks, vol. 12, No. 17, Unlando, FL, USA, 2007, pp. 1947–1952.
[47] Glaucya C. Boechat, Jeneffer C. Ferreira, Edson C.B. Carvalho Filho, Authentication personal, Proceedings of the International Conference on Intelligent and Advanced Systems (2007) 254–256, ISBN: 978-1-4244-1355-3.
[48] D. Umphress, G. Williams, Identity verification through keyboard characteristics, Man–Machine Studies 23 (3) (1985) 263–273.
[49] D. Gunetti, C. Picardi, Keystroke analysis of free text, ACM Transactions on Information System Security 8 (3) (2005) 312–347.
[50] S.J. Shepherd, Continuous Authentication by Analysis of Keyboard Typing Characteristics, European Convention on Security and Detection, Brighton, UK, Bradford University, 1995, pp. 111–114.
[51] Ricardo N. Rodrigues, Glauco F.G. Yared, Carlos R. do N. Costa, Joao Baptista T. Yabu-Uti, Fabio Violaro, Lee Luan Ling, Biometric access control through numerical keyboards based on keystroke dynamics, in: Proceedings of the International Conference on Biometrics, 2005, pp. 640–646, ISBN: 978-3-540-31111-9.
[52] Angela Lammers, Sharon Lagerfeld, Identity authentication based on keystroke latencies using neural networks, Computing Sciences in Colleges 6 (5) (1991) 48–51.
[53] S.A. Bleha, J. Knopp, M.S. Obaidat, Performance of the Perceptron algorithm for the classification of computer users, in: Proceedings of the ACM/SIGAPP Symposium on Applied Computing: Technological Challenges of the 1990s, Kansas City, MI, USA, 1992, pp. 863–866, ISBN:0-89791-502-X.
[54] S.A. Bleha, M.S. Obaidat, Computer users verification using the Perceptron algorithm, IEEE Transactions on Systems, Man and Cybernetics 23 (3) (1993) 900–902.
[55] Marcus Brown, Rogers Samuel J., User identification via keystroke characteristics of typed names using neural networks, Man–Machine Studies 39 (1993) 999–1014.
[56] M.S. Obaidat, D.T. Macchairolo, A multilayer neural network system for computer access security, IEEE Transactions on Systems, Man and Cybernetics 24 (5) (1994) 806–813.
[57] Marcus Brown, Samuel J. Rogers, A practical approach to user authentication, in: Proceedings of the Conference on Computer Security Applications, 1994, pp. 108–116, ISBN: 0-8186-6795-8.
[58] S. Cho, C. Han, D.-H. Han, H.-I. Kim, Web-based keystroke dynamics identity verification using neural network, Organizational Computing and Electronic Commerce 10 (2000) 295–308.
[59] S. Yong, W.K. Lai, G. Goghill, Weightless neural networks for typing biometrics authentication, Knowledge-Based Intelligent Information and Engineering Systems (2004) 284–293.
[60] J. Tou, R. Gonzalez, Pattern Recognition Principles, Addison–Wesley, 1981.
[61] R. Duda, P. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.
[62] S.A. Bleha, C. Slivinsky, B. Hussein, Computer-access security systems using keystroke dynamics, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (1990) 1217–1222.
[63] S.A. Bleha, M.S. Obaidat, Dimensionality reduction and feature extraction applications in identifying computer users, IEEE Transactions on Systems, Man and Cybernetics 21 (2) (1991) 452–456.
[64] M.S. Obaidat, A verification methodology for computer systems users, in: Proceedings of the ACM Symposium on Applied Computing, Nashville, TN, USA, 1995, pp. 258-262, ISBN:0-89791-658-1.
[65] Zhang Changshui, Sun Yanhua, AR model for keystroker verification, Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics 4 (2000) 2887–2890, ISBN: 0-7803-6583-6.
[66] Sajjad Haider, Ahmed Abbas, Abbas K. Zaidi, A multi-technique approach for user identification through keystroke dynamics, Proceedings of the IEEE International Conference on Systems, Man and Cybernetics 2 (2000) 1336–1341.
[67] Pin Shen Teh, Andrew Beng Jin Teoh, Thian Song Ong, Han Foon Neo, Statistical fusion approach on keystroke dynamics, Proceedings of the IEEE International Conference on Signal-Image Technologies and Internet-Based System (2008) 918–923, ISBN:978-0-7695-3122-9.
[68] A. Dahalan, M.J.E. Salami, W.K. Lai, A.F. Ismail, Intelligent pressure-based typing biometrics system, in: Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information & Engineering Systems, vol. 3214, Wellington, New Zealand, 2004, pp. 294–304, ISBN:978-3-540-23206-3.
[69] M. Fabian, A. Rubin, Authentication via keystroke dynamics, in: Proceedings of the 4th ACM Conference on Computer and communications security, Zurich, Switzerland, 1997, pp. 48–56, ISBN:0-89791-912-2.
[70] R. Jugurta, M. Filho, E.O. Freire, On the equalization of keystroke timing histograms, Pattern Recognition Letters 27 (13) (2006) 1440–1446.
[71] Salvador Mandujano, Rogelio 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 (2004) 181–187.
[72] Net Nanny Software International Inc., Technical Report on BioPassword Keystroke Dynamics, http://www.biopassword.com, 2001.