Keystroke Biometric

By:

Navid Bahrani, Niloufar Azmi, Majid Mafi

Submitted to Professor El Saddik

in partial fulfillment of the requirements for the course ELG 5121

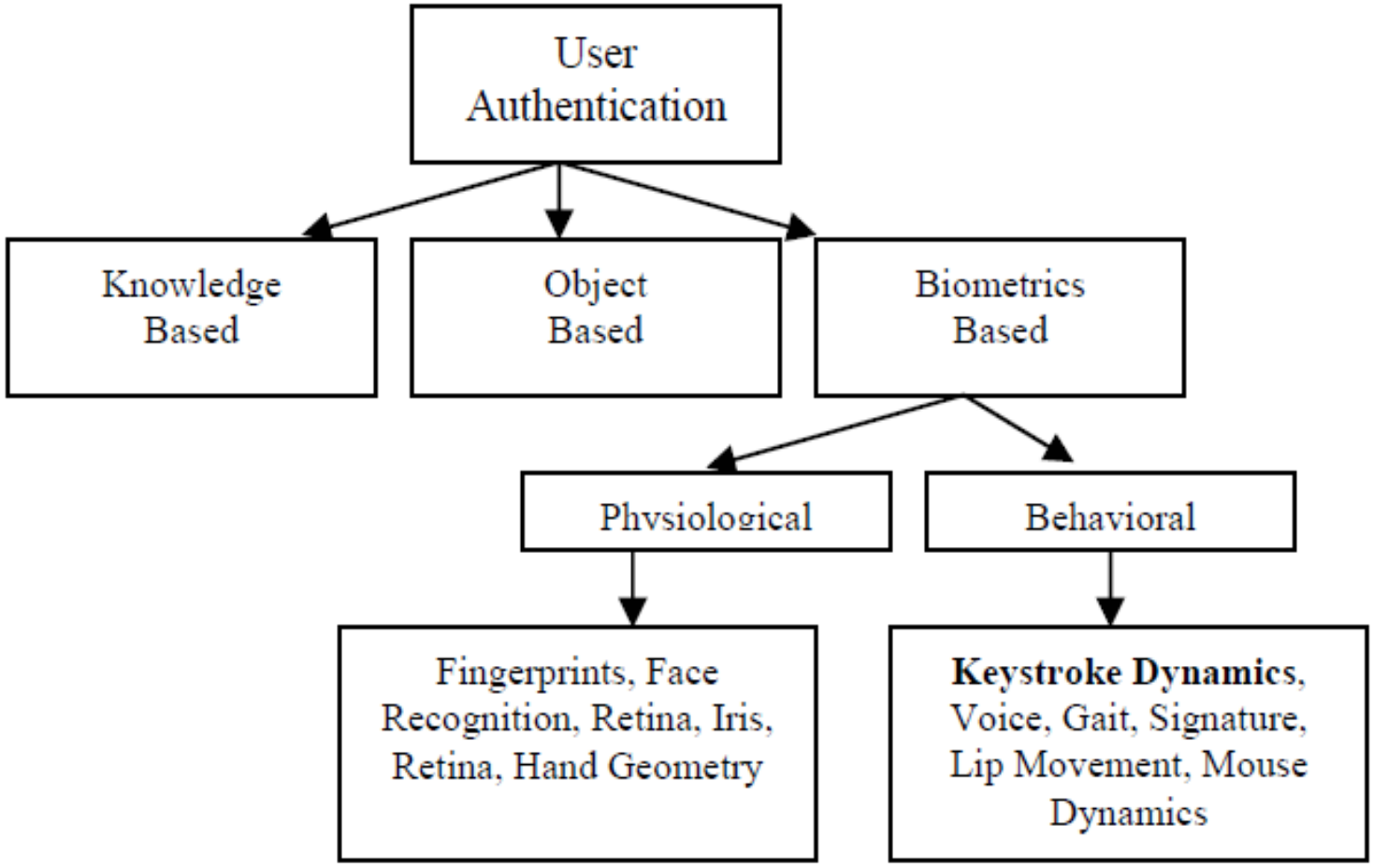
November 03, 2009

School of Information Technology and Engineering (SITE)

Outline

- Introduction
  - Overview of Biometrics
- Various approaches of research on keystroke dynamics
  - Features/Attributes
  - Feature Extraction
  - Classification methods
- Advantages of keystroke dynamics
- Conclusion
- Future Vision

User Authentication Approaches

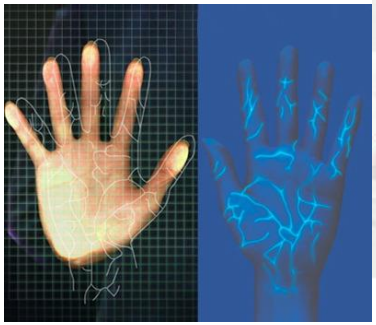


What is Biometric Authentication?

- An automatic method that identifies user or verifies the identity
  - Involves something one is or does
- Types of Biometric
  - Physiological
  - Behavioural

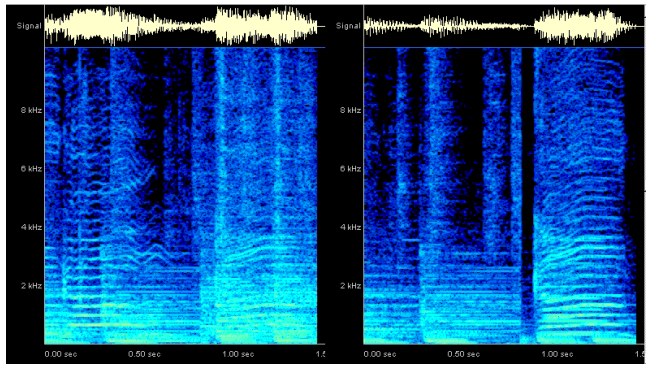
Physiological characteristics

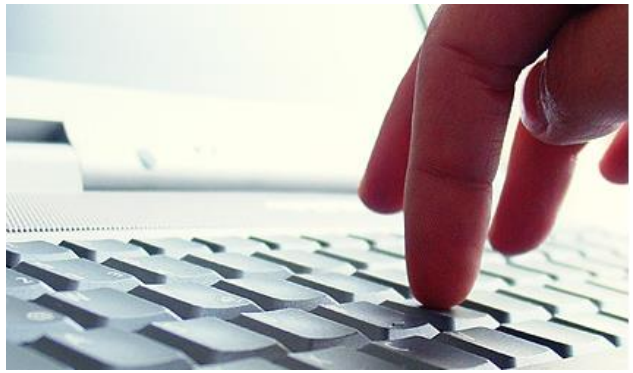
- Biological/chemical based
  - Finger prints
  - Iris, Retinal scanning
  - Hand shape geometry
  - blood vessel/vein pattern
  - Facial recognition
  - ear image
  - DNA



Behavioral characteristics

- A reflection of an individual's psychology
  - Hand written signatures
  - Voice pattern
  - Mouse movement dynamics
  - Gait (way of walking)
  - ***Keystroke dynamics***





Comparison of various biometric techniques

Table 1. Evaluation of biometric techniques [12]

Biometrics:	Universality	Uniqueness	Permanence	Collectability	Performance	Acceptability	Circumvention
Face	H	L	M	H	L	H	L
Fingerprint	M	H	H	M	H	M	H
Hand geometry	M	M	M	H	M	M	M
Keystrokes	L	L	L	M	L	M	M
Hand veins	M	M	M	M	M	M	H
Iris	H	H	H	M	H	L	H
Retinal scan	H	H	M	L	H	L	H
Signature	L	L	L	H	L	H	L
Voice	M	L	L	M	L	H	L
Facial thermogram	H	H	L	H	M	H	H
Odor	H	H	H	L	L	M	L
DNA	H	H	H	L	H	L	L
Gait	M	L	L	H	L	H	M
Ear recognition	M	M	H	M	M	H	M

Keystroke History

- Typing rhythms is an idea whose origin lies in the observation (made in 1897) that telegraph operators have distinctive patterns of keying messages over telegraph lines Behavioral biometrics
- In keeping with these early observations, British radio interceptors, during World War II, identified German radio-telegraph operators by their "fist," the personal style of tapping out a message.

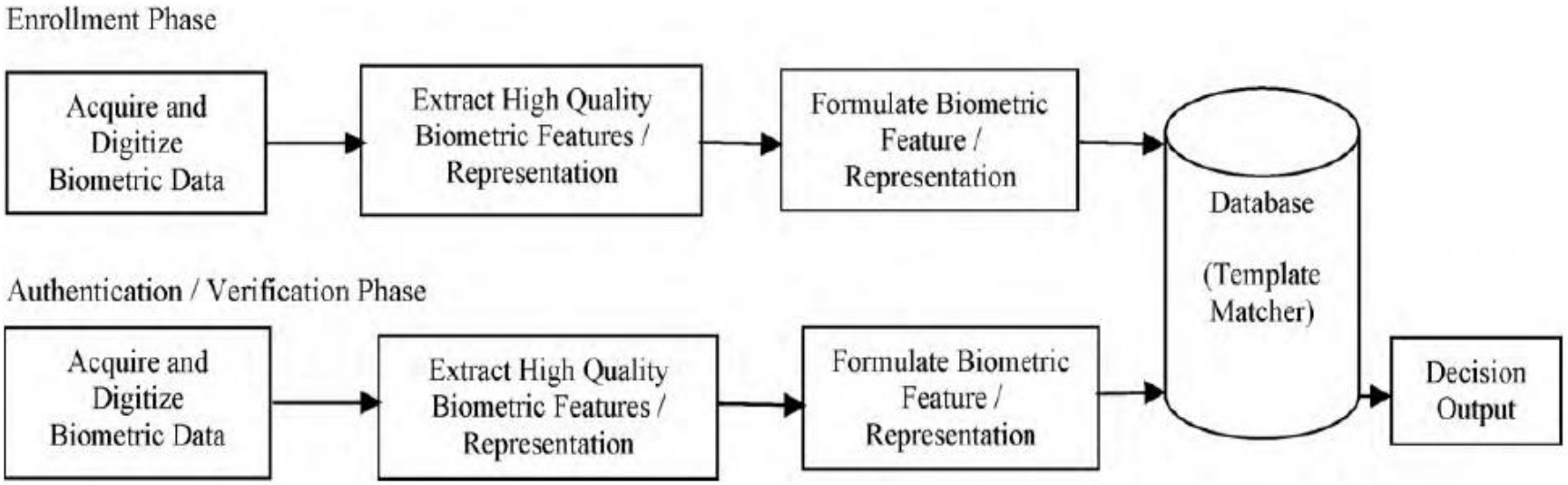
Keystroke Applications

- A Behavioral measurement aiming to identify users based on typing pattern/ rhythms or attributes
- Keystroke dynamics system different modes
- Identification mode (Find)
  - One-to-many
- Verification mode (Check)
  - One-to-one
- Non-repudiation

Keystroke Verification Techniques

- Static verification (Fixed text mode)
  - Only based on password typing rhythm
  - Authentication only at login time
- Dynamic verification (free text mode)
  - pattern regardless of the typed text
  - A continuous or periodic monitoring (On-the-fly user authentication)
  - not required to memorize a predetermined text (username & password)

Biometric System



Continuous Biometric User Authentication in online Examination (Dynamic):

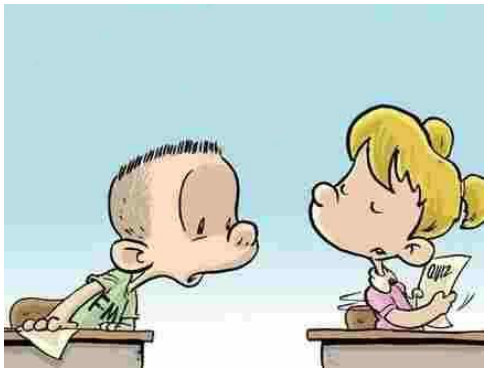
- Currently, there are 4 primary methods for user authentication:
  - Knowledge factors, or something unique that the user knows
  - Ownership factors, or something unique that the user hast
  - Something unique that the user is
  - Something unique that the user does

Secret question creation

The name of my first [Pet] was [Jaws]

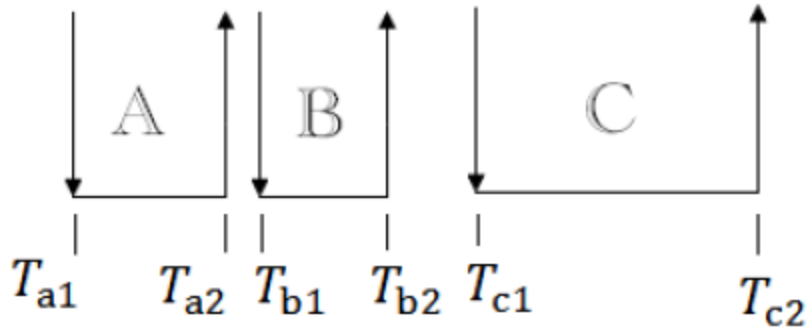
Secret question when password is forgotten

What was the name of your first [Pet] ?



Some metrics for user verification in online authentication:

- Typing speed
- Keystroke seek-time
- Flight time
- Characteristic sequences of keystrokes
- Examination of characteristic errors

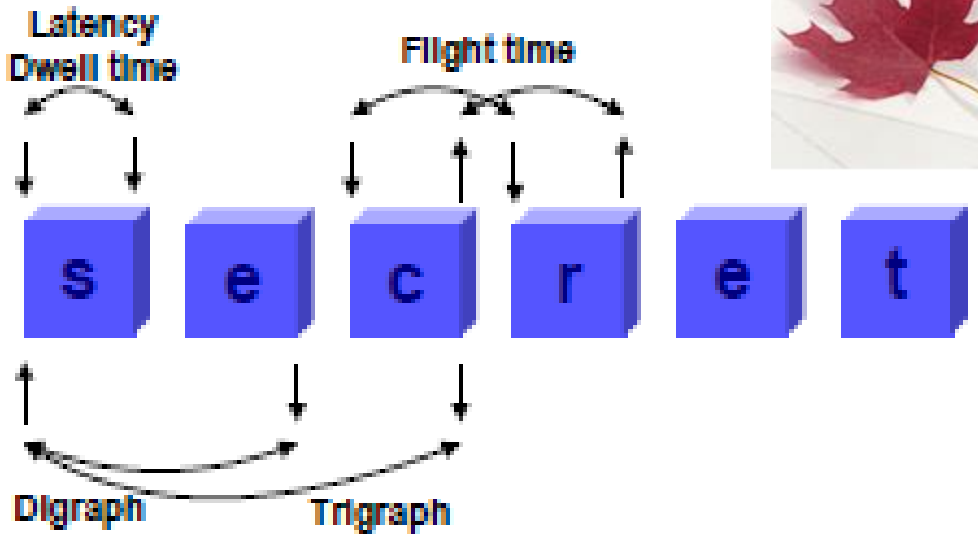


Keystrokes Dynamics (Features)

- Converts biometric data to feature vector can be used for classification
  - Keystrokes latencies (fight)
  - Duration of a specific keystroke (dwell)
  - Pressure (Force of keystrokes)
  - Typing speed
  - Frequency of error
  - Overlapping of specific keys combination
  - Method of error correction

Keystroke analysis

- Variety of methods
  - Mean typing rate
  - Inter-interval comparison
  - Digraph
  - Trigraph
  - Mean error rate
  - etc



Features & feature extraction method

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

Features & feature extraction method

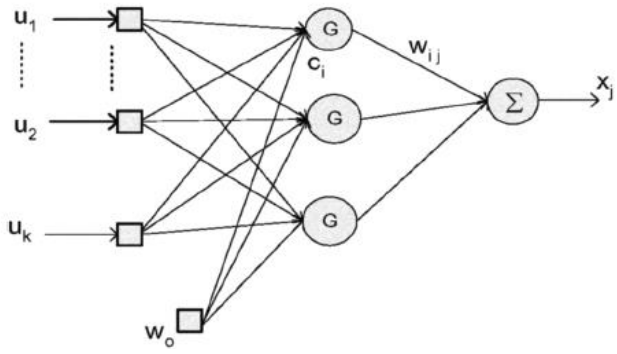
13	Digraph [27]	Average, Median and Standard deviation	The data is stored and the average, the median and the standard deviation of the times for each digraph is calculated and stored along with the statistical measures for the total time spent on each password/passphrase
14	Keystroke pressure [28]	Fast 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)

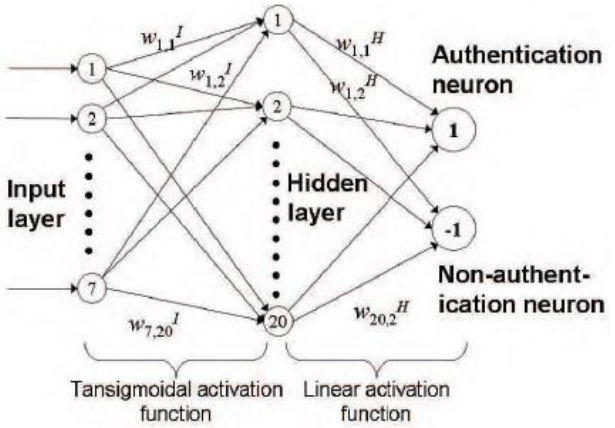
Figures of Merit

- False Rejection Rate - type I error – FRR
  - ➢ False alarm
- False Acceptance Rate - type II error – FAR
  - ➢ Missed alarm
- Equal-error rate (EER) or Crossover Error Rate (CER)
  - ➢ Different values of the operating threshold may result in different values of FRR and FAR
  - ➢ To ensure comparability across different systems

Classification methods

- Minimum distance
- Bayesian classifier
- Random forest classifier
- Neural nets
  - “combined” neural net
  - Multi-Layer Perceptron
  - RBFN
- Fuzzy (ANFIS)
- Support-vector machines
- Decision trees
- Markov models (hidden Markov model)
- Statistical Methods(mean, Std)





Classification Categories

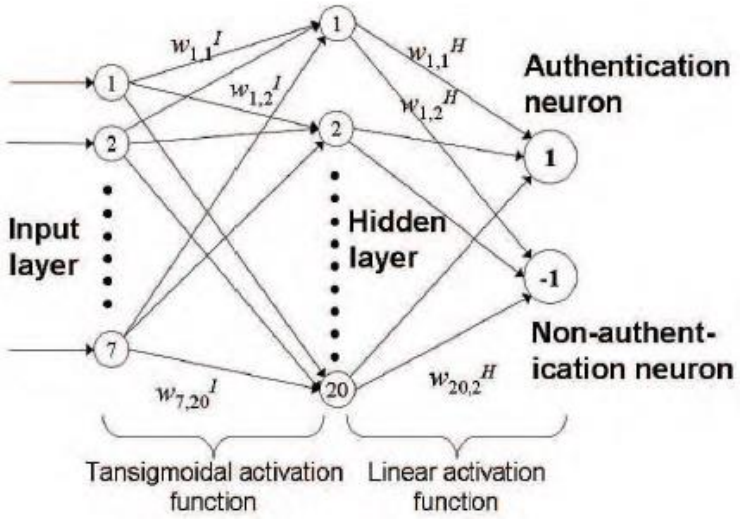
- Statistical Methods
- Neural Networks
- Pattern Recognition Techniques
- Hybrid Techniques
- Other Approaches

Statistical Methods

- Mean, standard deviation and digraph
- Geometric distance, Euclidean distance
- Degree of disorder
- k-Nearest neighbour approach
- Hidden Markov model
- N - graphys
- Manhattan distance
- Mean reference signature (mean & std)

Neural Networks

- Perceptron Algorithm
- Auto associative neural network
- Deterministic RAM network (DARN)
- Back Propagation model
- BPNN and RMSE
- Adaline and BPNN



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

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

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

Some Opportunities:

- Login information
  - Computer
  - Cell phones
  - Automated Teller Machine
  - Digital telephone dial
  - Digital electronic security keypad at a building entrance
- Continuous authentication
  - Online examination

Advantages of keystroke dynamics

- Software Only method. (No Additional Hardware except a Keyboard)
- Simple To Deploy and Use (username & passwords) – Universally accepted
- Unobtrusive, Non-Invasive, Cost Effective
- No End-User Training
- It provides a simple natural way for increased computer security
- Can be used over the internet

Keystroke drawbacks:

- User's susceptibility to fatigue
- Dynamic change in typing patterns
- Injury, skill of the user
- Change of keyboard hardware.

Keystroke Challenges

- Lack of a shared set of standards for data collection, benchmarking, measurement
- Which methods have lower error rate?
- Error rate comparison is difficult
- Work with very short sample texts
- There is no identical biometric samples
- Requires adaptive learning

Conclusions

- It seems promising , still needs more efforts specially for identification
  - Iris scanners provide the lowest total error rate - on the order of 10^{-6} in many cases
  - Even fingerprints provide an error rate on the order of 10^{-2}
- Extreme different typing patterns among examinees

Conclusions

- Several commercial systems on offer:
  - BioPassword (now AdmitOne), PSYLock, Trustable Passwords
  - but no evaluation data are publicly available for these systems
- Combined features of maximum pressure with latency → effective way to verify authorized user
- Combined ANN & ANFIS → greater promising result

Future work

- Using longer fixed texts
- Test on extensive database
- Combining many features
  - ➢increase the accuracy of keystroke analysis
- Find the most efficient features
- Adding mouse dynamic
  - ➢Helpful for identification
- Special characters & character overlapping
- Typing pattern as ***Digital Signature***

Future work

- Researchers focus rather on user verification, there is a little works on users identification
  - ➢Maybe an obstacle is gathering big database
- Also trends in classifiers shows that many people uses ANN
  - ➢work on black-box basis
  - ➢adding new user to the database
- Future research to reduce FAR & FRR

Comparison of Classifiers

- The random forest classifier is
  - robust against noise
  - its tree- classification rules enable it to find informative signatures in small subsets of the data (i.e., automatic feature selection)
- In contrast, SVMs
  - do not perform variable selection,
  - can perform poorly when the classes are distributed in a large number of different but simple ways.

Methods to measure the users typing biometric:

- Fuzzy logic:
  - There are many adjustable elements such as membership functions and fuzzy rules
  - *Advantage:*
    - many adjustable elements increase the flexibility of the fuzzy based authentication
  - *Disadvantage:*
    - increase the complexity in designing fuzzy-based authentication system.

A: Methods to measure the users typing biometric

- RBFN:(Radial basis function network)
  - Alternative neural network architecture
  - *Major advantage:* can be trained to allow fast convergence to solitary global minimum for a given set of fixed hidden node parameter.