**Research Summary**

**Biometric authentication and identification using keystroke dynamics: A Survey**

Data Acquisition:

1. Text entry -> Static or structured text and dynamic or free text
   * Static analysis: analysing keystroke behaviour of an individual on predetermined phrase at certain point in system
   * Dynamic analysis: continuous or periodic monitoring of keystroke behaviour
2. Environment -> Type of keyboard is important -> Rhythm or keystroke dynamics

* Can be Controlled or Uncontrolled

1. Features:
   * Key timings -> key pressed time and key release time
   * Latency: PP (press to press), RR (release to release) and RP (release to press also known as flight time)
   * Trigraph – time interval between the presses and releases of alternate keystrokes
   * N-graphs – ca be used for similar purpose as trigraph. The trigraph gave better classification results than digraphs or n-graphs

Key hold time or dwell time is the time for which each keystroke was pressed.

Keystroke latency is the combination of the hold and flight times. Hold time is more important than inter-key time. Total duration, the time it takes a user to type a certain string can also be measured. Speed is measured in words per minute (wpm).

1. Error metrics:
   * Matching algorithm used
   * Identification is the one-to-many process of comparing a submitted biometric sample against all of the biometric reference templates.
   * Error rates
     1. False Acceptance Rate – FAR – the percentage of the impostors inaccurately allowed as genuine users
     2. False Rejection Rate – FRR – number of genuine users rejected from using the system
     3. EER – Equal Error Rate – is defined as FAR/FRR at an operation point on ROC where FAR = FRR
2. Databases – soft biometric classification contain soft information such as age, gender and ethnicity.

**Approaches**

Classification of users is made based on similarities and dissimilarities among the templates

1. Statistical Algorithms: Simplest method – computing the mean and standard deviation of the features in the template.

Disadvantage - The lack of training stage which can be useful to identify patterns in the keystroke data.

1. Neural Networks - NN – Artificial Neural Networks
   1. Supervised learning or Backpropagation
   2. Unsupervised learning – Hopfield neural networks

Obaidat and Macchiarolo investigation to classify intercharacter times using artificial neural networks.

Tested architectures: Backpropagation, Sum of products and hybrid sum of products. Sum of products was found to perform better than other architecture -> Identification rate being 97.8%.

Yong et al suggested the use of weightless neural networks for classifying users. Data scaled into linear and non-linear intervals.

Adv of NN: can handle many parameters

Diss of NN:

* can be slow in the training and in the application
* Difficult to decide which features are important for classification due to it’s “black box”.
* A problem for continuous keystroke authentication where results are typically desired in real time

1. Pattern Recognition and learning based algorithms – the use of patterns which help to classify them into different categories based on certain algorithms.

Contains simple machine learning algorithms:

* Yu & Cho: used a three step approach to improve the performance of keystroke dynamics
* The SVN novelty detector achieved an average error rate of 0.81%
* Giot et al. proposed a method to identify computer users by using a Support Vector Machine(SVM) and achieved an identification rate of 95%. SVM is a supervised learning algorithm which shows promising results for both authentication and identification. It is an important algorithm which should be used as benchmark for future algorithms.
* The disadvantage of using probabilistic learning algorithms is that they provide a confidence value with the decision made and reduce the problem of error propagation by ignoring the outputs with low confidence values. Unsupervised learning techniques can identify patterns in the data automatically.

1. Search heuristics and combination of algorithms.

The genetic algorithms are used to find an optimum solution, they are part of evolutionary algorithms. They are used in Ant Colony Optimisation (ACO).

- SVM uses a Genetic Algorithm (GA) as an evolutionary algorithm. Min error rate of 5.18 %, FAR = 0.43% and FRR = 4.75%

- Using PSO (Particle swarm optimisation) with personal ang global acceleration of .5. Min error rate of 2.21%, FAR = 0.41 and FRR = 2.07

- Revett at al. used a bioinformatics approach to achieve FAR – 0.1 and FRR = 0.1. The algorithm can handle 40000 sample and provides promising results.

Adv of GA: can handle large databases

* Provides multiple solutions
* Can handle multi-dimensional, non-dimensional, non-continuous and even non-parametrical problems.

It was observed that histogram equalization of keystroke timing data led to an improvement in equal error rate (EER).

1. Factors affecting performance

* Joice and Gupta observed that shorter and easy to type login strings were easier to impersonate. They empathise the importance of timing accuracy of the machine
* Effect of clock resolution on keystroke dynamics
* EER increased by approximately 4.2% when using 15ms resolution clock instead of 1ms
* European standards for access control mandates a FAR of 1% and a miss rate of 0.001%
* System needs to be made robust to timing inaccuracies
* When users are unsure of what to type leads to distinctive pause during typing
* Structured vs unstructured text
* The use of user dominant hand can be used for identification
* Normal English words seemed less discernible from each other than those passwords which contain special characters.
* Godness measure – used to compute the quality of an word used in keystroke dynamics. Provides a reference of good non English words that can be used for testing purposes
* Non English words were identified with higher accuracy than English words
* Having a mix of alphanumeric characters in the password increased the chances of being identified with the valid user
* Longer passwords provides better means to identify the user accurately
* Size, type of password and number of samples needed for a person to enrol and authenticate has an impact on the error rate of the system.
* Peacock et al. – provided a graph for the cost to enrol and the cost to authenticate
* Chang – presented a way to increase the size of the training dataset using resampling techniques in time and wavelet domains. Computational costs were reduced from 2 hours to 3 seconds.
* The emotional state of the user has an impact on the typing speed. A negative state led to 70% reduction in speed and a positive state led to 80& increase.
* Health conditions, place where the person types and the type and brand of computer used can also have an impact on the tying speed

**Behavioural biometrics: a survey and classification**

1. Introduction

One of the defining characteristics of behavioural biometric is the incorporation of behavioural signature.

The behaviour has a beginning, duration and end.

Behavioural biometrics categories:

1. Authorship based biometrics – based on examining a piece of text or drawing produced by a person. The verification is accomplished by observing style peculiarities typical to the author.
2. HCI – users employ different strategies, use different styles and apply unique abilities and knowledge.

- HCI based biometrics can be subdivided in subcategories: HI with input devices and HCI based behavioural biometrics which measure advanced human behaviour such as strategy, knowledge or skill.

1. Indirect HCI-based biometrics. Evets that can be obtained by monitoring user HCI behaviour indirectly via observable low-level actions of computer software.
2. Motor-skills of users - An ability of users to utilise muscles. Muscle movements rely upon proper functioning of the brain, skeleton, joints and nervous systems and so on. Motor skills indirectly reflect the quality of functioning of such systems, making person verification possible. Motor skills are learned not inherited. Authors adapts definition for motor-skills based behavioural biometrics a.k.a. kinetics
3. Purely behavioural biometrics – measure human behaviour
4. Behavioural biometrics

Keystroke dynamics – typing patterns are characteristic to each person

* Some people are experienced typists utilising the touch-typing method while others utilise hunt-and-peck approach (only 2 fingers). Those differences makes verification possible based on users typing patterns. Some reports suggest that identification might also be possible
* For verification – input of user’s password might be sufficient but recognition needs a large amount of data and identification is based on comparisons with the profiles of all other existing users in the system
* Keystroke features are based on time durations between keystrokes, inter-key strokes and dwell time
* Keystrokes dynamics is probably the most researched type of HCI-based biometric.

1. Generalised Algorithm:

* Pick behaviour
* Break up behaviour into component actions
* Determine frequencies of component actions for each user
* Combine results into a feature vector profile
* Apply similarity measures function to the stored template and current behaviour
* Experimentally determine a threshold value
* Verify or reject user based on the similarity score comparison to the threshold value.