RESEARCH ARTICLE OPEN ACCESS

Keystroke Biometric For User Authentication - A Review

Pallavi Yevale, Mehul Jha, Swapnil Auti, Laukik Upadhye, Siddhrath Kulkarni.
(Department of Computer Engg, Alard College of Engineering, Pune University, India)

ABSTRACT
Conventionally, users are authenticated under the normal user name and password procedure. The latest trend in authentication is using Biometric as a feature for identifying users. By using biometrics as the integral part of authentication, the chances of imposters entering in a secured system becomes very low. Keystroke Biometric uses the behavioral typing pattern as a biometric for identifying a user.

In this paper, a survey is carried out on three different approaches that implements keystroke biometric system. 1) Clustering Di-Graph (CDG): This method uses clustering digraphs based on temporal features. This method joins consecutive keystrokes for representing keystroke pattern. 2) Hamming Distance-like Filtering (HDF): In this method, the dissimilarity has their EER(Equal Error Rate) depending upon filtering of predefined value of gathered data. It is based on resemblance to Hamming Distance. 3) Free Text and Euclidean Distance based Approach (FTED): In this method, keys are classified into two halves (left - right) and four lines (total eight groups) and then timing vectors (of flight time) are obtained between these key groups. For distinguishing the legitimate user from imposters, Timing Vectors are used.

***Keywords* -** Biometrics, FAR, FRR, EER, Digraph

I. INTRODUCTION

Keystroke Biometrics for user authentication uses the behavioral typing pattern of a user and authenticates him. As we all know Computers have become an ubiquitous part of the modern society. Everybody uses computers in one or other way. Computer is often used together with the space force we are all familiar with that is the internet. Everybody uses internet for personal reason or professional reason or may be for both the purposes. People today have become over dependent on internet. With the increase in the use of internet, there comes a factor that is of utmost importance and probably the riskiest term in the field of internet that is Security. It can be compromised under certain circumstances. For example in the early 2011, there was an online attack that happened on multiple companies which resulted in total network shutdown and all the important information and passwords of the employees were lost. This is the scale of attacks that can happen if there is not a good security present. With too much dependency on internet there is a need of protecting users and their information from the intruders. We all need a simple, low cost yet unobtrusive method for security purpose. All this led to the Keystroke Biometrics for user authentication coming into the picture. Biometrics basically is the science of measuring the unique physical characteristics of an individual. Biometrics is a property that cannot be shared with other individual. It is a property that cannot be lost. There are two types of Biometric properties:
1. Physiological properties (includes retina, hand, palm, etc) 2. Behavioral properties (includes typing pattern, speech, etc).

Keystroke Biometrics for user authentication uses typing pattern of an individual for authentication. It records the timing difference between the two keys pressed and when the same user logins later, the access is granted only if the timing matches the recorded data.

II. RELATED WORK

The use of passwords as the sole means of regulating access is a weak method of authentication. People choose passwords that are easy to remember, which generally means that they choose words or names that are familiar. This practice restricts the range of passwords to a fraction of what is possible, and by choosing passwords that might be found in a dictionary they become much more vulnerable to the most common techniques of computer hacking.

2.1. Clustering Di-Graph (CDG)

Tomer Shimshon [1] suggests that even after the user is authenticated, the logged station is vulnerable to imposters. So as to prevent that, they propose a method to continuously verify a user. This technique focuses on reducing the dimensionality of the features vector that describes a particular session. Later from each input stream of keystrokes, a feature vector is extracted. These feature vectors are used to create a model that represents the typing pattern of a user and the same is used for verification. This method suggests a technique that reduces the FAR (False Acceptance Rate) and FRR (False Rejection Rate) by using clustering di-graphs.

2.2. Hamming Distance-like Filtering(HDF)

Yoshihiro kaneko [2], proposed "Finding dissimilarity of two digraphs and obtained EER". Digraph is the term used to measure the dissimilarity between two consecutive keystrokes. Obtained EER from dissimilarity in users' digraph can be used identify weather he is legitimate or not. Measurement of difference can be close to zero if data is taken from same user. The EER can be obtained by generating Tune Parameter for multiple times.

This technique can combine with other statistical models to improve the result by reducing the EER.

2.3. Free Text and Euclidean Distance based Approach(FTED)

Saurabh Singh and Dr. K.V. Arya [ ] work is based on free text system in which user is supposed to type a string of his/her choice. The keys on keyboard were divided into 8 groups based on their location i.e. left side/right side and row number. Calculating the difference between the Flight Time (FT) obtained from the database and that from the user during trial using Euclidian Distance(ED) formula, the authentication is either granted or denied accordingly.

All the above papers that we surveyed aims at enhancing the security by further reducing FAR and FRR

III. STATISTICAL BASED KEYSTROKE BIOMETRIC ALGORITHMS

3.1. Clustering Di-Graph:

- Features Reduction-

This method consists of two phases: Training phase and Verification phase.

In the training phase, verification model is built that consists of a multi class classifier(C) and mapping function (M) for the user u based on all users' sessions(S). A vocabulary (Vu) that consists of n-graphs from a user's training sessions (Su). Then mean of the temporal feature for each n-graph in the vocabulary based on all its instances in the user training sessions (Su). Later a clustering technique is applied that clusters the means of the temporal features into k clusters. The result of the clustering is a mapping function from an n-graph in the original user vocabulary to a cluster (M<v,c>).Then the transformation of all the user's sessions(S) to features vectors(FVs) is done by first extracting for each user the means of their temporal features and then mapping them to a cluster corresponding to the mapping function. Then a multi-class classifier(C) based on those FVs.

In the verification phase, a session that is to be verified(St) is transformed to a features vector(FV) based on the mapping function that was created during the training phase and verify it based on the classifier(C). The process of creating the verification model(C and M) is performed for each user separately. Thus for each user, clustering of n-graphs is done in a different way, that leads to a different classifier. The motive behind this technique is that similar n-graphs can be considered as the same feature. Algorithm 1 and 2 present the pseudo code of the two algorithms: BuildUserModel that is called during the training phase and VerifySession is called during the verification[1].

```
Algorithm 1 - BuildUserModel(S,Su,k)
    Input: S is the users' sessions
           Su is the user's sessions (Su ε S)
           k is number of clusters
    Output: C - A multi-class classifier
  M<v, c> is mapping function from a n-graph to a cluster
1. Vu = CreateUserVocabulary(Su)
2. F= for each v ε Vu calculate the mean temporal feature of v
3. M<v, c>= Cluster (F,k)
4. FVs = TransformUserSessions(M,S)
5. C=Train (FV)
```

```
Algorithm 2 [1] - VerifySession(S,C,M,parameter)
    Input: St-The test session
       C is the user multi-class classifier
       M<v,c> - The user mapping function
       Parameter - the verification parameter
    Output: It Accepts or rejects the test session
  1. FV = TransformSession(M,St)
  2. Verify(C,FV,parameter)
```

- Clustering Technique

First, the n-graphs are sorted based on their temporal features. Then k similar n-graphs are grouped together into one cluster. The cluster temporal feature will be the average of the temporal feature will be the average of the temporal features of all the grouped n-graphs that it contains.

- User Verification by Classification

A model based on the users session has to be learned, that is later used to verify each session's features vector. For classification, binary classifiers are used. For continuous verification, samples of verified user are taken and since the alternative class is not clear because it may contain all imposters, multi class classification is used. In this technique, a user is classified after the classifier was trained on n users, and the verified user is one of them. For testing the method, data collection that was recorded in [6] was used. This dataset contains ten legitimate users who typed fifteen emails each. Also this dataset also contains 15 sessions which were typed by other users who represented attackers. The terms that were considered for evaluation were FAR, FRR, ER(Error Rate) that is the average of the FAR and the FRR measures. Error Curve was also used that represents the system measurements for various threshold and finally the concept of Area Under the Curve (AUC) was used to compare different number of clusters.

This technique introduces a new features reduction method that is used for each user separately based on the similar digraphs.

3.2. Hamming Distance-like Filtering

- Dissimilarity Measure

For dissimilarity measure it take the DOWN-DOWN time between the first key pressed and the next key pressed. Let *J* denote such fixed digraph set and let t_j denote the mean of the DOWN-DOWN time of $j \in J$ in a test datum. Let m_j and s_j denote the mean and standard deviation, respectively, of the DOWN-DOWN time of $j \in J$ in a template [2].

Formula for measuring dissimilarity is given by [2].

$\alpha(m_j - t_j) / (m_j(\alpha - 1))$ if $\alpha\text{-}1m_j \leqq t_j \leqq m_j$
$(t_j - m_j) / (m_j(\alpha - 1))$ if $m_j < t_j \leqq \alpha m_j$
1 otherwise

- EER Evaluation

EER evaluates dissimilarity measures as a user authentication method. Fig.1 shows how to obtain EER. In data collection, each subject is required to input a predefined text totally five times. For such collected data, data filtering is adopted [1]. Next, as the template of each subject, select one datum whose dissimilarity is lowest, the minimum sum among the same subjects. Then calculate the mean μ and standard deviation σ of dissimilarity measurements of its selected template and the rest four of the same subjects. Moreover, measure dissimilarity among each template and any other datum. With a tuning parameter *a*, set a threshold $\mu + a\sigma$ and authenticate them as follows. First, by adopting our proposal Hamming distance-like filtering, verify that some data are collected from different subjects. Next, among the same subjects again, if a measurement exceeds $\mu + a\sigma$, then the numerator of FRR, is counted. Moreover, among different subjects, if a measurement does not exceed $\mu + a\sigma$, then the numerator of FAR is counted. After the above authentication, tune *a*, reset a threshold, and re-authenticate them. Repeat this until the EER is obtained which is the cross-point of FRR and FAR.[2]

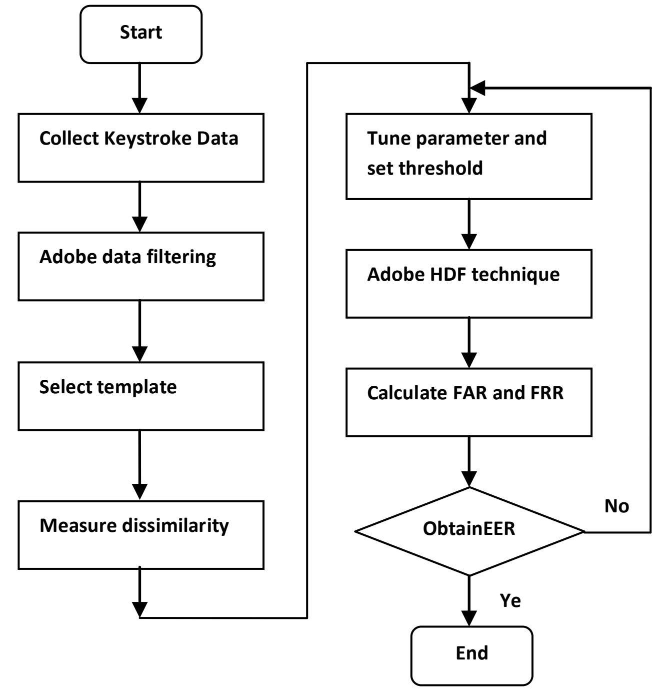

Figure 1: EER Evolution

3.3. Free Text and Euclidean Distance based Approach

- Flight Time

Saurabh Singh and Dr. K.V. Arya [3] used Flight Time (FT) as the key descriptor as the standard deviation provided by FT was quite significant. FT is the latency between release event of the previous key and press event of the current key.

- Key Grouping

It was very difficult to match the text entered by the user at login time with that stored in database as it should contain sufficient size of matched pattern vectors. Consider a pattern stored in database of user X-

ef	wm	nu	wk	ee	an	eo
25	23	35	32	23	21	21

at login time user enters- “we shall not k**ee**p **an**ything pending.” Here, the two matched sequences are “**ee**” and “**an**”, which are not sufficient for analysis purpose. To get sufficient sequences user has to enter a very long string, and system also has to maintain large number of sequences to increase the probability of matching which in turn would be computationally expansive [3]. The authors have therefore classified the key board into 8 parts, first in two halves i.e. keys in left hand and keys in right hand(L & R) and then in 4 lines starting from numbers row to last row (1,2,3,4) therefore the 8 parts are L1,L2,L3,L4,R1,R2,R3,R4.

With this approach the database looked like this-

L2L2	L2R4	R4R2	L2R3	L2l2	L3L4	L2R2
25	23	35	32	23	21	21

And the string gave the group sequence-
L2L2SPCL2R2R3R3SPCR4R2L2SPCR3L2L2R2SPC L 3R4R2SPCL2R2L2R3SPCR2L2R4L3R2R4L3
Now the matching sequences are
L2L2 R4R2 L2R2 L2R3 L2R4
Thus 5 matching sequences are found as compared to previous one (2 sequences). This information is sufficient for analysis. This justified the key grouping followed.

- Euclidian Distance

Once the FT in the database and that entered in the trial by the user are obtained, the Euclidian distance between the two vectors are calculated [3] as -

$$d(p,q)\ \sqrt{(p1-q1)^2+(p2-q2)^2+\cdots.(pn-qn)^2}=\sqrt{\sum_{i=1}^{n}(pi-qi)^2} \quad .....(1)$$

The distance d(p,q) is used to decide the authenticity of the user trying to login. As the human's behavioural characteristics are not consistent every time, distance d(p,q) may not be 0. Therefore, some marginal value α was designed such that if d ≤ α then the user is classified as legitimate user, and if d > α but less than or equal to another marginal value β, than the user is classified as suspect and he/she is asked to type another text to be verified, but if d > β, the user is classified as imposter.

- Profile Enhancement

When a user is classified as a legitimate user, his/her profile pattern vector is enhanced by adding new Key Group Pairs(KGP) which were not present in the database and were obtained from text entered by the user. So profile pattern vector is continuously improving the performance of the system.

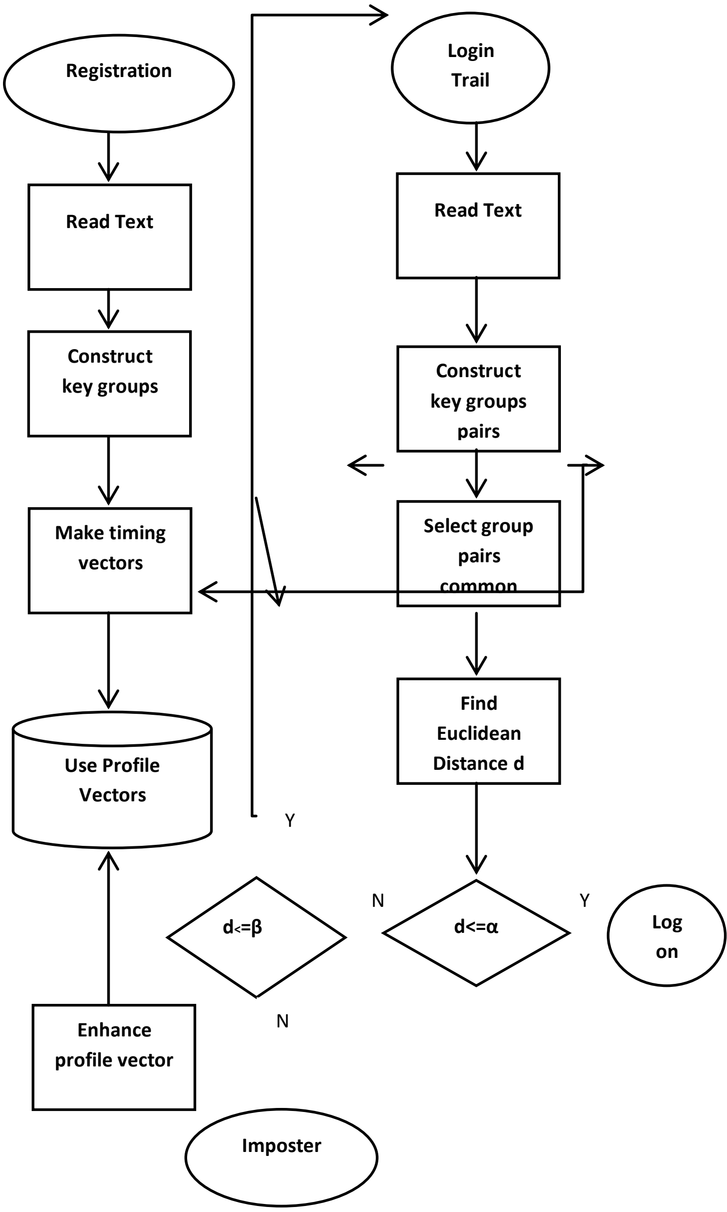

Figure 2: Flow Chart of the FTED System

IV. COMPARISONS OF STATISTICAL BASED KEYSTROKE ALGORITHMS

A comparison of the Statistical based algorithm is shown in Table 1

Papers	**CDG**	**FTED**	**HDF**
Heuristic	Tomer Shimshom et al.'s heuristic	Saurabh Singh et al.'s euristic	Yoshihiro Kaneko heuristic
Type of System	Free text	Free text	Structured text
Evaluation measure	• FAR • FRR • ER • AUC • DG	• Flight Time • Euclidian distance • FAR • FRR	• Hamming Distance • Dissimilar- ity measures • EER
Feature	digraph as two consecutive keystroke and uses clustering technique for evaluation. Also this method uses AUC as an enhanced feature for calculation. This technique continuously verifies a user.	Alpha-numeric keys are divided into 8 groups and Euclidean distance is found between trail and database vectors.	filtering is applied on collected digraph data. further on this filtered data is combine with tune parameter to lower the EER
Advantages	• This method can be used with any classification algorithms. • Less FAR and FRR as compared to other methods.	• Grouping of keys allows more patterns to be matched. • Profile enhancement mechanism enhances authentication process after every successful login.	• Data is filtered before used for analysis. • Tuning parameter is used to generate EER
Disadvantages	• This system continuously keeps user authenticating even after he is logged on • Lots of calculation is required to get the result	Comparatively more memory per account is required for profile enhancement	• After filtering data some digraph details may disappear • This technique reduces the FAR but not the FRR
Result	FAR is 0.41% and FRR is 0.63% with this method	FAR is 2.0 and FRR is 4.0 with this method	EER is 0.99 with this method

Table 1: Comparison of statistical based algorithms

V. CONCLUSION

CDG based algorithm shows the best performance of all three algorithms, with respect to FAR and FRR. Also continuous authentication is done even after user is logged in. FDED based algorithm is more suitable for users authenticating frequently as it enhances authentication process after every login for that particular account. HDF based algorithm filters the timing difference between two characters typed before analyzing it. HDF based algorithm can be easily integrated with other statistical models and thus is portable.

REFERENCES

[1] Saurabh Singh, Dr. K.V.Arya "*Key Classification: A New Approach in Free Text Keystroke Authentication System*" Circuits, Communication and System (PACCS),2011.

[2] Yoshihiro kaneko, Yuji Kinpara, Yuta Shiomi, "*A Hamming Distance-like Filtering in Keystroke Dynamic*", 2011 Ninth Annual International Conference On Privacy ,Security And Trust

[3] D.Gunetti and c.picardi , "*keystroke analysis of free text,*"ACM Trans. On information and system security,4,449-452,2006.

[4] H.Davoudi and E.Kabir , "A new distance measure for free text keystroke authentication",Proc . of the 14th international computer conference,570-575,2009.

[5] Tomer Shimshon, Robert Moskovitch, Lior Rokach, Yuval Elovici, "*Clustering Di-Graphs for Continuously Verifying*", 2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel

[6] K. Hempstalk. Continuous Typist Verification using Machine Learning, PhD thesis, University of Waikato, 2009.

[7] Jae-Wooklee,Sung soon choi,Byung-ro Moon, An Evolutionary Keystroke Authentication Based on Ellipsoidal Hypothesis Space, *GECCO'07,* July 7–11, 2007, London, England, United Kingdom.

[8] Rick Joyce and Gopal Gupta. Identity authentication based on keystroke latencies. Communications of the ACM, 1990.

[9] Gunetti and Picardi. Keystroke analysis of free text. In ACM Transactions on Information and System Security, volume 8, pages 312–347, 2005.