An enhanced distance metric for keystroke dynamics classification

Ha Nguyen Ngoc
dept. Mechatronics
Institute of vehicle engineering.
Ha noi,Viet Nam
Email: hayeuot@yahoo.com

Ngoc Tran Nguyen
DEPARTMENT OF INFORMATION SECURITY
Le Quy Don Technical University
Ha noi,Viet Nam
Email: tonono79@yahoo.com

***Abstract*—In this paper, the relationship between the distance metrics and the data model was analyzed and a new algorithm for keystroke dynamics classification was proposed. The results of the experiments on the CMU and GREYC keystroke dynamics benchmark and mobile devices datasets were evaluated. The classifiers using the proposed algorithm outperform existing top performing keystroke dynamics classifiers which use traditional approaches.**

***Index Terms*—Bio-password, Identification, Information security, Authentication.**

I. Introduction

Today, people more and more store sensitive data on their mobile devices. So, there is a strong authentication mechanism is a very important thing. Currently, the analysis of the sample type of users often use Keystroke Dynamics method (KD), this method is very useful to enhance the security of the password-based authentication. Furthermore, the touchscreen allows adding different features, from pressure on the screen or area of contact fingers on the characteristics based on time to use for the model of keyboard dynamics. Had the study verified the effectiveness of the addition of touch screen feature to identify and authenticate through the user's data. The results showed that this additional attributes enhance the accuracy of both processes. In the content of this study, we present a new measurement experiment on three data sets (CMU [4], GREYC [15] and group Margit Antall [1]), the results also showed that this new distance metric to further enhance the accuracy of the verification process.

The next section presents a summary of the research area KD, consider a few studies have been done on mobile devices that use the touch screen. Then, we refer to three data sets typically CMU [4], GREYC [15] and group Margit Antall [1] to perform tests and reviews a new measurement. The last part presents some conclusions and future research directions.

II. KEYSTROKE DYNAMICS

In contrast to other biometric methods, keystroke dynamics research does not require any specialized hardware devices. When taking a dynamic model of the keystroke is done only by software running in the background, which makes this method is very easy and does not impact the user. KD can be used for both cases according to the time of authentication static and continuous, compared to other authentication methods, the biometric method has a drawback is lower accuracy.

Data collected by the research reports KD uses many different input device, from the regular keyboard to the keyboard has a pressure sensor. The most common is to use time-based characteristics, they are Dwell time and Flight time. Dwell time is the period of time from the time press a key to release it out (sometimes called holding time), Flight time is the time period from the release of some key to press the next key.

Sometimes three or more consecutive time events related to key used as the characteristics (N-graphs) [9.10].

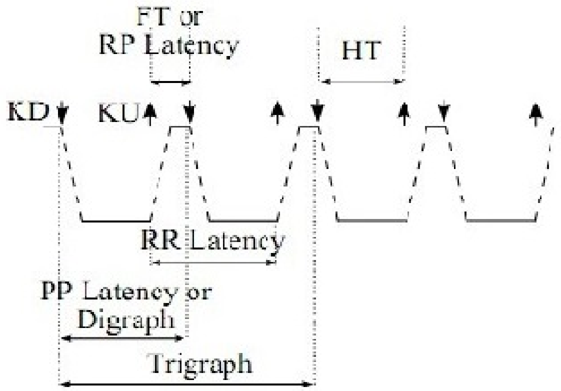

Fig. 1. N - graphs Keystroke Timing Information

KU: the time of release the key; KD: the time press the key; FT: is the time period from the release of this key to press next; HT: holding period; PP: is the period of time from the press came at the next key press; RR: about time from release some key to release next key; RP: about time from release some key to press next key.

In most of the papers [2,4] only use the characteristics of the two successive keys. Most of the proposed to identify model had already been tested in dynamic recognition of the keystroke, including the approach to statistics and machine learning.

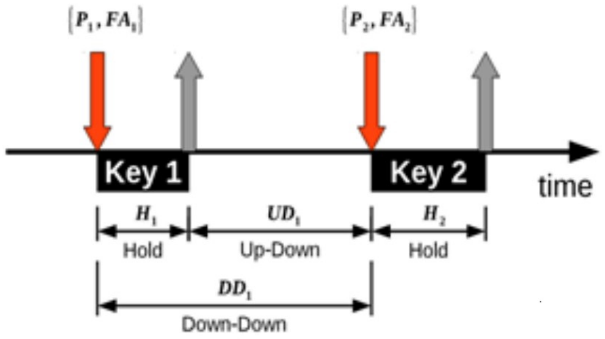

Fig. 2. Two - graphs from a mobile device [1]

H: Hold time; UD: Up-Down time; DD: Down-Down time (key transfer period); P: Pressure (finger pressure); FA: Finger Area (an area of exposed fingers)

Simple authentication methods is to build the model references for the users and the distance between the current model type and the reference model type accordingly. This method is called match the pattern and can be coordinated with other measurements, from simple measurement is the Euclidean metric to the Mahalanobis one [3]. Neural networks and support vector machines (SVM) is the best. In addition, a number of further detectors as follows: Nearest-neighbor, Neural-network, Fuzzy-logic, the Outlier-counting (z-score, SVM (one class), Fc-means.

In 2010, the research group of Romain Giot, Baptiste Hemery, Christophe Rosenberger [13] has proposed a multimodal biometric system combining keystroke dynamics and 2D face recognition. The objective of the proposed system is to be used while keeping in mind: low cost , good performances, acceptability, and respect of privacy. They have used different combined methods (min, max, mul, svm, weighted sum configured with genetic algorithms and genetic programming) on the scores of three keystroke dynamics algorithms and two 2D face recognition ones. This multimodal biometric system improves the recognition rate in comparison with each individual method. On a chimeric database composed of 100 individuals, the best keystroke dynamics method obtains an EER of 8.77 %, the best face recognition one has an EER of 6.38 %, while the best proposed fusion system provides an EER of 2.22 %.

March 2016, the research group of Abir Mhenni, Christophe Rosenberger [11] showed, in fact the characteristic attributes are extracted from the keystroke dynamics of an individual becoming less representative than from time to time. This can lead to error in biometric authentication tasks. Since the change over time of such properties, so the representative template must always be updated. They used the Windows growing and sliding as the template update method based on the classification of statistics. They also prove that user-specific threshold need to change according to each different updated version, that allows to reduce the rate of errors than when only updates a fixed threshold.

April 2016, the research group of Paulo Henrique Pisani, Romain Giot,.. in their paper [16] showed that the biometric features may undergo changes over time, which can reduce the predictive performance of the biometric system. Indeed, how the user types a password evolves with time and can be different in a short timespan. The reasons are numerous and cannot always be controlled: increased practice, changes on the environment, etc. For example, users can increase the speed to write the password due to more practice. These modifications increase the intraclass variability which, consequently, can increase the ratio of authentication failure. Template update adapts the user model to deal with these changes and, therefore, decreases the predictive performance loss. A strategy to reduce this performance decrease is to adopt a template update mechanism (sometimes referred to as an adaptive biometric system) [17,18]. The aim of the template update is to automatically adapt the biometric model/reference of the user to make it closer to the user current biometric data (i.e., decreasing the deviation due to template ageing). The majority of the papers in the area updates the user model only with biometric samples classified as genuine/positive, discarding those classified as impostors (negative). They usually employ a positive gallery, which is a set of biometric samples classified as genuine/positive. The paper [16] proposed to investigate if taking into account samples classified as impostors can improve the adaptive procedure. Thus, there would be a negative gallery too. This new approach, named Enhanced Template Update, uses all collected unlabeled samples to support the adaptation process. According to their experimental results, this new approach can improve the predictive performance when compared to current methods depending on the scenario. Some improvements on the visualization of results over time were also proposed during the analysis performed in their study.

III. THE PROPOSED NEW METRIC

Model KD is a promising way to get representation for the data of the user. Therefore, the majority of the work [5], [3] [6], focusing on the selection of the distance metric and the mean vector selection as a default model. However the mean vector [7] does not really guarantee the optimal distance, a well-known Weiszfeld method [8] for solving the facility location problem indicates that let $X := \{x_i : i \in \overline{1:N}\}$ be a set of N data points in R^n,in order to find a point $c \in R^n$ minimizing the sum of distances

$$dist(X,c) = min_{c \in R^n} \sum_{i=1}^{N} d(x_i, c) \tag{1}$$

where: $d(p,q) = \|p-q\|$ denotes the Euclidean distance of two vectors $p,q \in R^n$ should use the Weiszfeld iteration.

The gradient of $d(X,x)$ is undefined if c coincides with one of the data points x_i. For $c \notin X$

$$\bigtriangledown dist(X,c) = -\sum_{i=1}^{N} \frac{x_i - c}{\|x_i - c\|} \tag{2}$$

The optimal center c^*,if not in X , is characterized by $\bigtriangledown dist(X,c^*)=0$,expressing it as a convex combination of the points $x_i, c^* = \sum\limits_{i=1}^{N} \lambda_i x_i$ with weights $\lambda_i, c^* = \frac{1/\|x_i-c\|}{\sum\limits_{k=1}^{N} 1/\|x_k-c\|}$ that depends on c^*. This circular result gives rise to the Weiszfeld iteration

$$c_+ := T(c) \tag{3}$$

$$T(c) := \begin{cases} \sum\limits_{k=1}^{N} \frac{x_i/\|x_i-c\|}{\sum\limits_{k=1}^{N} 1/\|x_k-c\|} & \text{khi } c \notin X \\ c & \text{khi } c \in X. \end{cases} \tag{4}$$

Where c_+ is then updated center,c is current center, and $T(c) := \begin{cases} \sum\limits_{k=1}^{N} \frac{x_i/\|x_i-c\|}{\sum\limits_{k=1}^{N} 1/\|x_k-c\|} & \text{khi } c \notin X \\ c & \text{khi } c \in X. \end{cases}$

Furthermore, using similar scheme for Manhattan distance, when $d(p,q) = \|p-q\|_{LI} = \sum\limits_{i=1}^{N} |p_i - q_i|; p,q \in R$,the gradient of $dist(X,c)$ is represented as follows:

$$\bigtriangledown dist(X,c) = -\sum_{i=1}^{N} sign(x_i - c) \tag{5}$$

In this case, the optimal c^* is the median (not the mean) vector of X The above discussion attests to the fact that the KD model (the center of the KD point cloud) must be identified with the distance metric respectively. The process of the survey collected data on KD can realize the retrieved average value (mean) of the feature might not reflect well the nature of the data (see Figure 3), because there is more data in that other allocation far compared to the rest.

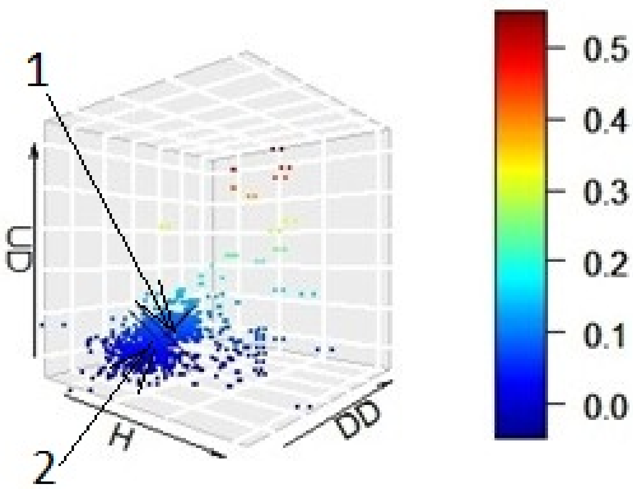

Fig. 3. KD data allocation on 3D graph; 1-Mean (average); 2-Median (Central)

Distance metric is a key issue in many classification algorithms. The commonly-used distance metrics as Euclidean, Manhattan e.g. assume that each feature of data point is equally important and independent from others. This assumption may not be always satisfied in real situations, especially when dealing with high dimensional data where some features may not be tightly related to the topic of interest. In contrast, a distance metric with good quality should identify important features and discriminate relevant and irrelevant features.

Therefore, providing a good distance metric is a particularly important issue and decides the success or failure of the learning algorithm or developed system. The basic idea of this paper is to select and transform the data characteristics into a new feature space are standardized, and then design a new distance metric has to mention the unexpected relationship between features and reduce the influence of the abnormal points in order to improve performance.

Formally, in the training phase, we apply the same idea as in [3] to filter and scale training data. First of all, the vectos ith the high deviations were removed from the training data using Mahattan distance:

$$X^* := \{x : x \in X, \|x-m\|_{Li} \leqslant \frac{\sum\limits_{k=1}^{N} \|x_k - m\|_{Ll}}{N}\} \tag{6}$$

where: $m = median(X)$ Once get the filtered training data set X^*, every feature in each vector $x_i \in X^*, i = \overline{1..M}$ is scaled by the mean absolute deviation $\sigma = (\sigma^1, \sigma^2, ..., \sigma^n)$:

$$x_i = (x_i^1, x_i^2, ..., x_i^n) := (\frac{x_i^1}{\sigma^1}, \frac{x_i^2}{\sigma^2}, .., \frac{x_i^n}{\sigma^n}) \tag{7}$$

where:

$$\sigma^k = \frac{\sum\limits_{j=1}^{M} |x_j^k - \frac{\sum\limits_{l=1}^{M} x_l^k}{M}|}{M} \tag{8}$$

The new proposed distance metric will adequately evaluate the distance by assigning different importance factors to the features of data points is defined by nonlinear function as follows:

$$dist(\overline{x}, \overline{y}) = \sum_{i=1}^{N} ln(1 + |x_i - y_i|/a_i) \tag{9}$$

Observe that the logarithm function gives distance d_N more robust with large changes (for imposters rejection purpose) and this function is more sensitive to small changes (for legitimate user acceptance purpose). On the other hand, the gradient vector and optimal model (center) can be easily calculated:

$$\bigtriangledown dist(X,c) = \bigtriangledown \sum_{i=1}^{M} d_N(x_i,c) = 2\sum_{i=1}^{M} \frac{x_i - c}{1+(x_i-c)^2} \tag{10}$$

Note that, the mathematical operators are applied element -by-element. Zeroing the gradient (10) we get:

$$\sum_{i=1}^{M} \frac{x_i}{1+(x_i-c)^2} = c\sum_{i=1}^{M} \frac{1}{1+(x_i-c)^2} \tag{11}$$

Equation (11) induces mapping

$$c_+ := \frac{\sum\limits_{i=1}^{M} \frac{x_i}{1+(x_i-c)^2}}{\sum\limits_{i=1}^{M} \frac{1}{1+(x_i-c)^2}} \tag{12}$$

Finally, convergence of the iteration (12) can be established as in the Weiszfeld algortihm [3]. In the next section the convergence of the training process will be illustrated by numerical experiments. The training process can be described in Algorithm 1 with the output is a template δ. This template δ is applied to construct a reference template for respective user and compute the distance between the current typing pattern and the reference template in the authentication stage as shown in Algorithm 2.

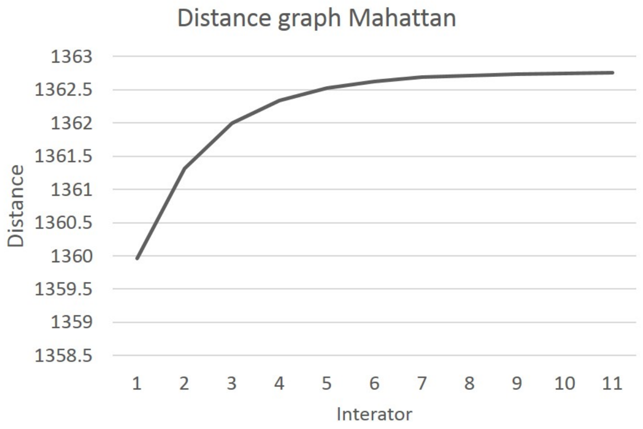

Fig. 4. The convergence of training process

Algorithm 1 KD model calculation

```
Input: $X := \{x_i : i \in \overline{1:N}\}$ set of feature vectors; the old thershold $\varepsilon$
Pre-processing: filter using (6); compute vector $\sigma$ (8) and new dataset $X^*$ with $M$ vectors using (7).
Initialization: assign model $c$ to the median vector of $X^*$.
Iteration:
compute distance $dist(X^*, c) = \sum_{i=1}^{M} d_N(x_i, c)$
update model $c_+$ using (12)
compute new distance $dist(X^*, c_+) = \sum_{i=1}^{M} d_N(x_i, c_+)$
If: $|dist(X^*, c) - dist(X^*, c_+)| < \varepsilon$ then
    return: template $\sigma = \{c, \delta\}$
end if:
```

Algorithm 2 KD verification

```
Input: $X := \{x_i : i \in \overline{1:N}\}$ set of feature vectors; KD template $\sigma = \{c, \delta\}$; threshold $\theta$
Pre-processing: scale $X$ using (7) with vector $\sigma$ from template $\delta$
Verification:
compute distance $dist(X^*, c) = \sum_{i=1}^{M} d_N(x_i, c)$
If: $dist(X^*, c) < theta$ then
    return TRUE
else
    return: FALSE
end if:
```

Biometric systems can have two distinct functions: verification and identification. Verification is a binary decision problem, in which the system accepts or rejects the identity claimed by the user. Identification, also called recognition, is a classification problem: the system classifies the input pattern into one of the N known classes.

The quality of biometric systems is usually characterized by three kinds of errors: FAR, FRR and EER. False Acceptance Rate (FAR) is the rate at which a biometric system accepts a sample as one belonging to the claimed identity when the sample belongs to an impostor. False Rejection Rate (FRR) is the rate at which a biometric system incorrectly rejects a sample provided by the genuine user. EER is the rate at which FAR is equal to FRR.

IV. THE EXPERIMENTS

In 2009, Killorhy and Maxion collected and published the test data set of Keystroke Dynamics (the CMU data) [4] which included 51 subjects with KD 400 is collected for each person. Moreover, they had reviews 14 KD algorithms available based on this data set. The different distance metrics, including the Euclidean distance and the Mahattan, were used.

In 2009, R.Giot, M. El-Abed, C. Rosenberger had collected and published the data set of keystroke dynamics (data GREYC) [14.15]. Some of the simulations were performed with SVM (support vector machine) for calculations in many different aspects of the data, in particular identification of ages, gender, positive hand, identity authentication ...

In 2014, Margit ANTAL, Lszl Zsolt SZAB, Izabella LASZLO had collected and published the test data collecteda from the keyboard on mobile devices with platform Androi [2] has a touch screen comprising 42 subject with 51 KD were collected for each person. Moreover, they have proved by experiments that the properties based on the touch screen, significantly improved the method of KD in the classification and authentication. During the measurements, the addition of the characteristic from the touch screen to default properties have increased over 10% accuracy for the classification. The improvement is very hard to explain in case the measurements of authentication because the error rate by only about 2.4% reduction (in the measurement of Manhattan).

A. *Experiments setup*

1) *CMU dataset:*

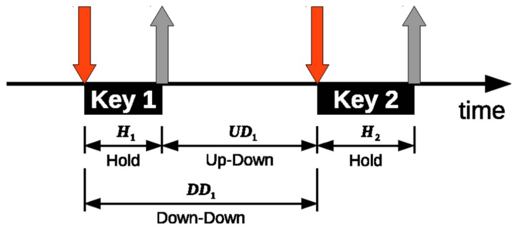

Fig. 5. Vectos features $(H_1, DD_1, UD_0, H_2, DD_2, UD_2, ...)$

Features name: Key hold time (H); Down-down time (DD); Up-down time (UD);

2) MARGIT dataset:

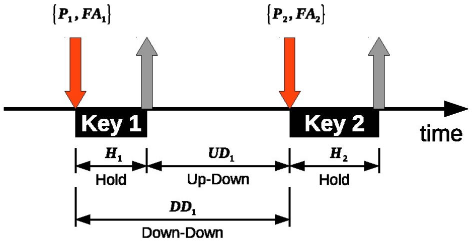

Fig. 6. Vectos features ($H_1, H_2, H_3, ..., DD_1, DD2, DD_3, ..., UD_1, UD_2, UD_3, ..., P_1, P_2, P_3, ..., FA_1, FA_2, FA_3, ..., MH, MFA, MP$)

Features name: Key hold time (H); Down-down time (DD); Up-down time (UD); Key press pressure (P); Finger area (FA); Mean hold time (MH); Mean finger area (MFA); Mean pressure (MP).

3) GREY-C dataset:

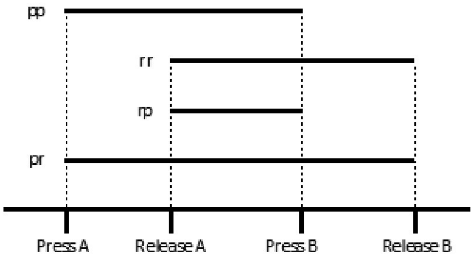

Fig. 7. Vectos features (PP,RR,PR, RP)

- ppTime (PP): the latencies of when the two buttons (keys) are pressed;
- rrTime (RR): the latencies of when the two buttons (keys) are released;
- prTime (PR): the durations of when one button (key) is pressed and the other is released;
- rpTime (RP): the latencies of when one button (key) is released and the other is pressed;

Features (PP,RR,PR,RP) convert to (H,DD,UD), where:
$H = PP - RP; DD = PP; UD = RP;$

B. Authentication result by the measurements for the data set of Killourhy Maxion Group

Measurement to verify is done by using the R language script provided by Killourhy & Maxion [3]. This script provides a calculation of rate EER for three unusual detector based on the measurements of Euclidean, Manhattan, and the Mahalanobis. The name of the data set datafile 'DSL-StrongPasswordData.txt'; Character strings are entered:**.tie5Roanl**; the number of attributes per login: 31 properties (**time based**) the number of people surveyed: 51; the number of input times: 400 times; a total of 8 session (50 times per session). The following is the table of results:

Measurement results in Fig.8 shows that with this new metric reduces ERR to 0.062 (i.e. increase efficiency was 43.63% compared to current best measurement is the Mahalanobis with ERR = 0.110). If continue to increase the number of training samples up to 375 times the efficiency increase is still up substantially.

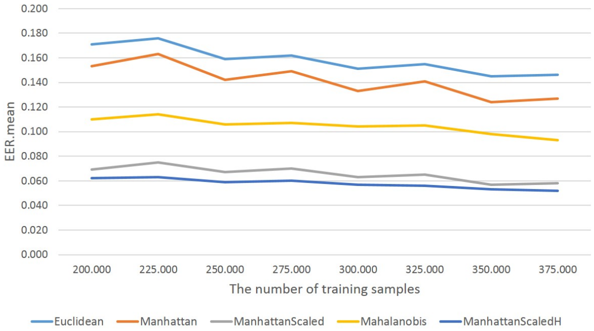

Fig. 8. EER ratios comparision Chart

C. Results of authentication using the measurements on the data set of the Group R.Giot, M. El-Abed, C. Rosenberger (GREYC data sets)

The measurements to verify made using the R language script provided by Killourhy & Maxion [3]. This script provides a calculation of EER for the three abnormal detectors based on measurements of Euclidean, Manhattan, and Mahalanobis. The name of the data set: datafile ”keystroke.db”; character strings are entered: **greyc laboratory**; the number of attributes per login: 50 properties (**time based**); the number of people surveyed: 100 persons; number of visits: 50 times; total 16 session. The following is the table I of the results achieved.

TABLE I
EER OF NEW MEASUREMENT RESULTS ON THE DATA SET GREYC

Detector	eer.mean	eer.sd
Euclidean	0.206	0.125
Manhattan	0.139	0.104
Mahalanobis	0.159	0.088
ManhattanScaledH	**0.067**	**0.055**

D. Authentication result by the measurements for the data set of Margit Antal Group

The measurements to verify is done by using the R language script provided by Killourhy & Maxion [3]. This script provides a calculation of rate EER for three unusual detectors based on the measurements of Euclidean, Manhattan, and Mahalanobis. The data is standardized and divided into 3 equal parts, each part contains 17 password of each user. The first two parts are used to create the user model and the rest to check the FRR. Five of the first passwords from the data of each person, except for the test model, which are used to test FAR. The name of the data set datafile 'keystroke normalized.arff'; Character strings are entered: **.tie5Roanl**; the number of attributes per login: 71 properties (**time based + touchscreen**); the number of people surveyed: 42; the number of input times: 51 times (resulting in three equal portions, 2 the first parts of the training data, the rest for verifying model, 5 the first samples (except for verifying model) of the data are used to check the data to impersonate), total 2 session.

TABLE II
EER OF NEW MEASUREMENT RESULTS ON THE DATA SET OF MARGIT ANTAL GROUP

. **Detector**	**eer.mean** (Margit Antal et al)	**eer.mean**	**eer.sd**
Euclidean	0.157	0.156	0.154
Manhattan	0.129	0.126	0.129
Mahalanobis	0.166	0.159	0.099
ManhattanScaledH		**0.069**	**0.071**

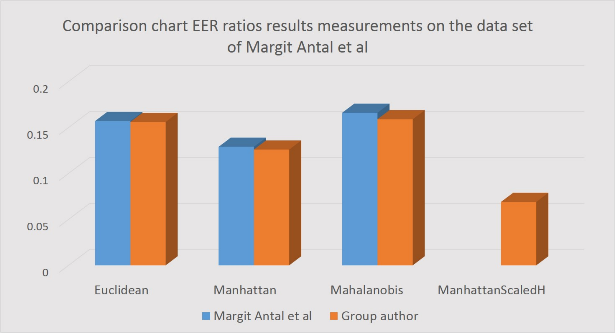

Fig. 9. Chart of EER results

We have table II and the comparison chart figure 9 of EER results as above. Through the chart we see that, with this new measurement the relative EER rates had fallen about 45%

V. CONCLUSION

In this paper, for the first time, we have studied the relationship between the data model of keystroke dynamics with the distance metrics. A new distance metric and the algorithm used for training data model has also been presented. We have conducted a series of experiments using the different distance metrics. The experimental results showed that with this new metric we can reduce ERR to 0.069.

The next study will focus on reducing the number of favorable patterns (the amount of data to be collected for the user), test measurements in the environment using the user's password and mobile devices in real environments, tested on Iphone devices not Androi, put more entropy review for each a user's password or adding new properties or improve this new distance metric. Furthermore we will focus on the differences between static typing and dynamic continuity in terms of creating a data model and the sorting algorithm.

We will also study using this new metric combines the advanced model update proposed in the paper [16].

REFERENCES

[1] M. Antal, L. Z. Szabo, and I. Laszlo., *Keystroke dynamics on android platform.*,19:820-826, 2015.

[2] M. Antal, L. Z. Szabo, and I. Laszlo., *Keystroke Dynamics Data Set.*, 2015.

[3] Kevin S. Killourhy and Roy A. Maxion., *"Comparing Anomaly Detectors for Keystroke Dynamics,"*,in Proceedings of the 39th Annual International Conference on Dependable Systems and Networks (DSN-2009), pages 125-134, Estoril, Lisbon, Portugal, June 29-July 2, 2009. IEEE Computer Society Press, Los Alamitos, California, 2009.

[4] R. M. Kevin Killourhy., *Keystroke Dynamics - Benchmark Data Set*, accessed October 10,2015.

[5] R. Giot, M. El-Abed, and C. Rosenberger., *Keystroke dynamics authentication.*, Biometrics, pages chapitre-8, 2011.

[6] Y. Zhong, Y. Deng, and A. K. Jain., *Keystroke dynamics for user authentication.*, In Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE,Computer Society Conference on, pages 117-123. IEEE, 2012.

[7] M. Ferrer, E. Valveny, F. Serratosa, I. Bardaji, and H. Bunke., *Graph-based k-means clustering: A comparison of the set median versus the generalized median graph.*, In Computer Analysis of Images and Patterns.

[8] A. Beck, M. Teboulle, and Z. Chikishev. *Iterative minimization schemes for solving the single source localization problem.*, SIAM Journal on Optimization, 19(3):1397-1416, 2008.42-350. Springer, 2009.

[9] Trojahn M, Arndt F, Ortmeier F.*Authentication with Keystroke Dynamics on Touchscreen Keypads - Effect of different N-Graph Combinations.*,In: MOBILITY 2013, The Third International Conference on Mobile Services, Resources, and Users, 2013, p. 114-119

[10] F. Bergadano, D. Gunetti, and C. Picardi.*User authentication through keystroke dynamics.*,ACM Transactions on Information and System Security, 5(4):367-397, 2002.

[11] Abir Mhenni, Christophe Rosenberger, Estelle Cherrier, Najoua Essoukri Ben Amara.*Keystroke Template Update with Adapted Thresholds.*,International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Mar 2016, Monastir, Tunisia.

[12] Syed Zulkarnain Syed Idrus, Estelle Cherrier, Christophe Rosenberger, Patrick Bours, *Soft Biometrics Database: A Benchmark For Keystroke Dynamics Biometric Systems.*,IEEE Conference BIOSIG, 2013, Darmstadt, Germany. 8 p., 2013.

[13] R. Giot, B. Hemery, and C. Rosenberger,*Low Cost and Usable Multimodal Biometric System Based on Keystroke Dynamicsand 2D Face Recognition,*,in Proc. IAPR International Conference on Pattern Recognition (ICPR), Istanbul, Turkey, 2010.

[14] R. Giot, M. El-Abed, C. Rosenberger,*GREYC Keystroke: a Benchmark for Keystroke Dynamics Biometric Systems*,IEEE Third International Conference on Biometrics: Theory, Applications and Systems (BTAS), Washington DC USA, Sept. 28-30, 2009.

[15] Greyc-keystroke-dataset

[16] Paulo Henrique Pisani, Romain Giot, Andre C.P.L.F. De Carvalho, Ana Carolina Lorena.*Enhanced template update: Application to keystroke dynamics.*Computers and Security,Elsevier, 2016, 60, pp.134-153.

[17] F. Roli, L. Didaci, G. Marcialis *Adaptive biometric systems that can improve with use, in: N. Ratha, V. Govindaraju (Eds).* Advances in Biometrics, Springer London, 2008, pp. 447-471.

[18] N. Poh, A. Rattani, F. Roli *Critical analysis of adaptive biometric systems, Biometrics.* IET 1 (4) (2012) 179-187.