Themed Section : Engineering and Technology
© 2018 IJSRSET | Volume 5 | Issue 3 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099

Biometric Authentication of Age and Gender prediction using GREYC Keystroke Dynamics Dataset

R. Abinaya[1], Dr. AN. Sigappi[2]
Research scholar[1], Associate Professor[2]
Department of computer Science and Engineering, Annamalai University, Chidambaram, TamilNadu, India

ABSTRACT

Keystroke dynamics allows to authenticate individuals through their way of typing on a computer keyboard. In this study, this paper interested in static shared secret keystroke dynamics (all the users type the same password). It can be combined with passphrases authentication resulting in a more secure verification system. This paper presents a new soft biometrics information which can be extracted from keystroke dynamics patterns: The Age and gender of the user when he/she types a given password or passphrase on a keyboard. This experiments were conducted on a web based keystroke dynamics database of 118 users and our experiments on keystroke authentication it exploits the features from 2D Discrete wavelet transformation (DWT) to characterize the keystroke dynamics, and provides results from classification algorithms. BPNN classifier it obtained best results achieved were 86.2% accuracy respectively.

Keywords: Biometrics, keystroke dynamics, Soft biometrics, authentication, Back Propagation Neural Network (BPNN), Discrete Wavelet Transforms (DWT).

I. INTRODUCTION

Keystroke dynamics is an interesting and a low cost biometric modality as it enables the biometric system to authenticate or identify an individual based on a person's way of typing a password or a passphrase on a keyboard [1], [2]. It belongs to the class of behavioral biometrics, in the sense that the template of a user reflects an aspect of his/her behavior. Among the behavioral biometric modalities, we can mention signature analysis, gait recognition, voice recognition, or keystroke dynamics. Generally speaking, the global performances of keystroke dynamics based authentication systems are lower than the popular morphologic modalities based authentication systems (such as fingerprints, iris, *etc*...) [3]. Besides, the main advantage of resorting to keystroke dynamics [4], [5] to authenticate a user relies in its low cost. Indeed, for this modality, no extra sensor is required. The fact that the performances of keystroke dynamics are lower than other standard biometric modalities can be explained by the variability of the user's behavior. One solution to cope with this variability is to study soft biometrics, first introduced by Jain *et al.* in [6]. In this paper "*soft biometric traits*" are defined as "*characteristics that provide some information about the individual, but lack the distinctiveness and permanence to sufficiently differentiate any two individuals*". For example, Jain *et al.* consider gender, ethnicity, and height as complementary data for a usual fingerprint based biometric system.

Soft biometrics allows a refinement of the search of the genuine user in the database, resulting in a computing time reduction. For example, if the capture corresponds to a male according to a soft biometric module, then, the standard biometric authentication system can restrict its research area to male users, without considering female ones.

Concerning keystroke dynamics, an original approach is presented in the work of Epp *et al.* [7], strongly linked with the behavioral feature of keystroke dynamics. The authors show that it is possible to detect the emotional state of an individual through a person's way of typing. In this case, detecting anger and excitation is possible in 84% of the cases. Gender recognition is dealt in the work of Giot *et al.* in [8]: they show that it is possible to detect the gender of an individual through the typing of a fixed text. The gender recognition rate is more than 90% and the use of this information in association to the keystroke dynamics authentication, reduces the Equal Error Rate (EER) by 20%. The work of Syed-Idrus *et al.* [9] show that it is possible to detect users' way of typing by using one finger (i.e. one hand) and more than one fingers (i.e. two hands) with 80% correct recognition accuracy performed on a dataset with three passwords.

The objective of this paper is twofold. First, we present a new data collection of 118 users. This propose an extended study of soft biometrics for keystroke dynamics on this Greyc database. We are interested in the criteria that can influence the way of typing of the users. We test if it is possible to predict if the user:

1. is a male or a female?
2. belongs to a particular age category

Indeed, predicting these soft features may help an authentication system (which is not considered here) to reduce the computing burden while in search of the genuine user in the database. This paper is organized as follows. Section 2 is devoted to the description of the proposed methodology: the characteristics of the database are described, together with the data collection process, and the tools that are used for analysis purposes. In Section 3, we present the obtained results while Section 4 presents the conclusions and the future works to be addressed.

II. Proposed work of Keystroke dynamics dataset for predicting age and gender category:

In this project, the authentication of person by keystroke dynamics by Greyc Web Based Keystroke Dynamics Dataset they proposed 2D-DWT feature Transformation BPNN classifier to predict the age and gender category of the users.

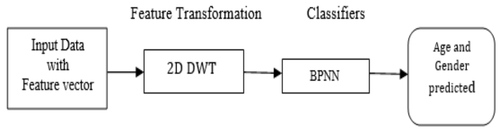

Fig.1: Flow Chart of Proposed Research Work

A. Dwt: 2d (Discrete Wavelet Transforms)

The decomposition is applied at different levels repeatedly on low frequency channel (LL) to obtain next level decomposition. The image is decomposed into four subbands LL, LH, HL, and HH subbands by applying 2D DWT on keystroke dynamics data passphrases [10]. The LL subband corresponds to low frequency components of an image and HL, LH and HH are high frequency components of a passphrases corresponds to vertical, horizontal and diagonal subbands respectively. The LL subband we obtain is half the original data. Figure 2 shows the typing speed passphrases decomposition based on wavelet scales [11], [12] 2D DWT gives dimensional reduction for less computational complexity, insensitive feature extraction, and multiresolution data approximation. The transform decomposes a typing passphrase and hence different facial expressions are attenuated by removing high frequency components. Wavelet coefficients are obtained by convolving a target function with wavelet kernels and mathematically DWT [10]

The 2-D-DWT binary-tree decomposition. For each level, the input signal is filtered along rows and the resulted signal is filtered along columns [20]. In this way, the 2-D decomposition of an input signal, with columns and rows, is described by the following equations:

High frequency coefficient of decomposition of level j are computed as follows:

$$L_{J+1}[n] = \sum_{i=0}^{N_{L-1}} w[i] \times L_J[2_{n-i}] \qquad (3)$$

$$H_{J+1}[n] = \sum_{I=0}^{N_{L-1}} h[i] \times L_J[2_{n-1-i}] \qquad (4)$$

When j= {0,1......L-1}, n= {0,1…N/2 $^{j+1}$-1} and L_O[n]=1N[n]

The 2D-DWT binary tree decomposition is illustration in the figure for each level the input signal is filtered along columns. In this way the 2D decomposition of an input signals 1N[M][N], with M columns and N rows is described by the following equation [17].

$$H_{j+i}[row][m] = \sum_{I=0}^{N_{L-1}} h[i] \times LL_j[row][2m - 1 - i] \qquad (5)$$

$$LL_{j+i}[n][col] = \sum_{I=0}^{N_{L-1}} w[i] \times L_j[2n - i][col] \qquad (6)$$

$$HL_{j+1}[n][col] = \sum_{i=0}^{N_{L-1}} w[i] \times H_j[2n - i][col] \qquad (7)$$

Where j={0,1…..L-1),row = {0,1….N/ $2^j - 1$ },m={0,1…M/ $2^{j+1} - 1, col = \{0,1 \ldots N/2^{j+1}$ -1}, n= {0,1…N/2^{j+1}-1},and LL_0[n][m].

In the rest of this document, we use the term layer to indicate both intermediate and output signals, while level is used for each decomposition stage. In the rest of this document, we use the term layer to indicate both intermediate and output signals while level is used for each decomposition stage [13], [14].

Learning in a multilayer network proceeds the same way as for a perceptron. a training set of input patterns is presented to the network that networks computes to its output pattern and if there is an error Origin other words a difference between actual and desired output patterns – the weights are adjusted to reduce this error

In a back-propagation neural network, the learning algorithm has two phases.

First, a training input pattern is presented to the network input layer. The network propagates the input pattern from layer to layer until the output pattern is generated by the output layer.

If this pattern is different from the desired output, an error is calculated and then propagated backwards through the network from the output layer to the input layer. The weights are modified as the error is propagated.

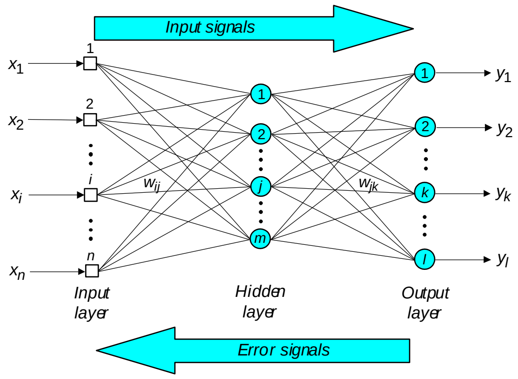

Fig: no.2 Three-layer back propagation neural network:

Step 1: initialization.

Set all the weights and threshold levels of the network to random numbers uniformly distributed inside a small range:

$$\left(-\frac{2.4}{F_I}, +\frac{2.4}{F_I}\right) \qquad (8)$$

where *Fi*is the total number of inputs of neuron *i*in the network. The weight initialization is done on a neuron-by-neuron basis.

Step 2: Activation

Activate the back-propagation neural network by applying inputs $x_1(p)$, $x_2(p)$, $x_n(p)$ and desired outputs $y_{d,1}(p)$, $y_{d,2}(p)$…, $y_{d,n}(p)$.

(a) Calculate the actual outputs of the neurons in the hidden layer:

$$y_j(p) = \mathrm{sigmoid}\left[\sum_{i=1}^{n} x_i(p)\,.\,w_{ij}\right](p) - \theta_j \quad (9)$$

Where *n* is the number of inputs of neuron *j* in the hidden layer, and sigmoid is the *sigmoid* activation function.

Step 2: Activation (continued)

Calculate the actual outputs of the neurons in the output layer:

$$y_k(p) = sigmoid\left[\sum_{j=1}^{m} x_{jk}\,(p).\,w_{jk}(p) - \theta_k\right] \quad (10)$$

where *m* is the number of inputs of neuron *k* in the output layer.

Step 3: Weight training

Update the weights in the back-propagation network propagating backward the errors associated with output neurons.

Calculate the error gradient for the neurons in the output layer

$$\delta_k(p) = y_k(p).\,[1 - y_k(p)].\,e_k(p) \quad (11)$$

$$e_k(p)\text{=}y_k(p).\,[1 - y_k(p).\,e_k(p)] \quad (12)$$

Calculate the weight corrections

$$\Delta w_{jk}(p) = \alpha.\,y_j(p).\,\delta_k(p) \quad (13)$$

Update the weights at the output neurons

$$w_{jk}(p+1) = w_{jk}(p) + \Delta w_{jk}(p) \quad (14)$$

Step 3: Weight training (continued)

Calculate the error gradient for the neurons in the hidden layer:

$$\delta_j(p) = y_j(p).\,[1 - y_j(p)] \times \sum_{k=1}^{1} \delta_k(p) w_{ij}(p)$$

$$\Delta w_{ij}(p) = \alpha.\,x_i(p).\,\delta_j(p)$$

$$\Delta w_{ij}(p+1) = w_{ij}(p) + \Delta w_{ij}(p) \quad (15)$$

Step 4: Iteration

Increase iteration *p* by one, go back to *Step 2* and repeat the process until the selected error criterion is satisfied.

B. Modelling Technique:

Recognition of Age and Gender Category by Keystroke Dynamics of 2d-Dwt Features with BPNN Classifier:

In BPNN classifier the Different letter password data are obtained from the dataset here 255 dimensions with 5 persons the input features are converted in to 1664 features by using 2D DWT transformation. in BPNN 20 hidden layer size with 70 training set and 30 testing for classification process getting 4class output. By different length of password for each user each layers trained the values as feed forward process. Hence by BPNN the iteration value is 100 and the 4 output is Male >20 obtained 87, the male <20 obtained 86% female>20 is 87 and female<is 87 respectively by these four age and gender category male <hast highest accuracy rate by 2 hidden layers with 118 no of users as input.

Table 1.BPNN –classifier in Unknown password data results

BPNN Classifier	TPR	FPR	Accuracy	Error detection
male above 20	86.26	13.74	87.23	87.23
Male below20	86.24	13.76	86.53	86.53
Female above 20	86.7	13.3	87.45	87.45
Female below 20	86.17	13.83	87.45	87.45

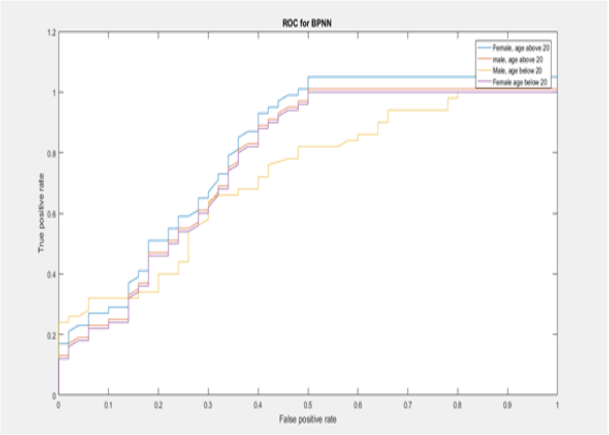

Fig 3. Roc curve of BPNN –classifier in different password data results

The above chart describes the different dimension level by DCT, 2D-DWT accuracy rate with various classifiers. Hence the performance matrix shown by the ROC curve.

III. Experimental Results and Discussion:

A. Dataset description:

The dataset is stored in files organized per directory for user the 'user' directory contains one file per user named user/user_xxx.txt, with xxx the id of the user [15],[16] Each user file contains the following information (one information per line) the login of the user, the name of the user, the gender of the user, age of the user

Login and name are chosen by the user during the first session. gender and age are filled during the account creation.

Common passphrases:

The 'passphrases' directory contains the inputs when the users typed the imposes login and password. There is a sub – dataset similar to all the public benchmarks: all the users type the same login and password. Samples of each user are stored in a directory name 'passphrases \user.xxx' where 'xxx' is the user id. The samples folder name is based on the timestamp of the sample. The list of the sample is presented in the text file named 'passphrases\user.xxx\captures.txt.\'. each line is the folder name of one sample. the index file is ordered by chronological order. unique password.

The password directory contains the inputs of the sub-dataset where each user possesses his own login and password. The genuine and imposter samples of user xxx are stores in the folder password and are indexed by password\user.xxx\genuine\capture.txt. The genuine samples folder nemesis based on the timestamp of the sample. Folder password\user.xxx\imposter contains the list of imposter samples (other users trying to impersonate xxx). the samples are indexed in passwords\user.xxx\imposter\captures.txt. the imposter samples folder name is based on the id of the imposter and the timestamp of the samples. thus, passwords\user_001\imposter \imposter_002_2010-10-18711:39:14 if the sample of an imposter wanting to impersonate user 001.sample representation. Each sample is stored in a folder containing meta-data, raw data (in order to let researchers to compute their own extracted features), extracted features commonly used in the literature.

Here is the display each file

UserAgent.txt: the user agent string of the web browser used to type (can be used to analyses the browser habits of the user)

Userid.txt: the id of the user who have typed the text

Data.txt: the acquisition date of the samples

Genuine .txt: A file containing 1for a samples typed by the user and 0 for a sample typed by an imposter

Login.txt: The string of the login

Password.txt: the string of the password

1_raw_press.txt: The press events of the login. one event per line with: the code of the key, the timestamp of event.

1_raw_release.txt: the release events of the login .one event per line with: the code of the key, the timestamp of the event.

P_raw_press.txt: The press events of the password. one event per line with: the code of the key, the timestamp of the event.

P_raw_release.txt: the release events of the password. One event per line with: the code of the key, the timestamp of the event.

1_pp.txt: the extracted press to press time of the login.

1_pr.txt: the extracted press to release time of the login.

1_rp.txt: the extracted release to press tie of the login

1_rr.txt: the extracted release to release tie of the login

1_release_codes.txt: the list of codes of key released.

1_total.txt: the total typing time of the login

P_pp.txt: the extracted press to press time of the password

P_rp.txt: the extracted release to press time of the password

P_rr.txt: the extracted release to release time of the password

P_release_codes.txt: list of codes of key release

P_total.txt: the total typing time of the password

Table .2: Dataset Description

Number of users:	118
Number of genuine sample:	9 087
Number of impostor samples:	10 043
Number of imposed samples:	9346

Table .3: Repartition of samples

User	118 Users
Gender	86 males 17 females
Age Category (between 16 and 35 years old)	≥ 20 years old (81men, 16 women); ≤ 20 years old (5 men, 1 women)

For the creation of the biometric database, some experimentation tools are required such as a laptop; two external keyboards (French keyboard for users in France and Norwegian keyboard for users in Norway) i.e. AZERTY and QWERTY, respectively; and an application to collect the keystroke dynamics data. The location and position of the hardware are in a fixed position and immoveable throughout the session for the authenticity of the outcomes. According to experts, the best password is a sentence [3]. Hence, for the purpose of this study for keystroke dynamics, we present 5 passphrases as shown in Table 2, which are between 17 and 24 characters (including spaces) long, chosen from some of the well-known or popular names or artists (known both in France and Norway), denoted P_1 to P_5. We asked all of the participants to type these 5 different passphrases 20 times. We use the GREYC Keystroke software [5] to capture biometric data. A screenshot of this software is shown in Figure 1. We define two classes of the way of typing; gender category; age category denoted as C_1 and C_2, respectively as follows:

Gender: C_1 = Male; C_2 = Female.

Age: $C_1 = \geq 20$ years old; $C_2 = \leq 20$years old.

The data that we had obtained from the 5 passphrases as listed in Table 2 are keystroke dynamics data, as mentioned earlier in the article. Keystroke dynamics data consist of information containing a field of four timing values namely: the timing pressure of when the two buttons are pressed (*ppTime*); the timing release of when the two buttons are released (*rrTime*); the timing of when one button is released and the other is pressed (*rpTime*) that is the latencies between keystrokes; and the timing of when one button is pressed and the other is released (*prTime*) that is the time durations of keystrokes [5,6]. These data however, do not match exactly to the durations and latencies of keys because their ordering is based on time and not key code [5]. We use the keystroke template *vector*, which is the concatenation of the four mentioned timing values to perform our data analysis by classifying two classes for each category

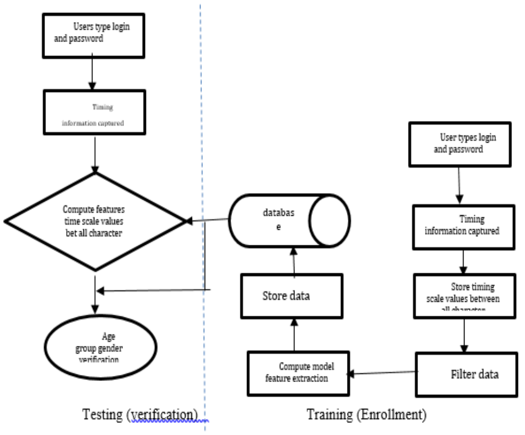
Users type login and password
Timing information captured
Compute features time scale values bet all character
Age group gender verification
databas e
Store data
Compute model feature extraction
User types login and password
Timing information captured
Store timing scale values between all character
Filter data
Testing (verification)
Training (Enrollment)

Fig.4: workflow diagram of age and gender prediction

B. Performance calculation:

The execution of a biometric framework is by and large described by the receiver operating characteristic (ROC) It can be condensed by the equal error rate (EER), the point on the bend where the false acceptance rate (FAR)and false rejection rate (FRR) are equivalents. Other framework assessment criteria incorporate productivity, flexibility, convenience, and comfort. Performance measures the execution of a biometric framework as far as procurement and recognizing error.

The end goal to assess the execution of a biometric framework, we by and large need a test benchmark and execution measurements. As per the International Organization for Standardization ISO/IEC 19795-1, the execution measurements are partitioned into three sets: Acquisition execution measurements, for example, the Failure-To-Enroll rate (FTE). Check framework execution measurements, for example, the Equal Error Rate (EER). Identification framework execution measurements, for example, the False-Negative and the False-Positive Identification Rates (FNIR and FPIR, separately).

Effectiveness: Effectiveness shows the capacity of a technique to accurately separate genuine and imposter. Execution pointers utilized by the inquires about are compressed as take after.

False Rejection Rate (FRR) refers to the rate proportion between erroneously denied honest to genuine clients against the aggregate number of authentic clients getting to the system. Once in a while known as False Nonmatching Rate (FNMR) or sort 1 error. A lower FRR infers less rejection rate and less access by genuine users. False Acceptance Rate (FAR) is characterized as the rate proportion between dishonestly acknowledged unapproved clients against the aggregate number of intruders getting to the framework. Terms, for example, False Match Rate (FMR) or sort 2 blunder alludes to a similar significance. A littler FAR shows an imposter accepted Equal Error Rate (EER) is utilized to decide the general exact accuracy and in addition a similar estimation against different frameworks. It might be here and there referred for as Crossover Error Rate (CER). Result examination depicted in the following segment will basically be express with FAR, FRR, and EER

$$FRR = \frac{Number\ of\ refused\ genuines}{Total\ no\ of\ genuines}$$

$$FAR = \frac{Number\ of\ Accepted\ imposter}{Total\ no\ of\ imposter}$$

Where P represents positive rate and N represents Negative rate, to predict accuracy =(tp+tn)/Sensitivity =tp_rate; specificity = tn_rate precision =tp/(tp+p); if f measure calculates = 2*(precision*recall)/(precision)

Table.4: performance calculation

Classification	True positive rate	False positive rate	recall	F-measure	Accuracy
Male above 20	86.1	13.8	91.1	88.5	87.23
Male below20	86.6	13.3	86.6	86.6	86.53
Female above 20	87.5	12.5	87.5	87.5	87.45
Female below 20	82.3	17.6	93.3	87.5	87.45

Classifiers Results:

We had performed several simulations with BPNN for computations on three different aspects of the data. The first results deal with the averaged (over 100 iterations) recognition rates for the four soft categories for different percentage of training data, from 1% to 90%. Then these results are completed by confidence intervals computation, based on a re-sampling and shuffling of the data.

Gender Recognition

The results of the recognition rates on different learning ratios with males (C_1) and females (C_2) for passphrases P_1 to P_5. The recognition rate, depending on the considered passphrase, is between 85to 95% for a ratio over or equal to 50%.

Age Category Recognition

The results of the recognition rates on different learning ratios age category for passphrases P_1 to P_5. The obtained recognition rate tends to vary more than for other soft categories, but stays between 85% and 90%, which are nevertheless quite good results.

IV. Conclusions and future work:

In this paper, we propose a new soft biometric approach for keystroke dynamics. It consists of predicting the user's way of typing by defining the gender, the age category whereby the 92% results were obtained. Another part of this work is the creation of a substantial database, with 118users, from France and Norway, with 100 samples per user. The obtained results could be used as a reference model to assist the biometric system to better recognize a user by a way he/she types on a keyboard. Hence, it would strengthen the authentication process by hindering an impostor trying to enter into the system. Having made a face image capture during the data collection session, we also plan to exploit the facial image capture to further enhance the performances by using a fusion method as our future work. Another work in progress consists in studying the fusion of several soft categories, to enhance the recognition.

V. REFERENCES

[1] Giot, R., El-Abed, M., Rosenberger, C." Greyc keystroke: a benchmark for keystroke dynamics biometric systems" IEEE Computer Society (2009)

[2] Giot, R., El-Abed, M., Rosenberger, C. "Keystroke dynamics overview" In Yang, D.J., ed.: Biometrics /Book 1. Volume 1. InTech 157–182, (July 2011)

[3] Junhong Kim, Haedong Kim, Pilsung Kang "Keystroke dynamics-based user authentication using freely typed text based on user-adaptive feature extraction and novelty detection" Applied soft computing, Elsevier, School of Industrial Management Engineering, Korea University, Seoul, South Korea, (2017)

[4] R. Joyce, G. Gupta, "Identity authentication based on keystroke latencies", Commun. ACM 33 ,Pg. No, 168–176. (1990)

[5] P. Kang, S. Park, S. Hwang, H.j Lee, S. Cho, "Improvement of Keystroke Data Quality Through Artificial Rhythms and Cues", Computers & Security, 27, pp. 3–11., (2008)

[6] Jain, A., Dass, S., Nandakumar, K.: "Soft biometric traits for personal recognition systems. In: Proceedings of International Conference on Biometric Authentication", (2004)

[7] Epp, C., Lippold, M., Mandryk, R." Identifying emotional states using keystroke dynamics. "In: Proceedings of the 2011 annual conference on human factors in computing systems., pg. No-715–724., (2011)

[8] Giot, R., El-Abed, M., Rosenberger, C.: Keystroke dynamics overview. In Yang, D.J., ed.: Biometrics / Book 1. Volume 1. InTech., pg.no 157–182, (July 2011)

[9] Syed-Idrus, S.Z., Cherrier, E., Rosenberger, C., Bours, P." A preliminary study of a new soft biometric: finger recognition for keystroke dynamics". In: 9th Summer School for Advanced Studies on Biometrics for Secure Authentication: Understanding Man Machine Interactions in Forensics and Security Applications. June 11-15 (2012)

[10] Nikos D. Zervas, Giorgos P. "Evaluation of Design Alternatives for the 2-D-Discrete Wavelet Transform" IEEE Transactions of circuit and systems and video Technology" volume 11, no.,12, December 2001

[11] M. Vishwanath, R. M. Owens, M. J. Irwin, "VLSI architectures for the discrete wavelet transform," IEEE Trans. Circuits Syst. II, vol. 42, pp. 305–316, May 1995.

[12] J. T. Kim et al., "Scalable VLSI architectures for lattice structure-based discrete wavelet

transform," IEEE Trans. Circuits Syst. II, vol. 46, pp. 1031–1043, Aug. 1998.

[13] C. Chakrabarti, M. Vishwanath, and R. M. Owens, "Architectures for wavelet transform: A survey," J. VLSI Signal Processing, vol. 4, pp.171–192, 1996.

[14] P. Wen-Shiaw and L. Chen-Yi, "An efficient VLSI architecture for separable 2-D discrete wavelet transform," in Proc. 1999 Int. Conf. Image Processing (ICIP99), vol. 2, pp. 754–758, Oct. 1999

[15] Syed zulkarnain, syed Idrus , Christophe Rosenberger " soft biometrics Database :A Benchmark for keystroke dynamics biometric system" conference paper "researchgate.net/publication/261040398., January 2013

[16] Romain Giot, Mohamad EL-Abed and Christophe Rosenberger, "Web-Based Benchmark for Keystroke Dynamics Biometric Systems:" A statistical Analysis", by published in the proceedings of the IIHMSP, Conferences 2012