FORM 2

THE PATENT ACT 1970
(39 OF 1970)
&
The Patent Rules, 2003

COMPLETE SPECIFICATION
(See Section 10 and Rule 13)

1 **TITLE OF THE INVENTION :**
A KEYSTROKE DYNAMICS BASED SYSTEM FOR SYNTHESIS OF IMPOSTOR DATA AND NORMALIZATION OF ENERGY LEVEL FOR USER AUTHENTICATION AND METHOD THEREOF.

2 **APPLICANT (S)**

(i) Name : **INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR.**

Nationality : An Indian Statutory Body.

Address : Sponsored Research & Industrial consultancy, Kharagpur-721302, West Bengal, India,

3 **PREAMBLE TO THE DESCRIPTION**

COMPLETE

The following specification particularly describes the invention and the manner in which it is to be performed.

FILED OF THE INVENTION:

The present invention relates to keystroke dynamics based authentication of a system user. In particular, the present invention is directed to develop a system and method for detection of behavioral biometric of the keystroke dynamics including the manner and rhythm in which an individual types characters on a keyboard or keypad of a computing device for future authentication to use the said device or the computing system.

BACKGROUND OF THE INVENTION:

The keystroke dynamics is basically includes detailed timing information that describes exactly when each key was pressed and when it was released as a person is typing at a computer keyboardor keypad of a computing device.

With the increase of technology users can transfer money, manage bank accounts, buy and sell stocks, does other online purchase etc. though their personal computing device or system such as PC, Laptop etc. These devices store huge amounts of sensitive data where unauthorized access could result in financial loss or unwanted disclosure or confidential information. User access to most computer systems is secured through possession of a username and password combination. Now to provide an inexpensive and more foolproof method than existing username and password approach for verifying the identity and the authenticity of a user, the keystroke dynamics is used.The keystroke dynamics based authentication can be integrated with existing login method to prepare a two level security system as shown in the accompanying figure 1.

Initially an enrollment phase is carried out in which the user is asked to enter his/her login ID/password for sufficient number of times. Features like hold time i.e. how long a key is pressed and latency i.e. the time between two successive keystrokes are calculated to prepare the reference model of a particular user using various machine learning classifiers which is later used for verifying the identity of the use. The process of enrolment is explained with the help of block diagram as shown in the accompanying figure 2.

Different classifiers are used for keystroke authentication and these classifiers are broadly classified under two categories namely supervised and unsupervised learning. Supervised learning requires both legal as well as impostor data whereas unsupervised learning requires only legal user data for its implementation.

Now, for collecting the impostor data, login details of the legal user is disclosed to other users who are already registered and are requested to enter the login details on behalf of the legal user. This approach of generating impostor data has two criticaldrawbacks which are stated below.

1) The login details of the user have to be disclosed to the other users already registered which compromises with the privacy of the user and makes the combination of username and password for protection altogether an illogical concept.

2) This approach of generating impostor data becomes infeasible if the number of users increases. Each time a new user is registered, his or her login details have to be entered by the rest of the users which are already registered. This process takes a lot of time because each time all the rest of the users have to be called and are requested for entering the credentials on behalf of the legal user which can also be frustrating for them.

It is found through experiments that classifiers based on supervised learning gives better results as compared to that of classifiers based on unsupervised learning [Richard O.Duda, Peter E. Hart, David G.Stork, *"Pattern Classification"*, 2nd edition]. But as stated herein above the use of supervised learning becomes impossible as collection of impostor data is an infeasible task for practical applications particularly in the computing systems where large number of user operates. Also, the typing behavior of a particular user may vary from time to time because of his or her different energy level at different times, also known as intra class variation which makes impostor data collection for a single user more difficult.

Thus, there has been need for developing a system and a method implementing keystroke based authentication technique involving supervised learning which would benefit synthesis of imposter data avoiding the limitations of the existing techniques as discussed above and can be applied in practical applications. Also importantly another need was synthesis of the imposter data independent of the user's energy level.

SUMMARY OF THE INVENTION:

Thus according to the basic aspect of the present invention there is provided a method for keystroke dynamics based user authentication for accessing a computing system or any information storage thereof comprising

detection of reference keystroke dynamics features for the user while enrolling for accessing the computing system or any information storage thereof through typing their login ID and/or password string involving a detector module with keypad of a computing system;

storing said reference keystroke dynamics features in a storage device and preparing reference model of a particular user involving machine learning classifiers based on the reference keystroke dynamics features for that user;

storing keystroke data of various users while typing free text characters on the keypad of the computing system and available by said detector module in a preferred electronic format in a storage device and involving a processing device in operative communication with the storage device to synthesize imposter data from the keystroke data for facilitating supervised learning for the keystroke dynamics based user authentication by searching equivalent transitions and movement of the hands corresponding to two keystrokes for each consecutive pair of keys in originally typed login ID and/or password string;

determining keystroke dynamics features of the user while requesting accessing said computing system or the information storage thereof through typing their login ID and/or password in the keypad involving said detector module and

comparing said keystroke dynamics features with the reference keystroke dynamics features based on the supervised learning to verify identity and authenticity of the user.

According to another aspect in the present method, the keystroke dynamics features detected by the detector module includes determination of hold time corresponding to time duration for pressing a key and latency corresponding to time duration between two successive keystrokes and determination of said hold time and latency involving selectively timer device disposed in said detector module.

According to a further aspect in the present method, the keystroke dynamics features for each user are scaled and shifted by depending upon the typing behavior of the user to make the features independent of energy level by involving

computing difference between original feature value and mean of the feature values of the entire string;

scaling and shifting the features by dividing the difference with standard deviation of the feature values of the entire string.

According to another aspect in the present method, synthesis of the imposter data from keystroke data of various users while typing free text characters on the keyboard or keypad of the computing system carried out by involving the processing device including determining positional parameter and direction of movement of hands in between two keystrokes for each consecutive pair of keys typed using single hand or two hands approach as applicable.

According to a further aspect in the present method, the determination of the positional parameter and the direction of movement of hands in between two keystrokes for synthesizing imposter data comprises

assigning two dimensional coordinates (x,y) to each of alphabetic letter keys according to their position in the keyboard;

determining the positional parameter including determining radius and angle for the each consecutive pair of keys typed to interpret direction of the movement of the hands changes in between the two keystrokes for the each consecutive pair of keys.

According to another aspect in the present method, the radius corresponds to the typed key is determined by involving

$$r = \sqrt{(x_2 - x_1)^2}$$

where r is the radius and x_2,x_1 are the assigned coordinates of two consecutive keys of the keyboards.

According to another aspect in the present method, the angle corresponds to the typed key is determined by involving

$$\theta = \tan^{-1} y/x$$

where θ is the angle and x,y are the assigned coordinates of the typed keys of the keyboards.

According to yet another aspect in the present method, similarity in transition time for movement of the hands two different pairs of keys having same radius, angle corresponds to equivalent transitions and movement of the hands for single hand typing approach.

According to yet another aspect in the present method,the keys of the keyboard are symmetrically or asymmetrically divided into two parts and assigning the keys on the left part to left hand and right part of the keyboard to the right hand for finding the equivalent transitions and movement based on sequence of hands typing the keys and variation of the transition time between pairs of keys lying in same parts and in different parts.

According to yet another aspect in the present method, the transition time between pairs of keys lying in same part is equivalent to transition time between pairs of keys having same radius, angle and lying in same part and the transition time between pairs of keys lying in same part is non-equivalent to the transition time between pairs of keys having same radius, angle and lying in different part of the keyboard.

According to a further aspect, the present method, is adapted to determine equivalent transitions and movement of the hands for two hand typing approach.

According to a further aspect in the present method, the equivalent transitions and movement of the hands is carried out by involving the processing device having time computing means for commuting the transition time between pairs of equivalent keys.

According to another aspect in the present method, the synthesizing of the imposter data comprises

collecting continuous keystroke data from various users while typing free text characters storing in an electronic format;

searching equivalent transitions in terms of radius, angle and direction of the movement of the hands in between the two keystrokes are found for each consecutive pair of keys corresponding to the string of the login ID/password using single hand or two hands approach;

assembling all possible equivalent transitions for the considered pair of key strokes to constitute user model for supervised learning.

According to another aspect in the present system there is provided a system for keystroke dynamics based user authentication for accessing a computing system or any information storage thereof involving the present method, comprising

detector module operative connected with keypad of the computing system having storage device and timer device to measure the hold time and the latency between two successive keystrokes to detect reference keystroke dynamics features for the user and storing the same;

means embodying machine learning classifiers based on the reference keystroke dynamics features for the user;

processing device for synthesizing the imposter data having time and position computing means for analyzing keystroke data of various users while typing free text characters on the keyboard or keypad of the computing system for determining equivalent transitions and movement of the hands corresponding to two keystrokes for each consecutive pair of keys in originally typed login ID and/or password string based on the positional parameters and the transition time in between the equivalent keys;

comparator means for the keystroke dynamics features of user while requesting accessing the computing system or the information storage thereof through typing their login ID and/or password in the keypad with the reference keystroke dynamics features for verifying authenticity of that user.

BRIEF DESCRIPTION OF THE ACCOMPANYING FIGURES:

Figure 1shows flow chart for a two level security system.

Figure 2 shows block diagram for enrollment of a user.

Figure 2(a) shows a Preprocessing and training for enrollment of a user in accordance with the present invention.

Figure 2(b) shows Block diagram of the authentication phase of continuous authentication system in accordance with the present invention.

Figure 3 shows the format of data collected in accordance with the present invention.

Figure 4 shows an arrangement of keys along with their coordinates.

Figure 5 shows the block diagram for impostor data synthesis.

Figure 6 shows comparison of results obtained for actual impostor data and impostor data synthesis.

Figure 7 shows a snapshot of data collection module (enrollment) in a designed prototype in accordance with the present invention.

Figure 8 shows a snapshot of testing module in the designed prototype in accordance with the present invention.

DESCRIPTION OF THE INVENTION WITH RESPECT TO THE ACCOMPANYING FIGURES:

The present invention discloses keystroke dynamics based user authentication for accessing a computing system or any information storage thereof. The keystroke dynamics based user authentication basically starts with detection of reference keystroke dynamics features for the user while he/she enrolled for accessing the computing system or any information storage thereof through typing their login ID and/or password string in keypad of the computing system. The reference model of a particular user is prepared involving machine learning classifiers based on the reference keystroke dynamics features for that user. Simultaneously imposter data is synthesized to facilitate the supervised learning for the keystroke dynamics based user authentication based on equivalent transitions and/or movement of the hands corresponding to two keystrokes for each consecutive pair of keys in typed login ID and/or password string. The equivalent transitions and/or movement are collected fromthe keystroke data of various users while typing free text characters on the keyboard or keypad of the computing system. The identity and authenticity of the user is then verified while he/she requesting accessing the computing system or the information storage thereof through typing their login ID and/or password in the keypad based on the supervised learning and comparing with the reference keystroke dynamics features. The enrolment process and the authentication phase of the present invention is shown in the accompanying figure 2(a) and 2(b).

In the present invention detection of behavioral biometric of keystroke dynamics data including the manner and rhythm in which an individual types text characters on a keyboard or keypad of a device or computing, is based on the keystroke dynamics of various users while typing free text data. For this a key logger executable program programmed such as in C sharp language is installed in the computers. It runs in the background while the person/users can do his normal work without any interruption. It captures the events such as key pressed, key released and timings of the corresponding events and is written in a

database in a preferred electronic format like notepad file. For maintaining the security of the data the name of the keys are not written directly in the original form, it is first encoded and is then written in the encoded form as shown in the accompanying figure 3.

Two dimensional coordinates i.e. the x and y coordinates are assigned to each of the alphabetic letter keys according to their position in the keyboard. It is considered as a 2-D array, each row and column assigned with an index according to the position. For e.g. the bottom row has an index 0 and is incremented as it moved up. Similarly each column is assigned with an index according to its positions; it is specifically scaled so as to avoid decimal coordinates. For avoiding decimal coordinates the spacing between two consecutive keys in the same row is kept as same. As a demonstration, an expanded view of generic keyboard along with the coordinates assigned to it is shown in the figure 4.

With the help of the assigned coordinates, positional parameter between any pair of keys is determined. Said positional parameter includes radius and angle can be found out between any pair of keys using simple mathematical formulae.

$$r = \sqrt{(x_2 - x_1)^2}$$

$$\theta = \tan^{-1} y/x$$

Now for a particular user, two different pairs of keys having same radius, angle and direction of movement of hands between the two key strokes has same variation in timing of the keystrokes and both pair of key strokes can be considered as equivalent. For example, variation of the timings in transition from E to F can be considered equivalent to variation of the timings in transition from R to G, T to H and similarly with other pairs of keys. Also the transition from E to F (for e.g.) is considered different from transition from F to E as the direction of the movement of the hands changes in between the two keys. However this is valid only if the user is made to type with a single hand and is known as Single

Hand Approach.

For a normal typing process i.e. a person typing with both his hands, the variation of timings between different pairs of keys may come out to be different due to the change of the hands in between the two keys. It is assumed here that if two keys are pressed with different hands then the time between the two events is very small because as soon as the user presses one key the other key will be pressed within a small time gap as the other hand of the user will almost be near to the second key when he is typing the first key with a different hand. Whereas if a pair of keys are typed by the same hand then the time between the two events is more as compared to that of the previous case as the same hand takes more time to go from one key to the other. This time depends upon the separation between the two keys which is represented in the form of radius and angle between the two keys.

For a normal typing process, another dimension is added to each of the keys in addition to x and y coordinates. The keys of the entire keyboard are symmetrically divided into two parts. The keys on the left part are assigned a position as left hand which assumes that user types these keys using his or her left hand and similarly the right part of the keyboard is assigned a position of right hand. For finding an equivalent pair, apart from same radius, angle and direction of the movement of the hands, the sequence of hands typing the keys should also be same as that of the original pair. For example, if a person types a pair of keys which falls in the different parts of the keyboard then the equivalent pair of keys should also be in different parts of the keyboard. For example, variation of the timings in transition from E to F can be considered equivalent to variation of the timings in transition from R to G but cannot be considered equivalent to transition from T to H because typing TH involves a change of hand. This approach is known as two hands Approach.

Procedure for Impostor data synthesis

In the present work 'electronics' is used as a password but the approach is valid for any typed string.

Step1. Free text or continuous keystroke data is collected from various persons for some days irrespective of the fact whether they are registered users or not. The procedure for this was already discussed earlier and the format of the data collected is shown in the figure 5.

Step2. New users are made to enroll and transition between each consecutive pair of key strokes in the typed password or string (electronics) is mapped to radius, angle and direction of the movement of the hands between the two key strokes.

Step3. An exhaustive search is made in each of the free text data file and equivalent transitions in terms of radius, angle and direction of the movement of the hands in between the two keystrokes are found for each consecutive pair of keys in the typed string using the Single hand approach or Two hands approach (discussed previously) as applicable. The timings of these equivalent pairs are written in a notepad file.

Step4: The entire file is searched so as to get as many equivalent transitions as possible for the considered pair of key strokes. If enough keystroke transitions are not available then the user model cannot be prepared.

Step5. Features like hold time and latency are calculated using the keystroke timings written in the file. After calculation of features, feature vectors are formed and various classifiers based on supervised learning can be used for preparing the user model. The system is ready for verifying the identity of the user and can be tested by making the users to login into their account.

Some of the results captured are shown with the help of a graph in figure 6. It can be observed that the supervised learning approach gives better results as compared to that of unsupervised learning approach. Supervised learning approach is tested with two techniques, with actual impostor data which is collected by disclosing the password of the legal user to the rest of the registered users in the system and with a new devised technique 'Impostor Data Synthesis'. The graph shows that the former technique gives slightly better performance as compared to the latter, however as discussed previously it is completely an

infeasible technique and can be only used for experimental purpose but the new technique gives approximately the same performance as compared to that of the actual impostor data using which supervised learning approach can be applied in practical applications.

Prototype Implementation

As discussed previously, continuous keystroke data is collected from some persons beforehand irrespective of the fact whether they are registered users or not. This data will be used for impostor data generation using the proposed technique 'Impostor Data Synthesis'. Now there is no need to disclose the login information of the legal user to the other users and making them type on behalf of the legal user. Whenever a new user is registered, enrollment technique is carried out wherein the user is entered his/her login ID/password for sufficient number of times. Features like hold time i.e. how long a key is pressed and latency i.e. the time between two successive keystrokes are calculated to prepare the reference model of a particular user using machine learning classifiers which is later used for verifying the identity of the use.

As seen from figure 7 user has to enter his first name as username, his roll and password which was taken as "electronics" for the present experiment. If any wrong details are emceed it can be corrected by clearing that specific field using a clear button provided for each field. Before pressing the register button, a user can edit the fields as many times he or she wants. There are two ways in which users can be made to enter, one is by making the user type with a single hand (Single Hand Approach) and the other is the normal typing process i.e. with both hands (Two Hands Approach). Once the user is made to enter his or her login derails for a sufficient number of times, a background process generates impostor data for the password string (it can be extended to all the strings typed i.e. username, roll number and password but in the present implementation only password string is used) using the continuous keystroke data collected and the corresponding model or reference signature is prepared and saved after which the system is ready for authentication which can be clone using the testing module.

In the testing phase the user has to enter his or her login details and on pressing the register button the system will decide whether to allow the access or not as shown in figure 8.

Making Feature vectors independent of the energy level

The features formed for each user are scaled and shifted by a certain amount depending upon the typing behavior of the user. The scaled and shifted feature vectors when tested with different classifiers gave better results as compared to that of unmodified feature vectors. One of the techniques for scaling and shifting is shown below:

New Feature value = (Original Feature Value - Mean of the feature values of the entire string)/ (standard deviation of the feature values of the entire string)

After testing the classifier with the above technique, rise in the overall accuracy was observed as compared to the testing performed with original feature vectors.

Advantages:

In the present work, with the ease in the collection of impostor data, classifiers based on supervised learning can be easily used which gives a high accuracy as compared to classifiers based on unsupervised learning.

Also features are made independent of the energy level of the user. With these two new techniques implemented, the overall accuracy of the system gets improved which makes the system more robust for practical applications.

WE CLAIM:

1. Amethod for keystroke dynamics based user authentication for accessing a computing system or any information storage thereof comprising

detection of referencekeystroke dynamics features for the user while enrolling foraccessing the computing system or any information storage thereof through typing their login ID and/or password string involving a detector module with keypad of a computing system;

storing said reference keystroke dynamics features in a storage device and preparing reference model of a particular user involving machine learning classifiers based on the reference keystroke dynamics features for that user;

storing keystroke data of various users while typing free text characters on the keypad of the computing system and available by said detector module in a preferred electronic format in a storage device and involving a processing device in operative communication with the storage device to synthesize imposter data from the keystroke data for facilitating supervised learning for the keystroke dynamics based user authentication by searching equivalent transitions and movement of the hands corresponding to two keystrokes for each consecutive pair of keys in originally typed login ID and/or password string;

determining keystroke dynamics features of the user while requesting accessing said computing system or the information storage thereof through typing their login ID and/or password in the keypadinvolving said detector module and

comparing said keystroke dynamics featureswith the reference keystroke dynamics features based on the supervised learning to verify identity and authenticity of the user.

2. The method as claimed in claim 1, wherein the keystroke dynamics features detected by the detector module includes determination of hold time corresponding to time duration for pressing a key and latency corresponding to

time duration between two successive keystroke sand determination of said hold time and latency involving selectively timer device disposed in said detector module.

3. The method as claimed in anyone of the claim 1 or 2, wherein the keystroke dynamics features for each user are scaled and shifted by depending upon the typing behavior of the userto make the features independent of energy level by involving

computing difference between original feature value and mean of the feature values of the entire string;

scaling and shifting the features by dividing the difference with standard deviation of the feature values of the entire string.

4. The method as claimed in anyone of the claims 1 to 3, wherein synthesis of the imposter data from keystroke data of various users while typing free text characters on the keyboard or keypad of the computing system carried out by involving the processing device including determining positional parameter and direction of movement of hands in between two keystrokes for each consecutive pair of keys typed using single hand or two hands approach as applicable.

5. The method as claimed in anyone of claims 1 to 4, wherein the determination of the positional parameter and the direction of movement of hands in between two keystrokes for synthesizing imposter data comprises

assigning two dimensional coordinates (x,y) to each of alphabetic letter keys according to their position in the keyboard;

determining the positional parameter including determining radius and angle for the each consecutive pair of keys by involving the assigned coordinates to

interpret direction of the movement of the hands changes in between the two keystrokes for the each consecutive pair of keys.

6. The method as claimed in claim 5, wherein the radius corresponds to the typed key is determined by involving

$$r = \sqrt{(x_2 - x_1)^2}$$

where r is the radius and x2,x1 are the assigned coordinates of two consecutive keys of the keyboards.

7. The method as claimed in claim 5, wherein the angle corresponds to the typed key is determined by involving

$$\theta = \tan^{-1} y/x$$

where θ is the angle and x,y are the assigned coordinates of the typed keys of the keyboards.

8. The method as claimed in anyone of the claims 1 to 7, wherein similarity in transition time for movement of the hands two different pairs of keys having same radius, angle corresponds to equivalent transitions and movement of the hands for single hand typing approach.

9. The method as claimed in anyone of the claims 1 to 8, wherein keys of the keyboard are symmetrically or asymmetrically divided into two parts and assigning the keys on the left part to left hand and right part of the keyboard to the right handfor finding the equivalent transitions and movement based on sequence of hands typing the keys and variation of the transition time between pairs of keys lying in same parts and in different parts.

10. The method as claimed claim 9, whereinthe transition time betweenpairs of keys lying in same part is equivalent to transition time between pairs of keys having same radius, angle and lying in same part and the transition time between pairs of keys lying in same part is non-equivalent to the transition time between pairs of keys having same radius, angle and lying in different part of the keyboard.

11. The method as claimed in anyone of the claims 9 or 10, is adapted to determined equivalent transitions and movement of the hands for two hand typing approach.

12. The method as claimed in anyone of the claims 8 to 11, wherein the equivalent transitions and movement of the hands is carried out by involving the processing device having time computing means for commuting the transition time between pairs of equivalent keys.

13. The method as claimed in anyone of the claims 1 to 12, wherein the synthesizing of the imposter data comprises

collecting continuous keystroke data from various users while typing free text characters storing in an electronic format;

searching equivalent transitions in terms of radius, angle and direction of the movement of the hands in between the two keystrokes are found for each consecutive pair of keys corresponding to the string ofthe login ID/password using single hand or two hands approach;

assembling all possible equivalent transitions for the considered pair of key strokes to constitute user model for supervised learning.

14. A system for keystroke dynamics based user authentication for accessing a computing system or any information storage thereof involving the method as claimed in anyone of the claims 1 to 13, comprising

detector module operative connected with keypad of the computing systemhaving storage device and timer device to measure the hold time and the latency between two successive keystrokes to detect reference keystroke dynamics features for the user and storing the same;

means embodying machine learning classifiers based on the reference keystroke dynamics features for the user;

processing device for synthesizing theimposter data having time and position computing means for analyzing keystroke data of various users while typing free text characters on the keyboard or keypad of the computing system for determining equivalent transitions and movement of the hands corresponding to two keystrokes for each consecutive pair of keys in originally typed login ID and/or password string based on the positional parameters and the transition time in between the equivalent keys;

comparator means for the keystroke dynamics features of user while requesting accessing the computing system or the information storage thereof through typing their login ID and/or password in the keypad with the reference keystroke dynamics features for verifying authenticity of that user.

Dated this the 7th Day of August, 2015

Anjan Sen
Of Anjan Sen and Associates
(Applicants Agent)
IN/PA-199

ABSTRACT

Title: A KEYSTROKE DYNAMICS BASED SYSTEM FOR SYNTHESIS OF IMPOSTOR DATA AND NORMALIZATION OF ENERGY LEVEL FOR USER AUTHENTICATION AND METHOD THEREOF.

The present invention disclosesa method for keystroke dynamics based user authentication for accessing a computing system or any information storage thereof. The method comprisingdetection of reference keystroke dynamics features for the user while enrolling for accessing the computing system or any information storage thereof through typing their login ID and/or password string, preparing a reference model of a particular user based on the reference keystroke dynamics features for that user, synthesize imposter data from keystroke data of various users while typing free text characters on the keypad of the computing system for facilitating supervised learning, determining keystroke dynamics features of the user while requesting accessing said computing system or the information storage thereof through typing their login ID and/or password in the keypad andcomparing said keystroke dynamics features with the reference keystroke dynamics features based on the supervised learning to verify identity and authenticity of the user.