ISSN: 0067-2904
GIF: 0.851

Keystroke Dynamics Authentication based on Naïve Bayes Classifier

Mays M. Hoobi*
Computer Science Department, College of Science, Baghdad University, Baghdad, Iraq.

Abstract

Authentication is the process of determining whether someone or something is, in fact, who or what it is declared to be. As the dependence upon computers and computer networks grows, the need for user authentication has increased. User's claimed identity can be verified by one of several methods. One of the most popular of these methods is represented by (something user know), such as password or Personal Identification Number (PIN). Biometrics is the science and technology of authentication by identifying the living individual's physiological or behavioral attributes. Keystroke authentication is a new behavioral access control system to identify legitimate users via their typing behavior. The objective of this paper is to provide user authentication based on keystroke dynamic in order to avoid un authorized user access to the system. Naive Bayes Classifier (NBC) is applied for keystroke authentication using unigraph and diagraph keystroke features. The unigraph Dwell Time (DT), diagraph Down-Down Time (DDT) features, and combination of (DT and DDT) are used. The results show that the combination of features (DT and DDT) produces better results with low error rate as compared with using DT or DDT alone.

Keywords: Biometric, Down-Down Time, Dwell Time, Keystroke Dynamics, Naïve Bayes, User Authentication.

مصادقة المستخدم باستخدام مصنف نيف بايز استنادا على ديناميكية ضغطة المفتاح

ميس محمد هوبي

قسم علوم الحاسبات، كلية العلوم، جامعة بغداد، بغداد ، العراق.

* Email:-Mais.shms@yahoo.com

الخلاصة:

المصادقه هي عملية تحديد ما إذا كان شخص في الواقع هو كما أدعى أن يكون. بما أن الاعتماد على أجهزة الكمبيوتر وشبكات الكمبيوتر قد ازداد هذا ما ادى الى ازدياد الحاجة للمصادقة حيث ان الهوية التي يدعيها المستخدم ممكن اثباتها بعدة طرق شائعه ، واحدة من أكثرهذه الطرق شعبية يمثله (شيء يعرفه المستخدم)، مثل معرفة كلمة السر أو رقم التعريف الشخصي.المقاييس الحيوية هي العلم والتكنولوجيا للمصادقه من خلال تحديد السمات الفسيولوجية أو السلوكية للفرد وان المصادقة من خلال تميز ضغطة المفاتيح هو نظام جديد لمراقبة سلوكية الدخول لتحديد المستخدمين الشرعيين عبر كتابة كلمة السر. في هذا البحث يتم تطبيق مصنف نيف بايز للمصادقة واثبات هوية المستخدم والخصائص المستخدمه هي وقت ضغطة المفتاح كخاصيه احاديه والخاصيه الثنائيه وقت ضغطة المفتاح الى وقت ضغطة المفتاح الذي يليه ومره اخرى بالدمج بين الخاصيه الاحاديه والخاصيه الثنائيه وبينت النتائج بأن الحاله الاخيره توصلت الى نتائج افضل وبأقل خطأ ممكن.

Introduction

Authentication is the process of verifying whether the digital identities of computers and the physical identities of people are authentic. There are multiple authentication technologies that verify the identity of a user before granting access to system resources. However, these technologies provide different levels of security, and none can be said to secure a system completely [1].

Biometric refers to technologies that measureand analyzethephysiological and/ or behavioralcharacteristicsofa humanforverificationoridentification. The biometric is very necessarily for robust, reliable, and foolproof personal authentication systems[2]. Biometric systems use either a person's physical characteristics (like fingerprints, irises or veins), or behavioral characteristics (like voice, signature or keystroke). Biometric data are highly unique to each individual, easily obtainable non-intrusively, time-invariant (no significant changes over a period of time) and distinguishable by humans without much special training[3].

Keystroke dynamics isa behavioral measurement and it utilizes the manner and rhythm in which each individual types. In addition keystroke dynamics is widely accepted biometric method for authentication ratherthanotherkindsofbiometric methods as it is a two factor biometric security system authentication on correctness of password and correctness of typing pattern [4]. Furthermore, keystrokedynamics features canbeusedinconjunction withothermechanisms,suchas generatinga"hardenpassword" [5]. There are two types of keystroke dynamics, the first one is based onanalysis performed on typing samples produced using predetermined text for all the individuals under observation.While the second type is dynamic analysis which implies a continuous or periodic monitoring of issued keystrokes.It is performed during the log-in session and continues after the session[6].

This paper is organized as follows: Section 2 discusses the simple overview of related work on keystroke dynamics, while section3 explains the most popular keystroke dynamic features.Section4 illustrates the description of NCB algorithm. Section 5 describes the proposed system including data collection, feature extraction, and user authentication based on NBC algorithm. Finally, Sections 6 and sections 7 discuss the experimental results and conclusion respectively.

Related Work

Over the last years, researchers have evaluated different features and classification methods in an effort to improve the classification capabilities of keystroke biometrics. Some of these features and classification are explained briefly in this section :-

In [7], a comparison between ADALINE (based on the single perceptron model) and the BPNN (Back Propagation Neural Network) model, using both the latency time and digraph latency time is presented.The work concluded that BPNN surpasses the ADALINE which. In [8], Probabilistic Neural Network (PNN) was presented for the strengthening of password security by employing the biometric feature of keystroke dynamics DT. The PNN was performed well and it was more suited to keystroke dynamic application than traditional BP model. In [9], a statistical-based comprehensive study is carried out keystroke dynamics-based user authentication system using neural network. The workconcluds that neural network-based methods gave better results as compared with statistical methods in keystroke patterns classification using Flight Time (FT)feature . In [10], a neural network was used to classify legitimate user from attackers. The proposed virtual key force feature was used and 43 users participated in the experiment. The conclusion of this work is a new feature (virtual key force) has been reduces the training and testing.

Keystroke Dynamic Features

Keystroke dynamics is a behavioral measurement aims to identify users based on the typing of the individuals or attributes. It provides an answer to the authentication and security problem. The principle behind keystroke dynamics is to extract and analyze the way an individual types as opposed to only what the individual types. Each person may have different styles to press the key because the typing style is based on user’s experience and individual skill which is difficult to imitate [11]. Number of different features of keystroke dynamics can be used for authentication. Some of these features are related to event of one character and called unigraph feature. Example is duration of a keystroke or Dwell Time (DT) which is the time interval where that key remained pressed.In another case the keystroke features are related to event of two keys, called diagraph feature. Examples are Down-Down time (DDT) and Up-Up time(UUT). Also there are other types of keystroke features such as typing error, force of keystrokes, Rate of typing, statistics of text etc [12].

Naïve Bayes Classifier (NBC)

TheBayesianClassificationrepresentsasupervised probabilistic learningmethodaswellasa statistical methodforclassification. Also itprovidesa usefulperspectivefor understandingandevaluatingmanylearningalgorithms.It calculatesexplicitprobabilitiesfor hypothesisanditisrobusttonoiseininputdata.There arethreewellknowncategories in classification methods,statisticallearning,rulebasedlearningandtreebased learning.A Naïve Bayes Classifier(NBC) is a simplprobabilisticstatistical classifier based on applyingBayes probability theorem[13,14]. Bayes theorem can be descried as follow [15]:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \qquad (1)$$

Where,

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \quad (2)$$

TheprobabilitiesofaneventAmaywelldepend onthepreviousorsimultaneousoccurrenceofanevent B . A issaidtobeconditioned on B .Theneedtoincorporatethis typeof dependenceintothetheoryresultedinthedefinition oftheconditionalprobability $P(A\mid B)$,whichisthe probabilitythat A willoccurgiventhat B alreadyhas. ThebasicideaofBayesruleisthattheoutcome ofan event A canbepredictedbasedonsomeevidences(x)that canbeobserved.TheBayesrulehas,

i) **Prioriprobability:**theprobabilityofanevent beforetheevidenceisobserved.

ii) **Posteriorprobability:**theprobabilityofan eventaftertheevidenceisobserved.

Informally,Bayesrulesays:

$$Posterior=\frac{Likelihoodprior}{Eivdence} \qquad (3)$$

$P(A)$ iscalledthepriorprobabilityof A i.e.beforehaving thedataorevidence x .Theterm $P(X\mid A)$ iscalledthe likelihood(probability densityfunction) and $P(A\mid X)$ is calledtheposteriorprobabilityi.e.afterhavingtheevidence. Theconditionalprobabilityisobtainedby [12]:

$$p(A\mid X)=\frac{P(X\mid A)P(A)}{P(X)} \qquad (4)$$

NBCalgorithmdescribeshowtheclassifierrecognizesthe pattern. Thedefinedsystem hasnumberofclasseseachof whichcontainshugeamountoftrainingsamples. The probability fora testpatternandeachclassiscalculatedto identify whichclass hashigher probabilityconditioned with inputfeatureprobability. A classwithhigherprobability isselectedasanexpected class which contains pattern related to the test pattern. By using training sets, probability density function is calculated for both test pattern and template. These probability density functions are used by bayes theorem to find conditioned probability. By comparing those probabilities, the bayes rule can find whether the new unclassified pattern is matched to template pattern[7]. In general NBC algorithm can be described as show in algorithm(1): [15]

Algorithm (1)

Step1: Establishatrainingset {x_j ,c_j },j=1, 2….. Nforeach**class**،

Where x_j number oftraining samples and c_j number ofclasses .

Step2:Compute a priori information such as probabilities for each template vector and probability density function $p(x\mid c_i)$ as show in equation (5).

$$p(x\mid c_i)=\frac{P(x\mid c_i)P(c_i)}{P(x)} \qquad (5)$$

Step3:Givenanewunclassifiedmeasurementy,useBayes theorem to obtain the measurement conditioned probability as show in equation (5).

$$P(c_i\mid Y)=\frac{P(Y\mid c_i)P(c_i)}{P(Y)} \qquad (6)$$

Step4:Choosec$_i$suchthat$P(c_i\mid Y)>P(c_j\mid Y)$ foralli≠j.

NBC for User Authentication

The proposed approachnamed as User Authentication Keystroke (UAK) focuseson static keystroke authentication. This system was, thisapproach mainly consists of three phases as illustrated in figure-1: First,data collection phase where a user registers or enrolls his/her timing vector patterns . Second, a feature extraction phase. Third, a NBC was performed by using timing patterns, either accept or reject user based on timing vector.Finally UAKS uses NBC algorithm to distinguish between legitimate and impostor user. The following subsections present each phase of UAKS.

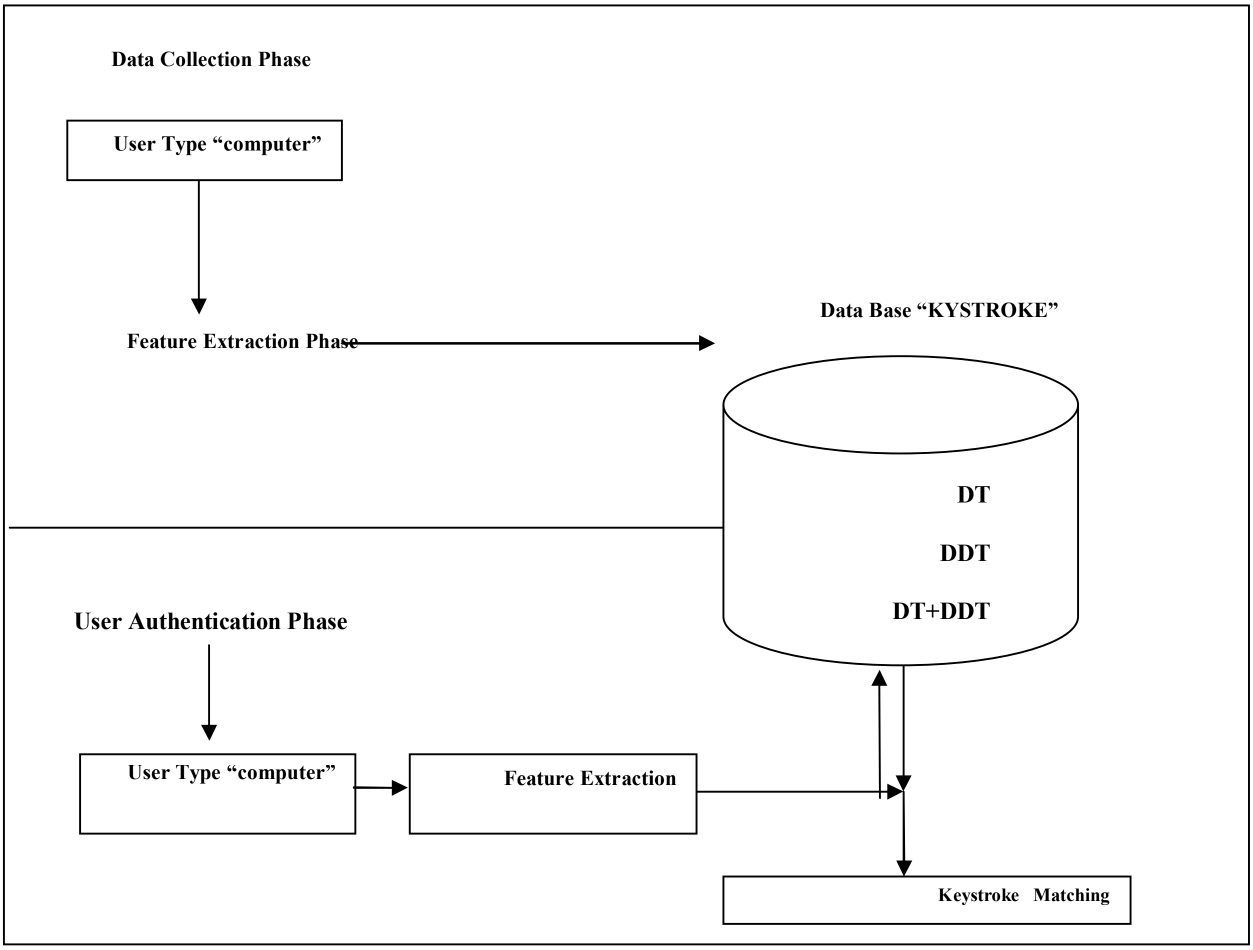

Figure 1- User Authentication Keystroke (UAK)

Data Collection Phase

Datacollectionis the first and acriticalstep in UAKS performance.Duetothehumanethicsissue,thedataused byotherresearchersarenotavailableforsharing. The data was collected from students and staff of University of Baghdad/College of Science with, total of 425 users participated in data collection phase. The participated users are divided into two classes: first the legtimate user class which contains 150 users, while the second one is impostor user class which contains 275 users. At the begning, each participant had to register the determined password “computer” during a login session and store in database named as (KYSTROKE). All participants were requested to enter the same password. Thus, a database of 425 user profiles is created. Each profile containing a sample of keystroke features (timing vector) measured in milliseconds.

Feature Extraction Phase

The feature extraction phase is used to distingushattributes common to all patterns belonging to a class. Complete set of discriminatory features for each pattern class can be found using feature extraction. As mentioned in the previous phase, the timeinformationwas collected andstored asrawdata in (KYSTROKE) databaseusedforlaterprocess. Thefeature extraction program was developed tocapturethreetypesofkeystroke features (timing information) fromusers'typingbehavior. In UAK, DT feature, DDT feature and combination of them are extracted as show in figure -2.

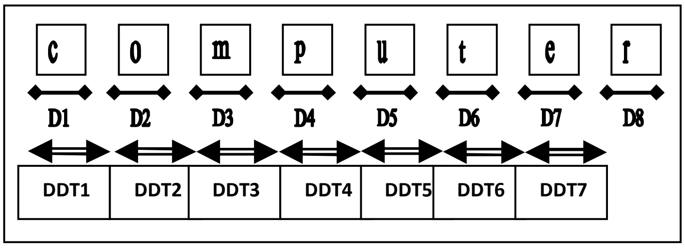

Figure 2- Features Representation of "computer" Password

The length of the timing vector is different and depending on the length of the password and the type of feature used. For example, a password "computer" which contains eight characters will result in eight DT, seven DDT and so on. Generally, a password with n character will yield n number of DT and n -1 number for DDT, and (n+(n-1)) for DT+DDT as illustrated in table(1).

Table1- Length of Timing Information of "computer" Password

Feature Name	DT	DDT	DT+DDT
Length of timing vector	**8**	**7**	**15**

NBC

NBC consists of two stages, training and classification. Trainingis theprocessoflearningamodel. After the extraction of timing vectors, the training process is performed to calculate the conditional probability $P(c_i \mid x)$ which requires mean (μi), variance (σ), and probability density function $P(x \mid c_i)$ as illustrated in equation(7) [18].

$$p(x \mid c_i) = \frac{1}{\sqrt{2\pi\sigma}} \exp\{\frac{-1}{2}(\frac{x-\mu_i}{\sigma})^2\} \qquad (7)$$

NBC can determine the mean (μi), variance (σ) and probability density function $P(x \mid c_i)$ by using training set at training process. These steps are to be repeated for all classes and choose which gives the highest probability. With the problem of keystroke dynamic classification, number of training samples together with the corresponding correct class for each sample are known. NBC classifies the typing style of users by comparing probability of input vector $p(x)$ with associated class probability $P(c_i)$. Theprobabilitydensityfunctionforfeaturevector $P(x \mid c_i)$ canbecalculated onlyfrom thetraining set.These probability densityfunctionshavesamevarianceσwith differentmeansμ1andμ2anditcanbeestablishedbyusing equation (7).In proposed system thereare twoclasses,thentheBayestheorem showsthatif $P(c_1 \mid x) > P(c_2 \mid x)$ implies,

$$\frac{P(x \mid c_1)P(c_1)}{P(x)} \rangle \frac{P(x \mid c_2)P(c_2)}{P(x)} \qquad (8)$$

so usingthe fact that $P(c_{1_i}) = P(c_2)$ yields,

$$P(c_1 \mid x) \rangle P(c_2 \mid x) \Rightarrow P(x \mid c_1) \rangle P(x \mid c_2) \quad (9)$$

andthisallowsa decisionrule,Choose c1if $P(x \mid c_i) > P(x \mid c_{i2})$ otherwise choose c2.

Experimental Results

This section demonstrates the results of UAKS with NBC and evaluates the efficiency of the extracted features DT, DDT and DT+DDT. All biometrics are measured under certain criteria, these are known as biometric performance measures. To evaluate the performance of the proposed keystroke dynamic system (UAKS) three types of these measures are used as illustrated [16;17]:-

1-**Detection Rate (DR):** is the rate, at which users are correctly classified as impostors when they should so.

2- False Alarm Rate (FAR): is the percentage of legtimate users incorrectly categorized as imposters.

3-Accuracy (Acc):which is the proportion of true results in the population.

In this paper the NBC isappliedintwostages:training and classification. In the training stage the algorithm was trained on 425 samples divided into 150 authenticated samples and 275 imposter samples. While in testing phase two experiments are performed. In the first experiment the NBC was tested on the samesamplesthatitwastrained on.While inthesecond experiment51samples was chosenrandomlyas 18(from150authenticusers)and33samples (from275 imposter users).The testingtrails and training trails are not the same. Tables (2 – 3)illustrate the results applying NBC with DT , DDT, and DT+ DDT when each user types “computer” password. The results show that (DT+ DDT) satisfied the higher DR and Acc compared with e other features.

Table 2- Results of Experiment1

Feature (s) Name	DR%	FAR%	Acc%
DT	90.7	0.6	76
DDT	82.5	0.3	77.9
DT+DDT	95	0.26	84.4

Table 3- Results of Experiment2

Feature (s) Name	DR%	FAR%	Acc%
DT	90	0.5	76.4
DDT	90.9	0.16	88.2
DT+DDT	90.9	0	94.1

For more illustration the comparison of the results of two experiments are demonstrated in figures-(3-4) for each case of used features.

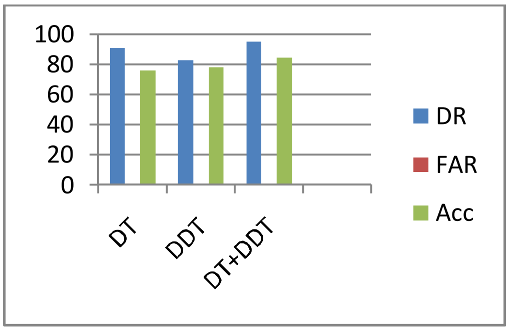

Figure 3-DR, FAR, and Acc of ExtractedFeatures in Experiment1

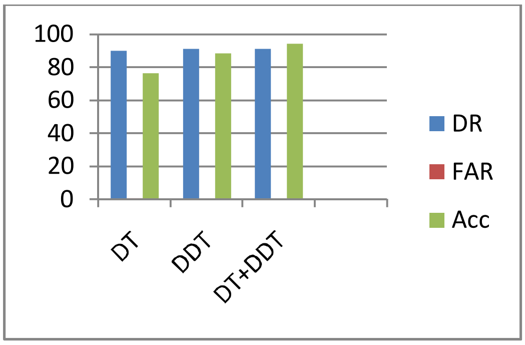

Figure 4-DR, FAR, and Acc of ExtractedFeatures in Experiment2

Conclusion

In this paper an authentication method with keystroke features was presented to improve the password user authentication method by applying NBC. The experimental results illustrate that when DT feature is used as unigraph feature the proposed system satisfied lowest DR and Acc. In other case when DDT feature (diagraph) is used, satisfied better results.Finally the best results (highest DR and Acc) are satisfied using (combine DT with DDT) features.

References

1. Joyce, R. and Gupta, G. **1990**. Identity authentication based on keystroke latencies, *Communications of the ACM 33 (2) 168–176.*

2.Matyas, S.M. ,andStapleton, J. **2000** .Abiometricstandardfor informationmanagement and security,*Computers&Security19(n. 2)428–441.*

3.Lawrence, O.Veridicom,I.Chat,m. 1998. Overviewof fingerprint verificationtechnologies,*(ElsevierInformationSecurity TechnicalReport*,Vol. 3,No.1).

4.Monrose,F.Reiter,M.K.andWetzel, S.**1999**. Passwordhardeningbased onkeystrokedynamics.In*Proceedingsofthe6thACM conferenceon Computerandcommunicationssecurity,volume 6,pages73–82.ACM Press*.

5.Obaidat ,M.S.andSadoun, B.**1997**. Verificationofcomputerusersusing keystroke dynamics. In *IEEETransaction on Systems, Man and Cybernetics,partB,volume27,pages261–269.*

6. Robinson, J. and Liang,V. and Chambers, J. and MacKenzie, C.**1998**. Computer user verification using login string keystroke dynamics.*In IEEE Transaction on Systems, Man and Cybernetics, Part A, volume 28, pages 236–241.*

7. Abdullah, N. Ahmad, A.M. **2000**. "User Authentication via Neural Networks", in Proceedings of the 9th International Conference on Artificial Intelligence: Methodology, Systems, and Applications, pp. 310-320.

8. Richard,S.**2003.** "A New Approach to Securing Passwords Using a Probabilistic Neural Network Based on Biometric Keystroke Dynamics", PhD thesis, University of Newcastle upon Tyne, UK, The Department of Electrical and Electronic Engineering.

9. Change, L.and Kin, C L. and C.Peng, L.**2007**. "Keystroke Patterns Classification Using the ARTMAP-FD Neural Network", In proceeding of 3rd International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIHMSP, Vol. 2, pp. 61-64.

10.Shanmugapriya, D. and Padmavathi,G.**2011.** "Virtual Key Force a New Feature for Keystroke Dynamics", *International Journal of Engineering Science and technology, Vol. 13, No. 10, pp. 7738-7743.*

11. Peacock, AK. and Wilkerson,M.**2004**. Typing patterns: a key to user identification, *IEEE Security and Privacy 2 (2) 40–47.*

12.Hataichanok,S. and Bhatarakosol, P.**2008**. user authentication using combination of behavioral biometrics over the touchpad acting like touch screen of mobile device. Proceedings of the International Conference on Computer and Electrical Engineering, Phuket, Thailand.

13.Balagani, K. and Phoha,SV. and Ray, VA. andPhoha, S.**2011**. *"On the Discriminability of Keystroke Feature Vectors Used in Fixed Text Keystroke Authentication, "*Pattern Recognition Letters.

14. Leung, K. C. and . Leung,C. H.**2011**. "Improvement of Fingerprint Retrieval by a Statistical Classifier", *IEEE Transactions on Information Forensics And Security,* Vol. 6, No. 1, Pp 59 -69.

15.LudmilaKuncheva, I. **2003**.Combining Pattern Classifiers Methods and Algorithms, Bangor, Gwynedd, United Kingdom.

16.Webopedia.com. FalseAcceptance, **2011**.http://www.webopedia.com/TERM/F/false_acceptance.html, accessed Jan. 2010.

17. Jain, AL. and SharathPankanti, H.**2003**. Biometric identification systems, *signal pro- cessing, ACM 83 (12) ,2539–2557.*