2009 International Conference on Computer Engineering and Applications
IPCSIT vol.2 (2011) © (2011) IACSIT Press, Singapore

Efficiency in the Design of a Biometric Identification System Based on Electrocardiogram Data

Mouhcine Guennoun [1], Zine E.A Guennoun [2] and Khalil El-Khatib [1]

[1] University of Ontario Institute of Technology
2000 Simcoe Street North, Oshawa, Ontario, Canada L1H 7K4
[2] Département Math-Info, Faculté des Sciences de Rabat
4 Avenue Ibn Battouta B.P. 1014 RP, Rabat, Maroc
{mouhcine.guennoun@uoit.ca , guennoun@fsr.ac.ma, khalil.el-khatib@uoit.ca }

Abstract. Feature selection is a critical step in building a large scale identification system based on electrocardiogram (ECG) signal. During this phase, the set of features that deemed to be the most effective attributes are extracted in order to construct suitable identification algorithms. A key problem that many researchers face is how to choose the optimal set of features since not all features are relevant to the identification algorithm, and in some cases, irrelevant and redundant features can introduce noisy data that distracts the learning algorithm and therefore severely degrade the accuracy of the identification system and cause slow training and testing process. In this paper, we present a complete framework to select the best set of ECG signal features that efficiently identify an individual. Our framework uses a hybrid approach for feature selection that combines the filter and wrapper models. In this approach, we rank the features according to the score assigned by an independent measure: the information gain ratio. Mahalanobis based classifier's predictive accuracy is used to reach an optimal set of features that maximizes the identification rate.

Keywords: Electrocardiogram, Feature selection.

1. Introduction

In recent years, advancement in computing and digital signal processing technologies have been achieved that allow automated identification of people based on their biological, physiological or behavioral traits. The technologies have also increased the number of traits that can be collected and used to identify people and to control access to resources. Systems that use any biological, physiological or behavioral trait to grant access to resources are called biometric systems.

Biometric systems include fingerprint, voice, facial recognition, voice, hands geometry and iris. Over the last few years, researchers have been investigating the use of heart electrocardiogram (ECG) signal as a biometric trait to identify individuals. Unlike the other human biometrics traits, the ECG signal is a characteristic that is hard to falsify, and can also be used to detect the liveness of the user. The validity of using ECG for biometric recognition is supported by the fact that the physiological and geometrical differences of the heart in different individuals display certain uniqueness in their ECG signals [1]. Studies in [2] and [3] showed the stability and distinctiveness of ECG as a biometric trait. An ECG based biometric recognition system can be applied in wide applications such as physical access control, medical records management, as well as government and forensic applications.

2. Related Work

Extensive work has been done to identify individuals based on the ECG signal. Signal processing techniques are used to extract the characteristic features of an individual. Existing solutions for biometric recognition from electrocardiogram (ECG) signals are based on temporal and amplitude distances between detected fiducial points. Shen et al. [4] extracted 7 temporal and amplitude features from the QRST wave. They combined a template matching method with decision based neural network (DBNN) to implement an identity verification system. The template-matching method is first applied to calculate the correlation coefficient for comparison of two QRS complexes. The DBNN is fed with the possible candidates resulting from the first step. The authors claimed that they have achieved a 100% verification rate using this combined method on a system of 20 subjects. They later extended the proposed method in a larger database that contains 168 individuals and 17 temporal and amplitude features were used [5]. The success rate was 95%.

In [6], the authors used specialized hardware equipments to extract 30 ECG features. A simple feature selection algorithm based on analysis of correlation matrix is employed to reduce the dimensionality of features to 21. Further selection of feature set is based on experiments. Soft Independent Modeling of Class Analogy (SIMCA) method based on Principal Component Analysis (PCA) is used for classification. The study involved 20 persons, and 95-100% identification rate was achieved by using empirically selected features. A major drawback of Biel et al's method is the high number of features used to identify a person. This limits the scope of applications to small size databases. Furthermore, the best set of features for high identification rate was found experimentally and was not proved using a feature selection algorithm.

In [7], the authors combined analytic and appearance based features to achieve high recognition accuracy. Twenty one (21) analytic features are extracted using known fiducial detectors [8,9]. The appearance features are extracted using unsupervised learning techniques based on Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). This work focused more on the extraction of features rather on reducing the dimensionality of features. Israel et al. [8] extracted a total number of 15 features, which are time duration between detected fiducial points. The number of features was reduced to 12 during the feature selection process. The feature selection was provided by a stepwise canonical correlation that used the Wilkes' lambda method. Linear Discriminant Analysis (LDA) for classification. This system was tested on a database of 29 subjects, and the authors reported 100% human identification rate and around 81% heartbeat recognition rate.

Most of the works cited above were tested on small size systems of 20 to 100 individuals; therefore, there was no focus on reducing the number of features to build an efficient identification system that can scale to a large database of individuals. In this paper we propose a feature selection algorithm to extract the most relevant set of features that can be employed in an ECG based identification system for a large dataset of users.

3. Feature Selection

There are currently two models in the literature for feature selection: the filter model and the wrapper model [10]. The wrapper model uses the predictive accuracy of a classifier as a mean to evaluate the goodness of feature set, while the filter model uses a measure such as information, consistency, or distance measures to compute the relevance of a set of features. These approaches suffer from many drawbacks: the first major drawback is that feeding the classifier with arbitrary features may lead to biased results and hence we cannot rely on the classifier's predictive accuracy as a measure to select features. A second drawback is that, for a set of N features, trying all possible combinations of features (2N combinations) to find the best combination to feed the classifier is not a feasible approach.

Different techniques were used to tackle the problem of feature selection. In [11], Sung and Mukkamala used feature ranking algorithms to reduce the feature space of the DARPA dataset from 41 features to 6 most important features. They used three ranking algorithms based on Support Vector Machines (SVM), Multivariate Adaptive Regression Splines (MARS) and Linear Genetic Programs (LGP's) to assign a weight to each feature. Experimental results showed that the classifier's accuracy degraded by less than 1% when the classifier is fed with the reduced set of features. Sequential backward search was used in [12,13] to identify the important set of features: starting with the set of all features, one feature is removed at a time until the accuracy of the classifier is below a certain threshold. Different types of classifiers were used with this approach including Genetic Algorithms in [13], Neural Networks in [12,14] and Support Vector Machines in [12]. Feature selection was proven to have a significant impact on the performance of the classifiers. Experiments in [15] shows that feature selection can reduce the building and testing time of a classifier by 50%.

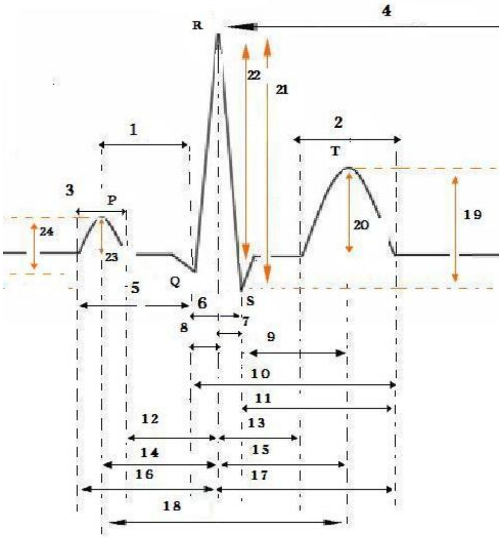

Fig. 1: Extracted Features

4. Feature Selection Algorithm

To select the best set of features, we collect, during the enrolment phase, around 2 minutes of heart beat data for 20 patients between 20 to 40 years old. Both women and men have been participating in the experiment. 24 features are extracted from the ECG signal (Figure 1 and Table 1). We then use the information gain ratio measure (see next section) to assign a score to each individual feature. The features are ranked based on this score. We then use a sequential forward selection algorithm to reach the optimal subset of features. The selection algorithm starts with an empty set S of best features, and then proceeds to add features from the ranked set of features F into S sequentially.

The "goodness" of the resulting set of features S is measured by the Mahalanobis based classifier accuracy. The selection process stops when the gained classifier's accuracy is below a certain selected threshold value or in some cases when the accuracy drops, which means that the accuracy of the current subset is below the accuracy of the previous subset.

Table 1: List of Extracted Features

Features				
Temporal	**1.**sPQ	**2.**T	**3.**P	**4.**RR
	5.PQ	**6.**QRS	**7.**sRS	**8.**sRQ
	9.SsT	**10.**QT	**11.**SfT	**12.**sRfP
	13.sRT	**14.**sRsP	**15.**sRsT	**16.**sRdP
	17.sRfT	**18.**PT		
Amplitude	**19.**TS	**20.**TS'	**21.**RS	**22.**QR
	23.PL'	**24.**amPQ		

5. Ranked Features

Using the data set of ECG signals collected from 10 patients, we could rank the features according to the score assigned by the IGR measure. The 12 top ranked features are shown in Table 2.

Table 2: Top 12 features

Rank	Feature	IGR
1	RS	0.81
2	QR	0.72
3	sRfP	0.52
4	sRsP	0.52
5	sRdP	0.50
6	sRfT	0.48
7	SfT	0.48
8	PQ	0.41
9	QT	0.40
10	SsT	0.38
11	RR	0.37
12	sPQ	0.36

6. The Best Subset of Features

The Mahalanobis based classifier was used to compute the identification rate for each set of features. Initially, the set of features *S* contains only the top ranked feature. After each iteration, a new feature is added to the list *S* based on the rank it's assigned by the IGR measure. Figure 3 shows the accuracy of each subset of features. We note that S_i is the *i* first features in the ranked list of features.

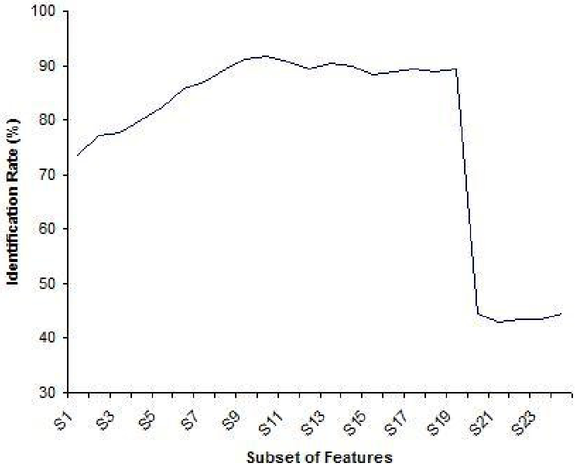

Fig 3: Mahalanobis based classifier's accuracy

Subset of features S_9 maximizes the accuracy of the Mahalanobis based classifier. We can conclude that the first 9 features (RS,QR,sRfP,sRsP,sRdP,sRfT,SfT,PQ,QT) are the best set of features to build a large scale biometric identification system based on the ECG signal. Increasing the number of features does not contribute to the improvement of the accuracy. Worst, irrelevant features distract the classifier and the accuracy drops to 43% with 21 features.

7. Discussion

By doing 1-to-N comparisons between an input ECG signature and all the stored user ECG templates, a user is identified. This raises the question of the applicability of this model in real security scenarios. Unlike finger print, a camera or video recording, or keystroke/mouse action behaviour logs, ECG is not something that could be traced or collected from a physical or digital crime scene. On the other hand, this 1-to-N verification model is not generally suitable for user authentication, where authenticating a user is based solely on an error threshold between the input signature and the stored template of the same user. One major issue that limits the deployment of this research in the field of security is the sensor complexity, that is, to be able to precisely sense all heartbeat embedded features. Furthermore, data values stability for the same individual is an issue. The heart beat characteristics continuously change according to the physical and emotional status of the subject. For example, an ECG profile built based on data collected from a relaxed user will not necessary match a new sample collected from the same user who just rushed upstairs to access the secure building.

8. Conclusion

In this paper, we have presented a novel approach to select the best set of features to efficiently identify individuals using ECG data. Our approach is based on hybrid approach that combines the filter and wrapper models for selecting relevant features. Experimental results show clearly the efficiency of this method. Indeed, we were able to reduce the number of features from 24 to 9 features.

9. References

[1] R. Hoekema, G. G. H. Uijen, A. van Oosterom, ”Geometrical aspect of the interindividual variaility of multilead ECG recordings”, IEEE Trans. Biomed. Eng., vol.48, pp.551-559, 2001.

[2] B. P. Simon and C. Eswaran, ”An ECG classifier designed using modified decision based neural network”, Comput. Biomed. Res., vol. 30, pp. 257-272, 1997

[3] G. Wuebbeler, et al., Human Verification by Heart Beat Signals, Working Group 8.42, Physikalisch-Technische Bundesanstalt (PTB).

[4] T.W. Shen,W. J. Tompkins, and Y. H. Hu, ”One-lead ECG for identity verification”, Proc. of the 2nd Conf. of the IEEE Eng. in Med. and Bio. Society and the Biomed. Eng. Society, vol. 1, pp. 62-63, 2002

[5] T.W. Shen, ”Biometric Identity Verification Based on Electrocardiogram (ECG)”, PHD Dissertation, University ofWisconsin, Madison, 2005

[6] L. Biel, O. Pettersson, L. Philipson, P. Wide, ”ECG analysis: a new approach in human identification”, IEEE Trans. on Instrumentation and Measurement, vol. 50, no. 3, pp. 808-812, 2001

[7] Y. Wang, F. Agrafioti, D. Hatzinakos, K. N. Plataniotis , “Analysis of Human Electrocardiogram (ECG) for Biometric Recognition”, EURASIP Journal on Advances in Signal Processing, Volume 2008, Article ID 148658.

[8] S. A. Israel, J. M. Irvine, A. Cheng, M. D. Wiederhold, and B. K. Wiederhold, ”ECG to identify individuals”, Pattern Recognition 38 (1): 133-142, 2005

[9] S. A. Israel, W. T. Scruggs, W. J. Worek, and J. M. Irvine, ”Fusing face and ECG for personal identification”, Proc. of 32nd Applied Imagery Pattern Recognition Workshop, pp.226-231, 2003.

[10] H. Liu, H. Motoda, “Feature Selection for Knowledge Discovery and Data Mining”, Boston: Kluwer Academic, 1998.

[11] A. H. Sung, S. Mukkamala, "The Feature Selection and Intrusion Detection Problems", In Proceedings of the 9th Asian Computing Science Conference, Lecture Notes in Computer Science 3029, Springer 2004.

[12] A.H. Sung, S. Mukkamala, "Identifying important features for intrusion detection using support vector machines and neural networks", Proceedings of the 2003 Symposium on Applications and the Internet (SAINT'03), January 2003.

[13] G. Stein, B. Chen, A.S. Wu, K.A. Hua, "Decision tree classifier for network intrusion detection with GA-based feature selection", In proceedings of the 43rd ACM Southeast Regional Conference – Volume 2, Kennesaw, Georgia, USA, March 2005.

[14] A. Hofmann, T. Horeis, B. Sick, "Feature selection for intrusion detection: an evolutionary wrapper approach", In Proceedings of the 2004 IEEE International Joint Conference on Neural Networks, July 2004.

[15] Y. Chen, Y. Li, X. Cheng, L. Guo, "Survey and Taxonomy of Feature Selection Algorithms in Intrusion Detection System", Inscrypt 2006.

[16] A. K. Jain, A. Ross, S. Prabhakar “An Introduction to Biometric Recognition”, In IEEE Transactions On Circuits and Systems for Video Technology, VOL. 14, NO. 1, JANUARY 2004.

[17] J.R. Quinlan, "Induction of decision trees", Machine Learning, 1, p81-106.