Features Selection for Automatic ECG Personal Identification

Yuliyan Velchev[1], Strahil Sokolov[2] and Ognian Boumbarov [3]

Abstract – **In this paper an approach for features selection for ECG personal identification is presented. The proposed approach uses a combination of the following types of features: temporal and coefficients obtained from PCA transformation. Temporal attributes of the ECG signals are extracted using signal segmentation based on HMM applied on wavelet transformation coefficients. The PCA features are obtained from a matrix, which is formed from signal portions for each detected cardiac cycle. In order to select the most representative signal portions a correlation matrix analysis is performed.**

Keywords – **ECG, Personal Identification, PCA.**

I. INTRODUCTION

Most of the systems for biometric identity recognition are based on unique information from fingerprints, signature, eye iris, human face etc. However, these modalities can't provide enough reliability, especially for systems with high level of security. The contemporary concept for personal identification system is to combine the well-known modalities (fingerprints, iris, face etc.) with biometrical modality, which is difficult to forge such as the electrocardiogram (ECG).

The ECG is a record of time varying bioelectric potentials generated by electrical activity of the heart. Normal human ECG consists of components named waves and complexes (Fig. 1).

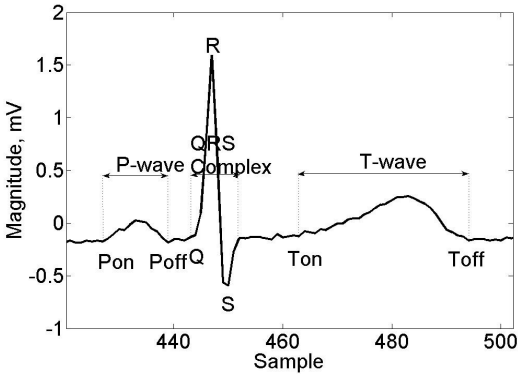

Fig. 1. Normal human ECG and its standard components (waves and complexes)

[1]Yuliyan Velchev is with the Faculty of Telecommunications at Technical University of Sofia, 8 Kl. Ohridski Blvd, Sofia 1000, Bulgaria, E-mail: julian_velchev@abv.bg.

[2]Strahil Sokolov is with the Faculty of Telecommunications at Technical University of Sofia, 8 Kl. Ohridski Blvd, Sofia 1000, Bulgaria, E-mail: s_sokolov@tu-sofia.bg.

[3]Ognian Boumbarov is with the Faculty of Telecommunications at Technical University of Sofia, 8 Kl. Ohridski Blvd, Sofia 1000, Bulgaria, E-mail: olb@tu-sofia.bg.

Sequential depolarization of the heart's right and left atrium reflects in ECG as P wave. The QRS complex corresponds to right and left ventricle depolarization. Ventricular repolarization forms the T wave in ECG. The regions between waves (complexes) are defined as isoelectric line or baseline.

It is proven that physiological differences in human heart for different individuals reflect in ECG as unique time intervals, amplitudes and shapes [3]. Also ECG analysis is attractive for human identification purpose because the ECG registration is relatively easy and noninvasive.

The remainder of this paper is organized as follows. In section 2 a brief overview of some existing methods for ECG personal identification is given. Section 3 presents a method for ECG segmentation for the purpose of extraction of temporal features. Section 4 describes a method for features extraction using Principal Component Analysis (PCA) applied on a complete cardiac cycle. Section 5 shows experimental results.

II. RELATED WORKS

An ECG person identification method using a set from temporal and amplitude features is described in [1]. The system has been tested with a database of about 20 individuals. The achieved recognition performance is near 100%, although the described system is not fully automatic.

Another method for ECG personal identification is given in [8]. It uses a combination of autocorrelation with discrete cosine transform (DCT). The described method refers as the AC/DCT method. The major advantage is avoiding the automatic ECG segmentation, which is one of the most complicated tasks in automatic ECG analysis. This method involves the following stages: signal segmentation by windowing in order to extract a portion with two or more cardiac cycles; normalized autocorrelation function (ACF) estimation for each segment; performing DCT over each ACF; classification using only significant DCT coefficients. One of the method's drawbacks is that the proper window size selection for ECG segmentation is critical and strongly depends on current heart rate.

In [9] is introduced a method, which utilizes the heart rate variability (HRV) as human identification feature. The HRV utilization for personal identification requires relatively long-term ECG recordings, which in some cases may be impractical.

In [2] is present an ECG identification method, which uses the Bayes' theorem. The authors claim that the Bayesian classifier outperforms the Mahalanobis' distance method for ECG identification purposes.

According to the performance comparison given in [8] the best performance for human ECG recognition is achieved by

using combination of different types of ECG features. For this reason, this paper describes a combination of temporal features with PCA coefficients.

III. ECG SEGMENTATION. TEMPORAL FEATURES EXTRACTION

ECG segmentation (ECG components onset and offset determination) is important stage in terms of extraction of temporal features as well as segmentation of regions (cardiac cycles) with high degree of similarity.

The ECG signal preprocessing is baseline drift reduction by high pass filtering and wavelet denoising by using soft threshold applied on wavelet coefficients [4].

The ECG segmentation is based on hidden Markov model (HMM) with Gaussian mixture model (GMM) core [7]. The HMM is probabilistic model which describes the statistical dependency between an observation sequence O and sequence of hidden states S. A HMM λ is denoted with:

$$\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{q}), \tag{1}$$

where $\mathbf{A}$ is transition probability matrix, $\mathbf{B}$ is observation probability distribution and $\mathbf{q}$ is initial distribution vector. The model used in this project is first order Markov process (Markov chain):

$$P(s_{t+1} \mid s_t, s_{t-1}, s_{t-2}, ...) = P(s_{t+1} \mid s_t), \tag{2}$$

where s_t is the state (ECG component) for time t. The ECG segmentation task is to find the optimal state sequence S' given the HMM parameters λ and observation sequence O:

$$\begin{aligned} S' &= \underset{S}{argmax}\{P(S \mid O, \lambda)\} = \\ &= \underset{S}{argmax}\left\{\frac{p(S, O \mid \lambda)}{p(O \mid \lambda)}\right\} = \\ &= \underset{S}{argmax}\{p(S, O \mid \lambda)\} \end{aligned} \tag{3}$$

The optimal state sequence S' in (3) is determined as the most probable state sequence given the observation. The final result is achieved using Bayes' rule. An effective way to calculate (3) is to use the Viterbi algorithm [5].

HMM learning (finding the HMM parameters λ) is a task which can be done using supervised or unsupervised techniques. Supervised learning is related to maximum likelihood:

$$\lambda' = \underset{\lambda}{argmax}\{p(\mathbf{O} \mid \lambda)\}, \tag{4}$$

where λ' represents HMM parameters, which maximize the probability and $\mathbf{O} = \{O_1, O_2, ..., O_N\}$ are the observation sequences.

Unsupervised learning generally employs the Expectation Maximization (EM) algorithm. It is only possible when more than one GMM component (mixture) is used. It has been proved through experiments that the EM algorithm can be applied successfully only to "refine" the already learned parameters that are obtained from supervised training.

One of the HMM requirements is the statistical independence between consecutive observations. Since ECG signals don't satisfy this assumption, the HMM can perform better when working with observation vectors resulting from appropriate time-frequency signal transform (wavelet transform is used in our approach) [7]. The coefficients from wavelet transform W serve as similarity measure between signal and a wavelet function at different scales s and time shifts τ as parameters:

$$W_x(s,\tau) = \langle x, \psi_{s,\tau} \rangle = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t-\tau}{s}\right) dt, \quad s \in \mathbb{R}^+, \tau \in \mathbb{R} \tag{1}$$

where $x(t)$ is the ECG signal and ψ is the wavelet function [6].

The discrete wavelet transform (DWT) is not suitable for this purpose, since the transform is not time-invariant and the resulting coefficients are decimated, so it is impossible to build observation vectors associated to a given signal sample. The continuous wavelet transform (CWT) doesn't have these limitations, but it can be used with reduced set of scales, because this transform has large amount of information redundancy. In the case, with reduced number of scales, the perfect signal reconstruction is not possible. The stationary wavelet transform (SWT) possesses compact support (perfect reconstruction is possible), but the number of available wavelet functions is limited. As the perfect signal reconstruction is not necessary in this case, the CWT with dyadic scales is used to build the observation vectors.

The wavelet function ψ should satisfy several requirements: maximum similarity between ψ and ECG components; minimal effective width; it is also desirable for ψ to be symmetric in order to avoid complicated phase corrections. The wavelet functions, which can fulfill these requirements, are: second derivative of the Gaussian function (Mexican hat), Morlet and all wavelets from Coiflet family. The final choice is made by segmentation performance comparison of a HMM combined with different wavelet functions. The segmentation performance criterion is receiver operating characteristics (ROC) plot. The ROC analysis shows that the best suited wavelet function for ECG segmentation is the second derivative of Gaussian function. In Fig. 2 is shown the ROC plot for T wave detection for different types of wavelet functions. Also after considering the ROC analysis the optimal dyadic wavelet scales are determined to be form 2 up to 64. So the observation vectors (features) $\mathbf{f}$ are built as follows:

$$\mathbf{f}_{O \equiv \tau} = \left[W_x(2,\tau), W_x(4,\tau), W_x(8,\tau), W_x(16,\tau), W_x(32,\tau), W_x(64,\tau)\right]^T \tag{5}$$

The number of components (mixtures) for HMM-GMM is chosen empirically to be 8. Some experiments have been made using HMM with Single Gaussian model (SGM) and conditional random fields (CRF), but the results are significantly worse than HMM-GMM.

The comparison is based on a synthetic ECG signal which is composed from Gaussian functions. This approach ensures

that ECG signal is noise free and eliminates the subjective criterion in terms to define onset and offset in ECG waves.

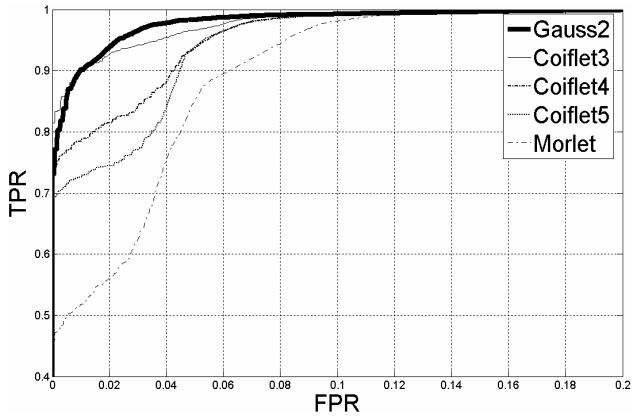

Fig. 2. T wave segmentation performance using HMM-SGM with different types of wavelet functions

Finally, the model performance is evaluated by segmenting the same synthetic signal corrupted with additive white Gaussian Noise (AWGN) (Fig. 3). The AWGN is often used to model the muscle artifacts (EMG noise).

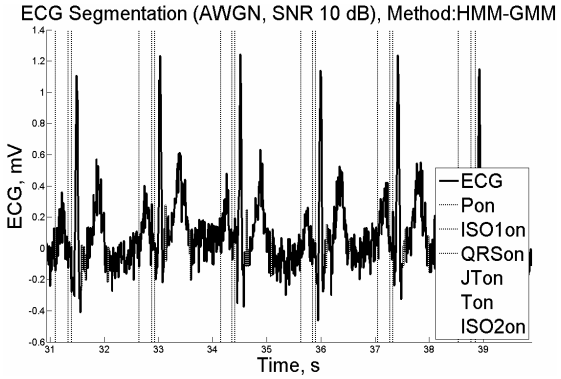

Fig. 3. ECG segmentation result with AWGN influence (SNR=10 dB)

The proposed HMM-GMM has 6 states: P wave; isoelectric line between P wave and QRS complex; QRS complex; JT segment (the segment between QRS complex and T wave); T wave and baseline between T wave and next P wave.

Experimental results in terms of segmentation performance for ECG segmentation model with HMM-GMM are given in section 5.

Temporal features such as duration of P wave, QRS complex, T wave and others are directly obtained by the segmentation process. The duration of isoelectric line between T wave and next P wave is excluded as temporal feature because it depends strongly on current heart rate.

At the end of this stage, the signal is segmented into complete cardiac cycles that contain all types of waves. All of the segmented and labeled cardiac cycles are of equal length. In the next stage they are subjected to feature extraction using a linear subspace-transformation.

IV. PCA FEATURES EXTRACTION

The principal component analysis (PCA) is a well-known technique for dimensionality reduction that concentrates the discriminative information into a low number of coefficients [10]. The main goal behind the reduction of a complex set of data to a lower dimension is to reveal the simpler structure that underlies it. The previous stage of our approach provides noise reduction and segmentation of the ECG signal. We consider each segmented and labeled cycle a row vector $\mathbf{X} = [x_1, x_2, ..., x_N]$ with a finite number of elements N, where each element x_i is a sample from the signal for a given cardiac cycle. For the purposes of our experiment we have selected also a fixed number M of training cycles per person that are strongly correlated. Thus in the training set we have M row vectors $\mathbf{X}_j, j = 1,...,M$. In order to prepare the transformation in PCA subspace, we calculate the mean vector and the difference vectors:

$$\mathbf{\Psi} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{X}_i,\ i = 1,...,M \tag{6}$$

$$\mathbf{\Phi}_j = \mathbf{X}_j - \mathbf{\Psi},\ j = 1,...,M \tag{7}$$

Difference vectors are stored in a matrix $\mathbf{C}$:

$$\mathbf{C} = [\mathbf{\Phi}_1, ..., \mathbf{\Phi}_M] \tag{8}$$

The next step is to calculate the covariance matrix $\mathbf{L} = \mathbf{C}\mathbf{C}^T$, and its eigenvectors $\mathbf{V}$ and eigenvalues $\boldsymbol{\lambda}$ that satisfy the equation:

$$\Delta(\boldsymbol{\lambda}) = \det\left|[\mathbf{L}] - \lambda_S[\mathbf{I}]\right| = \begin{vmatrix} L_{11}-\lambda_S & L_{12} & ... & L_{1M} \\ L_{21} & L_{22}-\lambda_S & ... & L_{2M} \\ ... & ... & ... & ... \\ L_{M1} & L_{M2} & ... & L_{MM}-\lambda_S \end{vmatrix} = 0 \tag{9}$$

Then, the eigenvectors $\mathbf{v}_l, l = 1,...,M$ of the matrix $\mathbf{L}$ are calculated according to:

$$\begin{aligned} &(L_{11}-\lambda_s)v_1^s + L_{12}v_2^s + ... + L_{1M}v_M^s = 0 \\ &L_{21}v_1^s + (L_{22}-\lambda_s)v_2^s + ... + L_{2M}v_M^s = 0 \\ &... \\ &L_{M1}v_1^s + L_{M2}v_2^s + ... + (L_{MM}-\lambda_s)v_M^s = 0 \end{aligned} \tag{10}$$

Each eigenvector $\mathbf{v}_l^S$ contains M components: $\mathbf{v}_S = \left[v_1^S, v_2^S, ..., v_i^S, ..., v_M^S\right]^T$ and the following rule is valid:

$$\left(v_1^S\right)^2 + \left(v_2^S\right)^2 + ... + \left(v_i^S\right)^2 + ... + \left(v_M^S\right)^2 = 1 \tag{11}$$

The transformation matrix $\mathbf{U}$ is calculated from the difference matrix and the eigenvectors:

$$\mathbf{U} = \sum_{i=1}^{M}\mathbf{\Phi}_i\mathbf{v}_i = \mathbf{C}\mathbf{V}, i = 1...M \tag{12}$$

Finally, the PCA-projections of the cardiac cycles are calculated as follows:

$$\mathbf{\Omega}_j = \mathbf{U}^T\left(\mathbf{X}_j - \mathbf{\Psi}\right), j = 1,...,M \tag{13}$$

In the PCA sub-space we create person-specific clusters using LBG algorithm. Classification is realized with

computation of distance between input cycle projection and the centers of the clusters.

PCA is an optimal method which delivers the opportunity to perform back-projection to signal space - the mean square error (MSE) between the restored signal and original signal is minimized.

V. EXPERIMENTAL RESULTS

The ECG segmentation performance is evaluated by using mean value μ_e and standard deviation σ_e of the absolute error between expert annotated and detected ECG characteristic point (onset and offset of a given ECG wave or complex). The segmentation performance results for noise free synthetic ECG and synthetic ECG with AWGN can be seen in Table 1.

TABLE I
STATISTICAL PARAMETERS OF THE ABSOLUTE ERROR FOR ONSET (OFFSET) DETECTION OF SOME THE STANDARD ECG COMPONENTS

	P_{on}	$ISO1_{on}$	QRS_{on}	T_{on}
	μ_e, ms σ_e, ms	μ_e, ms σ_e, ms	μ_e, ms σ_e, ms	μ_e, ms σ_e, ms
Noise free ECG	-3.8 6.1	5.8 4.2	-11.1 5.1	-6.9 7.2
ECG with AWGN (SNR = 10 dB)	7.5 48.1	-4.8 91.6	-16.2 7.1	4.4 12.3

The ECG-cycles projections for two different persons form clusters in PCA subspace are shown in Fig. 4. The two clusters according to the first three principal components can be clearly seen on the image.

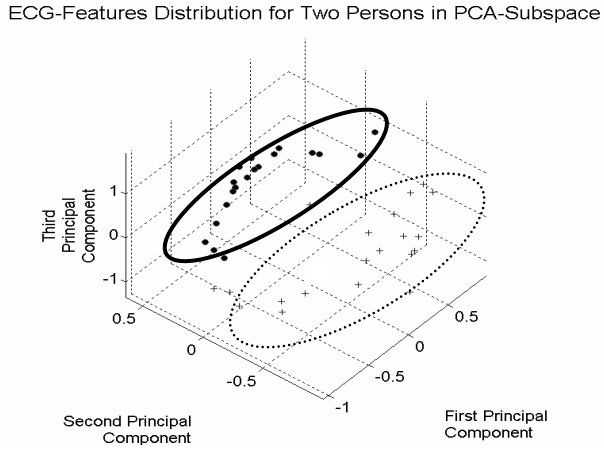

Fig. 4. ECG-Features Distribution for two persons (o and +) according to the first three principal components in PCA subspace.

In our experiments we have used real-world ECG data from 10 persons, which is acquired with industry-standard ECG hardware, equipped with PC-interface. We used MATLAB and LabView as software platforms.

VI. CONCLUSION

In this paper we presented an approach for features selection for ECG personal biometric identification. Features are selected as combination from temporal parameters of the ECG signal and PCA coefficients. The PCA transformation is applied on ECG portions which represent complete cardiac cycles (without baseline between T and P waves). For these purposes the ECG is segmented by using HMM with Gaussian mixtures. The model uses observation vectors which are formed from wavelet transform coefficients. In our future work we plan to implement our approach as a part of a personal identification system by applying neural network classifiers.

ACKNOWLEDGEMENT

This work was supported by National Ministry of Science and Education of Bulgaria under contract BY-TH-202/2006: "Multimodal biometric analysis: methods and algorithms".

REFERENCES

[1] L. Biel, O. Pettersson, L. Philipson, P. Wide, "ECG Analysis: A New Approach in Human Identification", IEEE Transactions on Instrumentation and Measurement, vol. 50, no. 3, pp. 308-312, 2001.

[2] Zhaomin Zhang, Daming Wei, "A New ECG Identification Method Using Bayes' Theorem", 2006 IEEE Region 10 Conference TENCON 2006, pp. 1-4, 2006.

[3] Y. N. Singh, P. Gupta, "ECG to Individual Identification", 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS 2008), pp. 1-8, 2008.

[4] V. Georgieva, R. Kountchev, "An Influence of the Wavelet Packet Decomposition on the Noise Reduction in Ultrasound Images", ICEST 2006, pp. 185-188, 2006.

[5] R. L. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, 1989.

[6] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, 2nd edition, 1999.

[7] N. P. Hughes, L. Tarassenko, S. J. Roberts, Markov Models for Automated ECG Interval Analysis, In Advances in Neural Information Processing Systems, vol. 16, NIPS, 2003.

[8] Y. Wang, F. Agrafioti, D. Hatzinakos, K. N. Plataniotis, Analysis of Human Electrocardiogram for Biometric Recognition, Journal on Advances in Signal Processing (EURASIP), vol. 2008, no. 1, 2008.

[9] J. M. Irvine, B. K. Wiederhold, L. W. Gavshon, et al., Heart rate variability: a new biometric for human identification, Proceedings of the International Conference on Artificial Intelligence (IC-AI '01), pp. 1106–1111, 2001.

[10] F. Castells, Principal Component Analysis in ECG Signal Processing, EURASIP Journal on Advances in Signal Processing, vol. 2007, 2007.