Bio-Medical Materials and Engineering 24 (2014) 2883–2891
DOI 10.3233/BME-141107
IOS Press

Feature extraction from a novel ECG model for arrhythmia diagnosis

Junjiang Zhu, Lingsong He* and Zhiqiang Gao
[a]*Mechanical and engineering, Huazhong University of Science & Technology, 1037 Luoyu Road, Wuhan 430000, China*

Abstract. Feature extraction is a crucial aspect of computer-aided arrhythmia diagnosis using an electrocardiogram (ECG). A location, width and magnitude (LWM) model is proposed for extracting each wave's features in the ECG. The model is a stream of Gaussian function in which three parameters (the expected value, variance and amplitude) are applied to approximate the P wave, QRS wave and T wave. Moreover, the features such as the P-Q intervals, S-T intervals, and so on are easily obtained. Then, a mixed approach is presented for estimating the parameters of a real ECG signal. To illustrate this model's associated advantages, the extracted parameters combined with R-R intervals are fed to three classifiers for arrhythmia diagnoses. Two kinds of arrhythmias, including the premature ventricular contraction (PVC) heartbeats and the atrial premature complexes (APC) heartbeats, are diagnosed from normal beats using the data from the MIT–BIH arrhythmia database. The results in this study demonstrate that using these parameters results in more accurate and universal arrhythmia diagnoses.

Keywords: ECG, feature extraction, classification, arrhythmia diagnosis, parametric Gaussian functions

1. Introduction

Arrhythmia, also known as cardiac dysrhythmia or an irregular heartbeat, refers to a large and heterogeneous group of conditions characterized by abnormal activities. Some types are life-threatening, whereas others might not be as critical but still require attention and treatment. In this paper, these non-fatal types are the research focus. In clinical situations, diagnosing arrhythmia is primarily based on analyzing electrocardiogram (ECG) signals, which are both economical and non-invasive [1]. Since a beat-by-beat human-based examination can often be very time-consuming and tedious [2], automatic ECG analysis has become an area of intense research in recent years.

Abnormal activities of different origin sites define different kinds of arrhythmias. For example, when another region of the atria depolarizes before the sinoatrial node, atrial premature complexes (APC), also called premature atrial contractions (PACs) or atrial premature beats (APB), occur. However, if the heartbeat is initiated by Purkinje fibers in the ventricle, the arrhythmia is classified as a premature ventricular contraction (PVC). PVC can be diagnosed from normal heartbeats by checking the spacing and time between the PVC and the preceding QRS wave. It can only be distinguished from

*Corresponding author: Lingsong He, Mechanical and engineering, Huazhong University of Science & Technology, 1037 Luoyu Road, Wuhan 430000, China. Tel.: +86-027-87543970; Fax: +86-027-87543970; E-mail: helingsong@hust.edu.cn.

APC by the compensatory pause. Therefore, determining how to diagnose different kinds of arrhythmias is a challenging process.

In literature, diagnosing arrhythmias by the pattern recognition scheme is generally composed of three blocks: the preprocessing of raw biological signals, the feature extraction procedure to highlight some of the most important patterns, and finally, a classification stage in which different classes are identified. Among these blocks, feature extraction is a crucial step because inaccurate features could lead to errors. Over the past years, there has been extensive literature detailing the investigation of the performance of features extracted in the time domain [3], frequency domain [4], time-frequency domain [5], and wavelet package decomposition [6]. The methods can be divided into three types: the subspace analysis (such as the principal component analysis (PCA), independent component analysis (ICA) and sparse representation), statistical methods (such as higher order spectral features) [8], and key points. For subspace analysis, the time domain, frequency domain or wavelet details of the original signal are projected into different subspaces, and the main components are usually taken as the features. The main challenge associated with this method lies in the fact that ECG signals are nonlinear, highly subjective and fuzzy [7]; therefore, the features extracted by these methods may contain subjective, and thus highly misleading, information. For statistical methods, the signals' mean value, mean square deviation, peak-to-peak value, higher order statistics of the time domain, frequency domain or wavelet details are taken as the features. The main drawback of the statistical method is the ability to miss nonlinear features. Even more so, features that perform well in some cases may be lost in another situation. Later on, ECG signals were analyzed more and more by using advanced signal processing techniques, such as recurrence quantification analysis [9], fractal theory [10] and so on. Although these methods are excellent for mining nonlinear features, the medical sense of these features is unclear. In this situation, the key point's analysis is a good solution. However, the primary challenge is extracting the key points. In this paper, we modeled the P wave, QRS wave and T wave as well as extracted the parameters as features to be fed to classification.

2. Material and methods

2.1. Material and Pre-processing method

All the data is from the MIT–BIH arrhythmia database [11], which contains over 109,000 labeled ventricular beats from 15 different heartbeat types including the PVC and APC. These beats are contained in 48 recordings, each of which has a 30 minute duration and includes two leads: the modified limb leads II and one of the modified leads V1, V2, V4 or V5. Additionally, the sampling frequency is 360 Hz. The data had already been filtered by a band-pass filter at 0.1–100 Hz, and the resolution is 200 samples per mV. We used beats that are contained in recording 100, 101, 102, 103, 104, 105, 107, 109, 112, 113, 114, 115, 116, 117, 119, 121, 122, 200, 201, 205, 207, 208, 212, 220, 222, 223, 228, 232, 233, and 234 and extracted a total of 18,234 normal beats, 1987 APC beats, 3369 PVC beats, and additional beats with typical beats.

In most literature, a certain fraction of each recording is used to train the dataset, and another fraction of the heartbeats is used for the testing. However, since ECG signals are highly subjective, the classifier trained by the mentioned kinds of features may perform well for a given database while simultaneously failing to predict a new person's ECG signals. Therefore, in this paper, heartbeats extracted from some persons (recordings 103, 104, 105, 107, 109, 111, and 112) were used for training while the other persons' were used to test the results.

In this paper, wavelet filtering was used to denoise the original signal. The basic wave is DB4 from the Daubechies wavelet family since it is more similar to the QRS wave in shape than the other basic waves.

The Pan Tompkins algorithm [12] was used to detect the QRS complex on the denoised ECG signal, and then the ECG data was segmented so that each segment consisted of 99 samples before the QRS mid-point and 200 samples after the QRS mid-point. Finally, each beat was normalized.

2.2. Features extraction

2.2.1. A new ECG model

ECG reflects a procedure of biological electrical changes which is inspired by a consequential excitement of a pacemaker, atrium and ventricles. In a clinical situation, a single heartbeat is interpreted as the sum of five pluses, which are called the P wave, Q wave, R wave, S wave and T wave, where a P wave represents atrial depolarization; a Q wave is any downward deflection after the P wave; an R wave follows as an upward deflection; the S wave is any downward deflection after the R wave; and a T wave follows the S wave. The present feature extraction method is based on the idea of parameterizing each of these waves with a determined shape.

Since the Q, R and S waves reflect a single event, they are taken as a union and are investigated from the viewpoints of both the time-domain and frequency-domain. In the time domain, QRS lasts 0.06-0.1s for adults, whereas kids may just last 0.04-0.08s. Moreover, many clinical results reveal that a QRS complex will differ from another by the R-amplitudes, Q-amplitudes and S-amplitudes. Therefore, a model with three parameters (the Locations, Widths and Magnitudes, and LWM model) may be enough to describe a QRS complex. Figure 1 illustrates a QRS complex from the MIT–BIH arrhythmia database from both the time and frequency-domain view. In the low frequency section, the relationship existing between the frequency and log-magnitude resembles a quadratic curve. Therefore, a profile following a Gaussian distribution is a suitable shape to start with:

$$Z_{QRS}(t) = \sum_{K=1}^{3} d_k e^{-\frac{(t-t_k)^2}{\sigma_k^2}} \tag{1}$$

where d_k refers to a magnitude, t_k represents the locations and σ_k denotes a single wave's width. A synthesized QRS complex is shown in Figure 2.

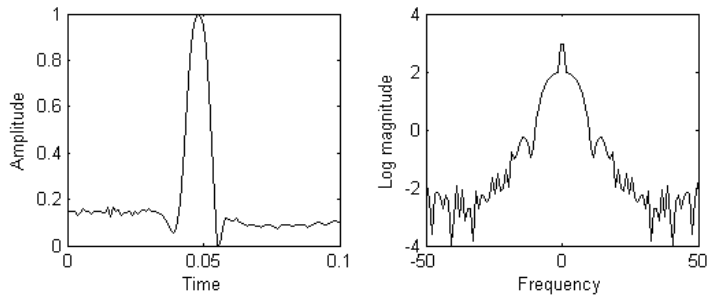

Fig. 1. Typical QRS complex from the MIT–BIH arrhythmia database.

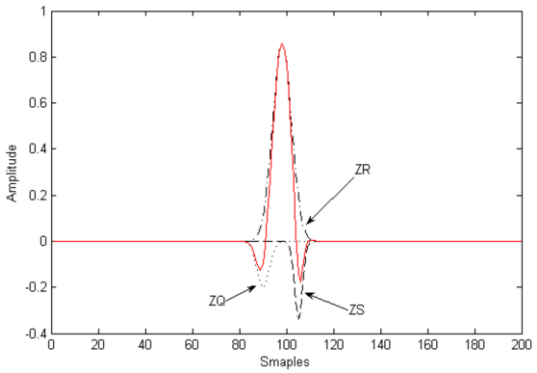

Fig. 2. A synthesized QRS complex (red line) consisting of a Q wave (ZQ), R wave (ZR) and S wave (ZS) using the LWM model.

Note that the relationship between the locations of the three waves is mostly fixed; therefore, the model in Eq. (1) can be simplified as:

$$Z_{QRS}(t) = d_{k1}e^{-\frac{[t-t_{k2}+a(\sigma_{k1}+\sigma_{k2})]^2}{\sigma_{k1}^2}} + d_{k2}e^{-\frac{(t-t_{k2})^2}{\sigma_{k2}^2}} + d_{k3}e^{-\frac{[t-t_{k2}-b(\sigma_{k1}+\sigma_{k2})]^2}{\sigma_{k3}^2}} \tag{2}$$

where a and b are constant and can be obtained from a statistical analysis.

The P wave is yet another important part of a heartbeat. A normal P wave lasts 0.05s-0.1s, although some abnormal beats may have a P wave that lasts more than 0.11s. Additionally, the amplitudes vary among the 0.05mV to 0.25mV range. For consistency with the QRS complex model, using a LWM to describe the P wave is the best choice. Again, we analyze the P waves from both a time-domain view and a frequency-domain field.

Although most P waves are quite smooth and symmetric, there are other existing kinds of profiles, such as dual-peaks waves and bidirectional waves. Therefore, a single Gaussian function is not enough to describe P waves. A combination of two Gaussian functions may cover:

$$Z_P(t) = d_{k1}e^{-\frac{(t-t_k)^2}{\sigma_k^2}} + d_{k2}e^{-\frac{[t-t_k+c\sigma_k]^2}{c\sigma_k^2}} \tag{3}$$

where d_k refers to a magnitude, t_k represents the location, σ_k denotes a single wave's width and c is a constant. In this paper, c=0.45 is experimentally set. Synthesized P waves for each type are shown in Fig. 3.

T waves usually last 0.05-0.25s and are morphologically very similar to P waves. Therefore, the same expression is used to describe T waves. Thus, the model is then given as:

$$Z(t) = Z_P(t) + Z_{QRS}(t) + Z_T(t) \tag{4}$$

where $Z_T(t)$ has the same expression as $Z_P(t)$. Then, determining how to extract the parameters becomes the core problem.

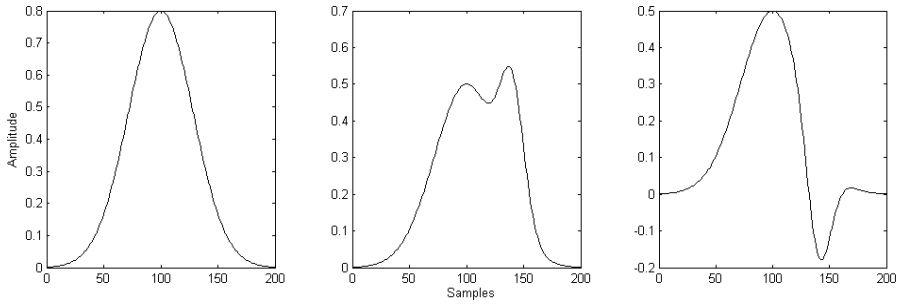

Fig. 3. Synthesized P waves for a smooth P wave with a single peak (a), a dual-peak wave (b) and a bidirectional P wave (c).

2.2.2. Parameters recovery method

From a statistical view, to obtain unknown parameters is to infer the parameters $\theta = \{d_1, \cdots d_7, t_1 \cdots t_3, \sigma_1, \cdots \sigma_5\}$, given the knowledge of the noisy discrete ECG signal, $\mathbf{y}$. In order to find the best set of $\boldsymbol{\theta} = \{\mathbf{d}, \mathbf{t}, \boldsymbol{\sigma}\}$, we define the error as in Eq.(5).

$$\varepsilon = \|\mathbf{y} - \mathbf{z}\|^2 \tag{5}$$

Ideally, the goal would be to directly minimize ε as defined in (5); however, this does not seem to be suitable as the dependence of $z(t)$ on $\boldsymbol{\theta}$ is highly nonlinear. During the experiments, we used a greedy algorithm in an attempt to find a global, optimal solution. Even though the global optimal algorithm generates quite a low error, time-domain aliasing was the primary issue encountered. In other words, the estimated parameters did not clearly reveal medical features as was expected. Thus, determining a local, optimal solution would address the problem.

To find a local solution, a heartbeat must be further segmented. The order of each wave was fixed, and it was already known that the R-peak is the 100th sample; therefore, in fitting the QRS, the P and T wave in a local region becomes possible. In this paper, the trust-region method, a simple yet powerful tool for optimization, was used. Then, the process of parameter estimations was improved and was segmented into two phases, as shown in Figure 4.

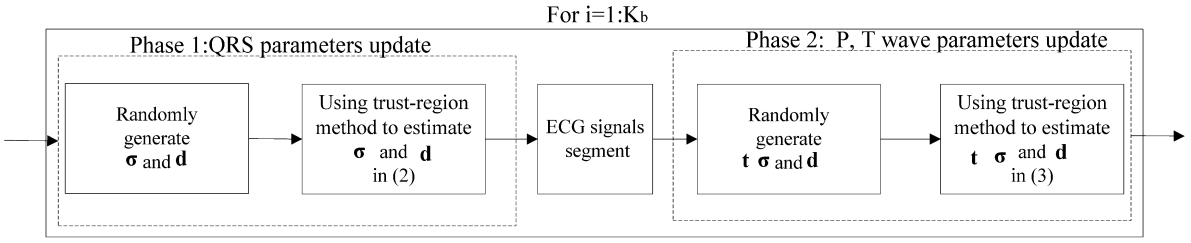

Fig. 4. Flowchart of the parameters extraction method.

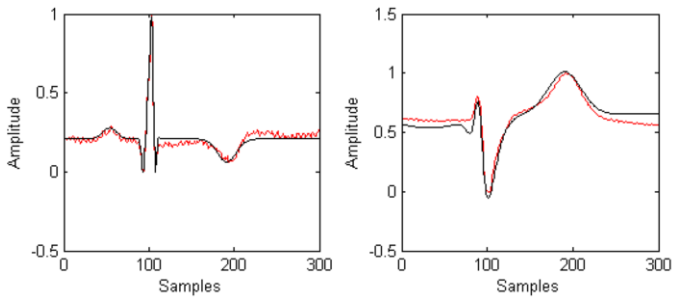

Fig. 5. The fitting result for the APC beat (left) and PVC beat (right). The red lines refer to the original signals, and the black ones refer to the synthesized signals.

Because $t_R = 100$ was already known, we only needed to estimate $\boldsymbol{\sigma}$ and $\mathbf{d}$ in Phase 1. In Phase 2, in order to cut the computational load, we took a segment from the first sample to $(100 - r_{\sigma_R})th$ (not a sample for P wave extraction) and another segment from $(100+r_{\sigma_R}+1)th$ to the last sample for the T wave segment with the fact that the order of waves are fixed. A final estimation result can be seen in Figure 5.

2.3. Classifiers

2.3.1. SVM Classifiers

A support vector machine is a non-linear network based upon statistical learning principles. In particular, it has the ability to classify unseen patterns with a small set of data as the training material and, therefore, is widely used in the pattern recognition scheme. The SVM finds the hyper-plane that maximizes the separation margin between the two different classes while also minimizing the structure risk. Samples are seen as points, and if the patterns are not linear, these points would be mapped into a higher dimension space. This transformation is carried out by kernel function. Literature has incorporated the use of many kinds of kernel functions, such as the Sigmoidal function and the Polynomial and Radial Basis Function kernels. In this paper, a Least Square SVM (presented by Suykens and Vandewalle) [1] was used as was the case in [2].

2.3.2. Neural network classifier

The neural network is famous for its robustness, memory capacity, nonlinear mapping ability and strong self-learning ability. In this paper, two kinds of neural networks are used. One is a common three-layer feed-forward neural network (MLP), and the other is an adaptive three-layer feed-forward neural network, which we called A-MLP. The only distinguishing factor between the two is the activation function. For MLP, we used an S-function as its activation function, but for A-MLP, we used:

$$F = \frac{\alpha}{1+e^{-\beta x}} \tag{6}$$

where α and β are parameters that are decided during the training process. The input layer contains 16 nodes because 16 features are selected from the high dimension feature space. The output layer contains 4 neurons that correspond to the four classes that were used. For MLP, we chose 10

Table 1

Result for training set

		Normal beats			APC			PVC		
		Acc(%)	Spe(%)	Sen(%)	Acc(%)	Spe (%)	Sen (%)	Acc (%)	Spe (%)	Sen (%)
Train-ing set	MLP	97.49	96.27	98.62	96.13	94.34	97.15	96.83	96.34	96.18
	SVM	98.65	97.43	98.76	95.42	93.53	93.27	97.17	97.18	97.03
	A-MLP	99.63	99.74	99.68	98.17	96.87	94.37	94.65	99.01	97.12
Testing set	MLP	87.37	82.17	81.27	90.39	92.03	90.22	88.51	92.02	87.25
	SVM	89.34	83.32	82.21	91.62	93.07	91.19	92.15	91.33	88.13
	A-MLP	92.37	93.53	91.89	93.63	95.29	94.56	92.55	92.41	92.01

neurons for the hidden layer in order to obtain the highest accuracy, and for A-MLP, the hidden layer consisted of 8 neurons by trail. The learning algorithm for the MLP used in this paper was the back propagation. The error between the desired response and obtained response was calculated and then used as a feedback for upgrading the network weights. The training process stops until the error reduces to less than a threshold.

3. Results and discussion

3.1. Simulation result

In this section, the proposed method was tested for three types of heartbeats: normal beats, PVC beats and APC beats. Firstly, for each heartbeat, 16 features were extracted (including Locations, Widths and Magnitudes of each pulse and R-R intervals). The extracted parameters were further tested by being fed to three classifiers. Three measures were taken to evaluate the results for both the training set and testing set: the accuracy, sensitivity and specificity, the definitions of which can be found in [17]. The classification results are shown in Table 1.

Table 1 shows the classification result for the training set. Although the leaning algorithm for A-MLP is more complicated than the other two's leaning algorithm, it achieves the highest accuracy for both the training set and testing set. For the testing set, A-MLP obtains a maximum accuracy of 92.37%, 93.63%, and 92.55% for normal beats, APC and PVC, respectively, meaning that, when our method is used for other people whose ECG signals are not concluded in the training set, we can also obtain a high probability for an accurate diagnosis.

3.2. Comparison between other feature extraction methods

In this section, we present some popular feature extraction methods and then compare them with our work. In paper [13], the Wavelet analysis was chosen as a feature extraction method; the wavelet coefficients are fed to three dimensionality reduction techniques before being used as the input of the classifiers. They found that Daubechies 8 (Db8) yielded the highest accuracy (96.8%). In Amit and Shantanu's work [14], the local fractal dimension (LFD) at each sample point of the ECG waveform is taken as the feature, and a nearest neighbor is used as a classifier. They finally generated a sensitivity of 93.15% for normal beats and 91.07% for V arrhythmia. In paper [15], both linear (such as normalized low frequency power, normalized high frequency power, and so on) and nonlinear features (such as approximate entropy, hurst exponent, and so on) are used. They yield accuracy as high as 84.6%. In paper [16], the authors use higher order statistics (HOS) of the wavelet packet decomposition (WPD)

Table 2

Comparative results of different feature selection methods

Authors	Features/techniques	Classifiers	Arrhythmia	Accuracy (%)
Donna Giri et. al	Wavelet coefficients/ICA	Neural network	2 beats types	98.6
Kim et. al	linear and nonlinear feature	Different Classifiers	3 beats types	84.6
Amit and Shantanu	Local fractal dimension	Nearest neighbor algorithm	6 beats types	93.15(Sen)
Kutlu and Kuntalp	WPD coefficients/HOS	k-NN	5 beats types	90(Sen), 98(Spe)
In this work	Parameters of LMW	A-MLP	3 beats types	99.98

coefficients for the purpose of automatic heartbeat recognition. The classification accuracy of the proposed system is measured by an average sensitivity of 90%, average selectivity of 92% and average specificity of 98%.

As shown in Table 2, it is clear that, in comparison with other methods, the new methodology yields the highest possible accuracy in classification.

4. Conclusion

In this paper, we addressed the problem of extracting features of ECG signals. We presented a novel ECG model whose parameters have a clear medical sense. We introduced a local extract method to find the locations, widths and magnitudes. Through comparing the synthetic and real ECG signals, we observed that this approach performed well after having been improved. We also show that this medical information can be fed to classifiers for arrhythmia diagnosis. Perhaps the most important observation made is the following: this presented approach's success does not depend on how similar they are to normal heartbeats. Thus, it can be used for arrhythmia diagnosis. In the frequency-domain, the fitting LMW low-frequency coefficients can coincide well with those of the originals. This suggests that the LWM model could also be a potential method for ECG signal denoising or ECG compression.

However, there are some remaining constraints. The parameter extracted method is sensitive to prior knowledge. For some types of arrhythmias that miss some waves, the parameters can just represent noise information instead of real waves. Further work can also be conducted to improve the computational load of the parameters extracted method.

Acknowledgement

This paper is supported by the 'Fundamental Research Funds for the Central Universities', (HUST: CXY12Q023) and the ministry of science and technology of the People's Republic of China, within the framework of the project the CNC products innovation demonstration (Contract number: 2012BAF13B06)

References

[1] R.J. Martis, U.R. Acharya and L.C. Min, ECG beat classification using PCA, LDA, ICA and discrete wavelet transform, Biomedical Signal Processing and Control **8** (2013), 437–448.

[2] E. Luz, T.M. Nunes, V. de Albuquerque, J.P. Papa and D. Menotti, ECG arrhythmia classification based on optimum-path forest, Expert Systems with Applications **40** (2013), 3561–3573.

[3] B.M. Asl, A.R. Sharafat and S.K. Setarehdan, An adaptive backpropagation neural network for arrhythmia classification using R-R interval signal, Neural Network World **22** (2012), 535–548.
[4] L. Citi, E.N. Brown and R. Barbieri, A real-time automated point-process method for the detection and correction of erroneous and ectopic heartbeats, IEEE Transactions on Biomedical Engineering **59** (2012), 2828–2837.
[5] M.G. Tsipouras, V.P. Oikonomou, D.I. Fotiadis, L.K. Michalis and D. Sideris, Classification of atrial tachyarrhythmias in electrocardiograms using time frequency analysis, Computers in Cardiology **31** (2004), 245–248.
[6] M. Korurek and A. Nizam, Clustering MIT-BIH arrhythmias with ant colony optimization using time domain and PCA compressed wavelet coefficients, Digital Signal Processing **20** (2010), 1050–1060.
[7] R.J. Martis, U.R. Acharya, K.M. Mandana, A.K. Ray and C. Chakraborty, Cardiac decision making using higher order spectra, Biomedical Signal Processing and Control **8** (2013), 193–203.
[8] G. Swapna, U.R. Acharya, S. VinithaSree and J.S. Suri, Automated detection of diabetes using higher order spectral features extracted from heart rate signals, Intelligent Data Analysis **17** (2013), 309–326.
[9] M. Krishnan, S.V. Sree, D.N. Ghista, E. Ng, Swapna, A. Ang, K.H. Ng and J.S. Suri, Automated diagnosis of cardiac health using recurrence quantification analysis, Journal of Mechanics in Medicine and Biology **12** (2012), 1240014-1–1240014-24.
[10] W. Bucaoto, H.J. Kim and A. Lenskiy, Fractal analysis and the effect of aging on the heart rate and breathing frequency relationship, **260** (2011), 430–437.
[11] G.B. Moody and R.G. Mark, The impact of the MIT-BIH arrhythmia database, IEEE Eng. in Med. and Biol. **20** (2001), 45–50.
[12] Jiapu Pan and Willis J. Tompkins, A real-time QRS detection algorithm, IEEE Transactions on Biomedical Engineering **32** (1985), 230–236.
[13] D. Giri et al., Automated diagnosis of coronary artery disease affected patients using LDA, PCA, ICA and discrete wavelet transform, Knowledge-Based Systems, **37** (2013), 274–282.
[14] A.K. Mishra and S. Raghav, Local fractal dimension based ECG arrhythmia classification, Biomedical Signal Processing and Control, **5** (2010), 114–123.
[15] W.S. Kim et al., A study on development of multi-parametric measure of heart rate variability diagnosing cardiovascular disease, International Federation for Medical and Biological Engineering (IFMBE) **14** (2007), 3480–3483.
[16] Y. Kutlu and D. Kuntalp, Feature extraction for ECG heartbeats using higher order statistics of WPD coefficients, Computer Methods and Programs in Biomedicine, **105** (2012), 257–267.
[17] T. Tanantong, E. Nantajeewarawat and S. Thiemjarus, Toward continuous ambulatory monitoring using a wearable and wireless ECG-recording system: A study on the effects of signal quality on arrhythmia detection, Bio-Medical Materials and Engineering, **24** (2014), 391–404.