Clustering of Heartbeats from ECG Recordings Obtained with Wireless Body Sensors

A. Rashkovska*, D. Kocev ** and R. Trobec*
* Department of Communication Systems, Jožef Stefan Institute, Ljubljana, Slovenia
** Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
{aleksandra.rashkovska, dragi.kocev, roman.trobec}@ijs.si

Abstract - Long-term electrocardiographic (ECG) recordings can be beneficial for detection and diagnosis of heart diseases, in particular arrhythmias. A wireless multi-function biosensor that measures a potential difference between two proximal electrodes on the skin enables monitoring of vital functions, like heart rate, respiration and muscular activity. It can thus make long-term ECG measurements while users are performing their everyday duties and activities. These measurements are significantly longer and heterogeneous than the measurements performed in a controlled hospital environment. Consequently, their inspection for identification of different groups/clusters of heartbeats, either manual or computer supported, is obligatory. In this paper, we propose a method for automatic clustering of heartbeats from an ECG obtained with a wireless body sensor. We use state-of-the-art data mining methods for time series clustering - hierarchical agglomerative clustering in conjunction with dynamic time warping distance. The results show that the proposed methodology is robust and comparable to the classical Holter algorithms and therefore worth to be further evaluated.

I. INTRODUCTION

Long-term electrocardiographic (ECG) recordings are intended to help in detection or diagnosis of heart diseases. These measurements are significantly longer and heterogeneous than the measurements performed at a controlled hospital environment. Consequently, manual inspection of these recordings in order to identify different groups/clusters of heartbeats (that can be used for better describing the health status of the subject) is a tedious, hard and expensive job. An alternative is to use computational techniques for automatic clustering, like neural networks [1], data mining methods for clustering of ECG features [2] or time series clustering methods [3]. In this paper, we address the task of automatic heartbeat clustering using data mining methods for time series clustering.

Today the most standard ECG device used in medicine is the well-known 12-lead ECG, where wires are connected to the electrodes placed on 10 locations of the body. More long-term ECG recordings are usually acquired by a Holter monitor where reduced number of electrodes are connected with wires to a small portable recorder that acquires continuous ECG measurement throughout several days. On this type of equipment, more electrodes are used to obtain the signal, in which one of them serves as a reference for the others. Usually one to three wired leads are utilized for automatic tasks of ECG analysis. Related studies for ECG clustering utilize this type of measurements. However, to align with future trends in e-health, in this paper, we utilizes a unique ECG data acquisition that promises better future for long-term ECG recordings – unobtrusive measurement with a single ECG lead without wires. A wireless multi-function biosensor that measures a potential difference between two proximal electrodes on the skin enables monitoring of vital functions - heart activity and respiration [4].

In this paper, we propose a method for automatic clustering of heartbeats from an ECG obtained with a wireless body sensor. We exploit state-of-the-art data mining methods for time series clustering, namely hierarchical agglomerative clustering in conjunction with dynamic time warping distance. The obtained clusters will then be analyzed and used to derive meaningful heartbeat categories.

The remaining of this paper is organized as follows. Section II introduces the fundamental steps in heartbeat clustering and describes the selected methods for each step; Section III presents and discusses the results; and finally, Section IV concludes the paper.

II. METHOD

An automatic system for clustering of heartbeats from ECG recordings can be divided into four steps (see Fig. 1) as follows: 1) ECG data acquisition; 2) ECG signal preprocessing; 3) heartbeat segmentation; and 4) heartbeat clustering. The selection of methods for each of the four steps can have a crucial implication on the final result – the heartbeat clusters and consequently, the medical interpretation of such results.

A. ECG Data Acquisition

The data analyzed in this work has been obtained from unobtrusive ECG body sensor (dimensions: 2x9cm, weight: 14 g) with two electrodes at the distance of 8 cm. The placement of electrodes on the chest can be easily

The authors Aleksandra Rashkovska and Roman Trobec acknowledge the financial support from the Slovenian Research Agency under the grant P2-0095. Dragi Kocev is supported by the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (Grant number ICT-2013-612944).

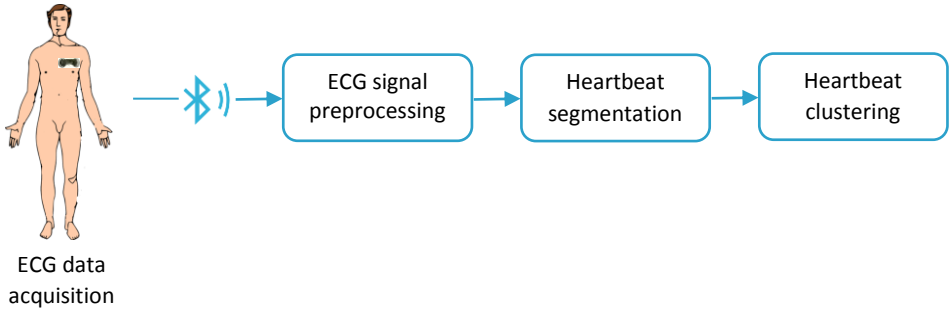

Figure 1. Diagram of the heartbeat clustering system.

fine-tuned to maximize the quality of the ECG recording [5]. Note that besides ECG, other features can be extracted from the measured potential, e.g., muscle activity and respiration [6]. In addition to the ECG, the sensor can also sense the information about the measurement conditions, e.g., movements and temperature, thus providing information that allow for ambient intelligence [7, 8]. A moderate sampling rate of 125 Hz with 10 bit analogue/digital converter is used as an optimum between medical value and amount of generated data. The sensor has a long autonomy (up to 7 days) [9], a low power wireless connection (BT4) to a Smartphone or other personal device, and a corresponding software for visualization and interpretation of measurements.

The ECG measurements obtained with the sensor are suitable for medical use, e.g., screening of patients with potential heart rhythm disturbances, reconstruction of the standard 12-chanell ECG [10], Holter-like investigations for long-term monitoring of patients after cardiac surgery, monitoring in cardio oncology [11], etc. The device can support solutions to every-day problems of the medical personal in hospitals, health clinics, homes for the elderly and health resorts. Its exceptionally lightweight design allows for unobtrusive use also during sports activities or during exhaustive physical work. Currently, there are only a few similar devices, but none has competing services on such advanced level [12-15]. A prototype of the wireless multifunctional body sensor is shown in Figure 2.

Figure 2. Prototype of the wireless multifunctional body sensor

B. *ECG Signal Preprocessing*

The preprocessing step usually concerns the noise removal from the ECG signal. The simplest and most widely used technique for reducing noise in ECG signals is the application of finite impulse response (FIR) digital filters [16]. Such filters perform well for the attenuation of known frequency bands, such as the noise coming from the electrical network (50 Hz or 60 Hz). However, filters must be applied with caution, since they can significantly distort the morphology of the ECG signal and make it unusable for cardiac diagnostic. Therefore, most state-of-the-art methods for ECG signal analysis do not even apply preprocessing on the signal.

In order to investigate the potential influence of the preprocessing task on the final clustering results, we examine scenarios with preprocessed and not preprocessed (raw) ECG signal. The applied preprocessing technique on the ECG signal is a low pass FIR filter with cutoff frequency set to 50 Hz.

C. *Heartbeat segmentation*

The heartbeat segmentation step refers to the detection of R peaks and generation of heartbeat segments from the ECG signal. Heartbeat segmentation methods have been studied for more than three decades [17, 18, 19, 20] and vary in complexity, i.e., from very simple methods to more elaborated ones. The method used for R peak detection in this work has been described and evaluated in [21]. The cut-off time for the segments is 0.2 *s* before the R peak time.

D. *Heartbeat Clustering*

Clustering is concerned with grouping objects into classes of similar objects [22]. Given a set of examples (object descriptions), the task of clustering is to partition these examples into subsets, called clusters. The goal of clustering is to achieve high similarity between examples within individual clusters (intra-cluster similarity) and low similarity between examples that belong to different clusters (inter-cluster similarity). Consequently, the notion of similarity (and conversely distance) is of a crucial importance in clustering.

A cluster is typically represented with a prototype, such as the mean/centroid or the medoid of the examples belonging in the cluster. The aim is to obtain compact clusters that have a low (intra-cluster) variance. Methods like k-means clustering or hierarchical agglomerative clustering (HAC) can be used to find sets of clusters with low intra-cluster variance and low inter-cluster similarity.

Here, we focus on HAC as a clustering method. HAC builds a hierarchy of clusters in a “bottom-up” approach: at the lowest level each example belongs in its own cluster and then pairs of clusters are merged together as one moves up the hierarchy. Typically, merging of the examples is done in a greedy manner [23].

The result of hierarchical clustering can be visually presented using a dendrogram (see Fig. 3). The dendrogram allows for cutting the hierarchy in two major

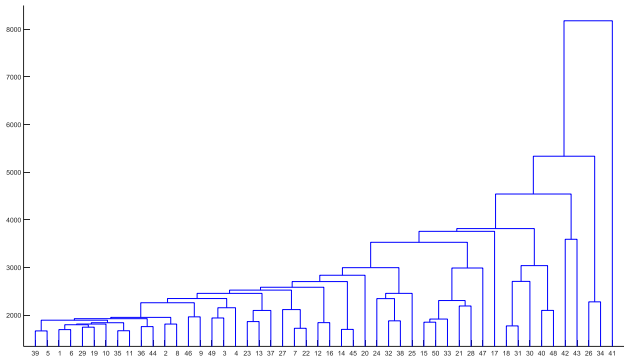

Figure 3. Dendrogram of hierarchical cluster tree

ways: one could specify the desired number of clusters or the maximal distance between two clusters to allow merging. This allows for closer inspection of the clusterings observed in the hierarchy of clusters.

A very important part of the HAC method is the definition and calculation of the distance between two clusters, i.e., the linkage function. There are several types of linkage that could be selected specifically for a given task. These different linkage types include maximum or complete linkage (maximum distance between any two examples from these clusters), minimum or single linkage (minimum distance between any two examples from the clusters), mean or average linkage (average distance between the examples of the two clusters), centroid linkage (the distance between the centroids/prototypes of the two clusters) and the decrease in variance for the cluster being merged (Ward's criterion). Here, we used average linkage for constructing the hierarchy of clusters.

In this work, we consider the heartbeat signals as time series, hence, for use methods for HAC clustering of time series data. More specifically, we use distance-based clustering methods that rely on the definition of an appropriate distance measure between two time series. There are several distance measures for time series that could be applied in this context. When selecting a distance, we should consider the properties of the data at hand and the properties of the specific distances.

For example, if the time series have equal length, then one can apply some standard distance measures such as the Euclidean distance or the Manhattan distance. These measures could, however, not be appropriate in this context because (1) they assume that the time series are synchronized, and (2) they mainly capture the difference in scale and baseline. Considering the properties of the heartbeat signal analyzed here, use of the above distances is not recommended: the duration of the heartbeats varies and the main interest is not as much the amplitude of the signal as its shape.

Considering the specific properties of the heartbeat signals (different length, shape over amplitude), we need to resort to more sophisticated distances that are able to capture the desired dynamics of the heartbeat signals. A distance that supports the required flexibility is the Dynamic Time Warping (DTW) distance [24]. The DTW distance "stretches" the time axis to obtain better matching between two time series. It accomplishes this by assigning multiple values of one of the time series to a single value of the other. This means that a time series that is delayed with respect to another one or otherwise temporally stretched (but maintains the approximate magnitude) will still be considered similar to the original series. If there is no stretching of the time axis, the DTW distance behaves as the Euclidean distance. Finally, the DTW distance can be used for time series of different length and not synchronized, which makes it a very flexible distance function.

In this work, we instantiate the parameters for the clustering algorithms as follows. As mentioned, we used HAC with average linkage implemented in Matlab. The DTW distance was calculated using the guidelines provided by Ratanamahatana and Keogh [25]: when calculating the distance between two time series, the window size is set to 10% of the length of the longer time series. For efficient calculation of the distance matrix between all of the heartbeats, we used the FastDTW implementation provided by Salvador and Chan [26].

III. RESULTS

A. *Experimental setup*

The ECG data analyzed in this work consists of four 30-minute measurements from a single volunteer, acquired during different activities. Therefore, each measurement is clustered separately. The statistics of the measurements are given in Table 1.

The dendrogram from the HAC clustering was cut to obtain 50 clusters. This high value is a precaution measure since the clustering method will cluster outliers in separate clusters (with very few examples), leaving the number of meaningful clusters reasonable. A cluster is considered meaningful if it contains more than 3 examples. The number of 3 was selected considering the length of the investigated measurements.

TABLE I. ECG MEASUREMENTS STATISTICS
Sinus – Normal sinus beat; SVES – Supra Ventricular Extra Systole; AF – Atrial Fibrilation

ID	**Activity**	**Mean BPM**	**# Beats**	**# Clusters [a]**		**Clustered beats [%]**		**Type of rhythm**
				Raw ECG	***Filtered ECG***	***Raw ECG***	***Filtered ECG***	
M1	walking	91.0	2727	21	23	98.57	98.72	**Sinus (< 1% SVES)**
M2	sitting	82.1	2462	20	19	98.62	98.33	**87%AF + 12%Sinus + 1%SVES**
M3	laying	57.3	1717	11	11	97.50	97.44	**82 % Sinus + 18 % SVES**
M4	sitting	64.1	1917	19	21	98.07	98.33	**Sinus (< 0.1% SVES)**

a. With more than 3 examples

B. *Raw vs. Filtered ECG*

The results showed that the numbers of meaningful clusters from raw and filtered ECG measurements do not differ much (see Table 1). Also, the percentage of clustered heartbeats from raw and filtered ECG does not differ much for each measurement. The percentage of clustered beats is around 98% for all measurements, regardless of the preprocessing. Visual inspection of the clusters showed that similar clusters are often generated from the raw and filtered ECG. Therefore, it can be concluded that the used method works well on the ECG signal without any preprocessing done beforehand. However, clusters from filtered ECG tend to be more coherent.

C. *Clustered ECG interpretation*

For visual inspection of the results, the clusters obtained from the filtered ECG measurements M1, M2, M3 and M4 are given in Fig. 4, 5, 6 and 7, respectively. The clusters are order by their size (number of examples). Only the clusters containing more than 3 examples are presented. For each cluster, the medoid heartbeat of the cluster is represented with black curve, accompanied with maximum 20 examples/heartbeats closest to the medoid represented with gray curves. The medoid is an example of the cluster whose average distance to the other examples is the smallest or, in other words, it is the example that is the closest to the "average" example in the cluster. The medoid is considered as a cluster representative/prototype in applications where mean is not attainable, which is the case in heartbeats clustering with DTW.

Results show that the clustering method identified different cardiac events in the measurements, even in the measurement acquired while walking. In all cases, the clusters with the highest number of examples clearly correspond to a normal ECG beat, except for the measurement M2. Namely, the P wave is not present in the ECG beats from the biggest cluster from M2, indicating atrial fibrillation. In M2, also some other smaller clusters, e.g., C11 and C5, can be identified as atrial fibrillation. The rest of the clusters (12%) from M2 belong to normal sinus beats, except for C6, C12 and C24, which indicate SVES (Supra Ventricular Extra Systole). The results for measurement M3 clearly identify clusters with SVES beats (around 18% - C2, C3, C16 and C25), namely clusters with the QRS looking very different from normal (the QRS has spikes). A small percent of SVES beats (< 1 %) is also present in M1, in particular, identified by C30, C20, C17, C38 and C11. Measurement M4 is clearly sinus ECG rhythm – all the clusters indicate normal ECG beat, except for a very small number of artifacts/outliers (C29 and C14).

The cardiac events identified in each measurement are also summarized in Table 1. For the ultimate medical interpretation of the clusters, a medical professional should be consulted.

IV. CONCLUSION

In this paper, we propose a method for automatic clustering of heartbeats. Unlike most related studies for ECG clustering which utilize long-term ECG measurements from 12-lead ECG or Holter monitor, here we utilize a unique ECG data acquisition - noninvasive measurement with a single channel of bipolar ECG without wires. After appropriate preprocessing and segmentation of the ECG measurements, we perform hierarchical agglomerative clustering using Dynamic Time Warping (DTW) as distance measure.

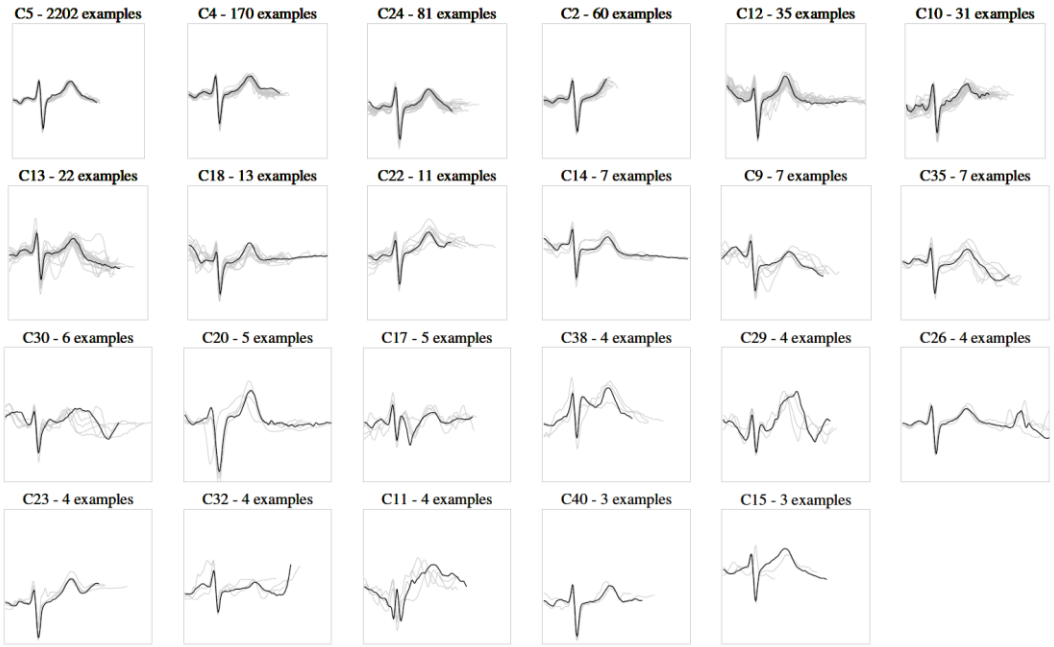

Figure 4. Clusters from filtered measurement M1. Black line represents medoid, gray lines represent maximum 20 closest examples to the medoid.

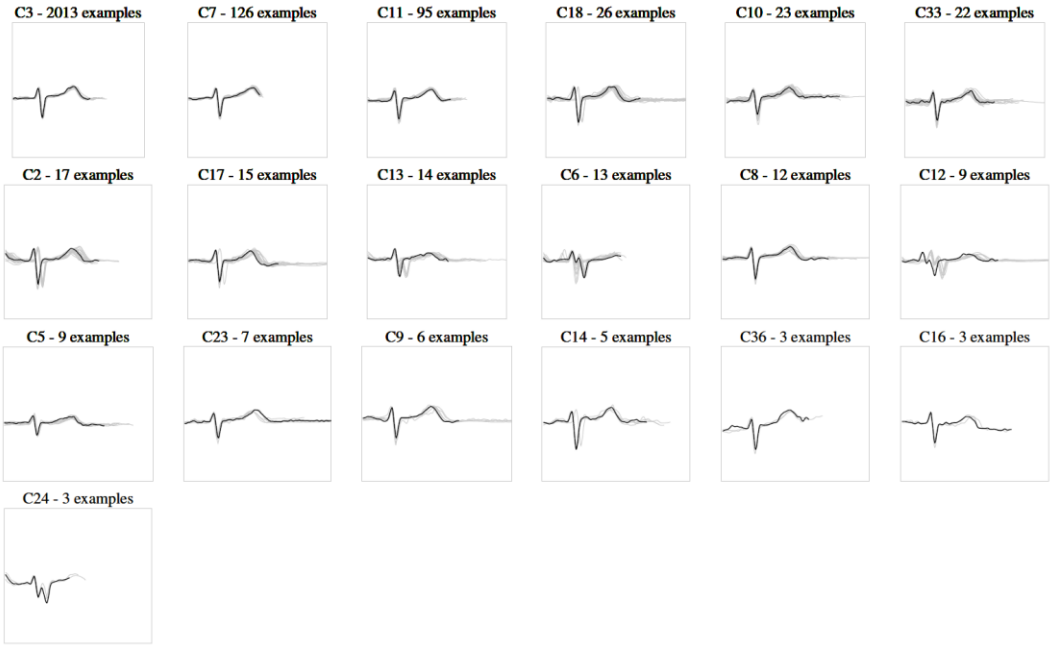

Figure 5. Clusters from filtered measurement M2. Black line represents medoid, gray lines represent maximum 20 closest examples to the medoid.

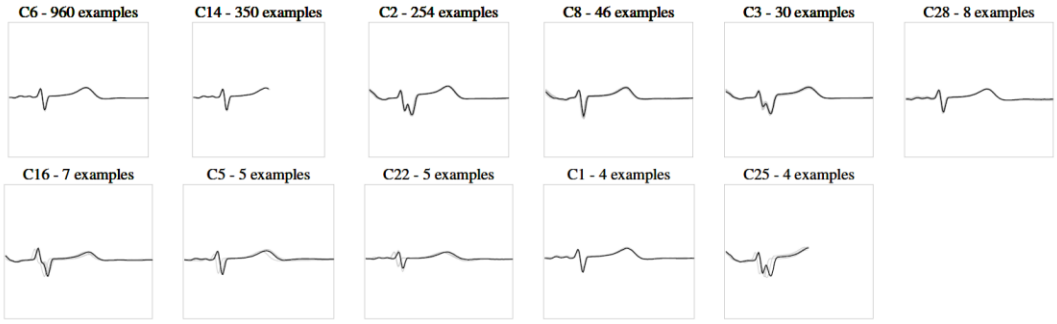

Figure 6. Clusters from filtered measurement M3. Black line represents medoid, gray lines represent maximum 20 closest examples to the medoid.

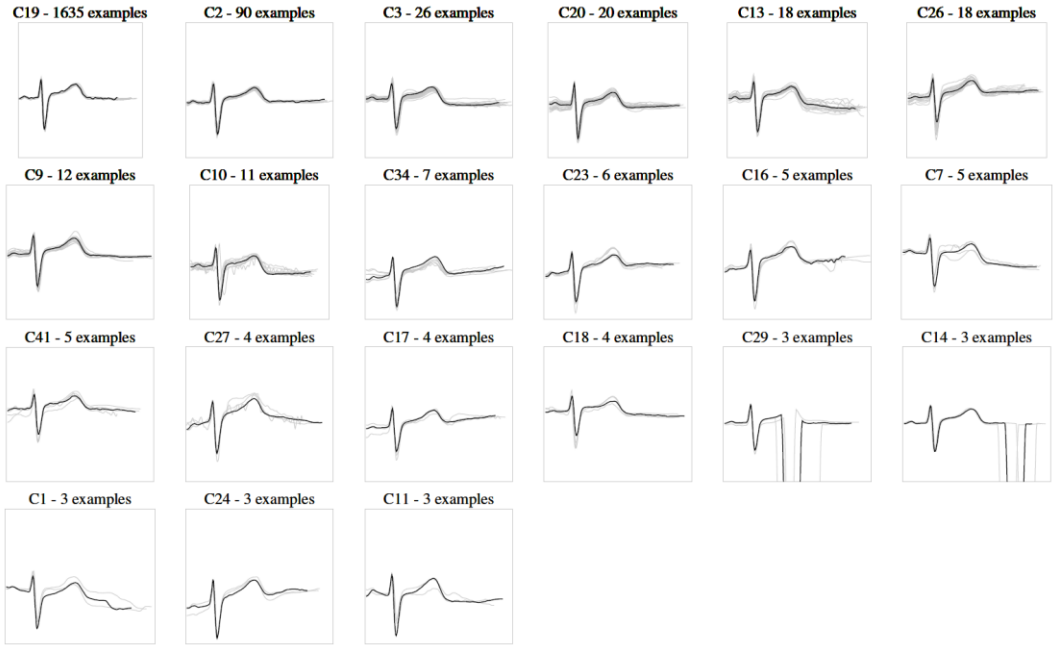

Figure 7. Clusters from filtered measurement M4. Black line represents medoid, gray lines represent maximum 20 closest examples to the medoid.

The results show that we were able to identify meaningful clusters, confirming that the clustering from a differential ECG can provide an interesting insight into the medical condition based on the measurements. However, these methods should be accounted only as assistance for the medical professional and should not be trusted blindly.

In further work, we plan to investigate different segmentation techniques and their potential influence on the final clustering results. Furthermore, other distances over time series considering the specifics of heartbeat signals will be designed and explored, as well as other clustering methods, including k-medoids and predictive clustering trees. The obtained clustering results will then form the base for annotation of the ECG beasts as a pre-step for the future task of ECG beats classification.

REFERENCES

[1] M. Lagerholm, C. Peterson, G. Braccini; L. Edenbrandt; and L. Sornmo, "Clustering ECG Complexes Using Hermite Functions and Self-Organizing Maps," IEEE Transactions on Biomedical Engineering, vol. 47, pp. 838–848, August 2002.

[2] A.Vágner, L. Farkas, and I. Juhász, "Clustering and Visualization of ECG Signals," Advances in Intelligent and Soft Computing, vol. 101, Editors: D. Dicheva, Z. Markov, E. Stefanova, Eds. Springer, 2011, pp. 47-51.

[3] J. R. Annam, S. S. Mittapalli, and R. S. Bapi, "Time series Clustering and Analysis of ECG heart-beats using Dynamic Time Warping," Proceedings of INDICON 2011, Annual IEEE India Conference, December 16-18, 2011, pp. 1-3.

[4] A. Rashkovska, I. Tomašić, K. Bregar and R. Trobec, "Remote Monitoring of Vital Functions - Proof-of concept System," Proceedings of MIPRO 2012, 35th International Convention, May 21-25, 2012, Opatija, Croatia, pp. 446–450.

[5] I. Tomašić, S. Frljak, and R. Trobec, "Estimating the universal positions of wireless body electrodes for measuring cardiac electrical activity," IEEE Transactions on Bio-medical Engineering, vol. 60, pp. 3368–3374, December 2013.

[6] R. Trobec, A. Rashkovska, and V. Avbelj, "Two proximal skin electrodes - a respiration rate body sensor," Sensors, vol. 12, pp. 13813–13828, October 2012.

[7] R. Trobec, V. Avbelj, and A. Rashkovska, "Multi-functionality of wireless body sensors," The IPSI BgD transactions on internet research, vol. 10, pp. 23–27, January 2014.

[8] H. Gjoreski, A. Rashkovska, S. Kozina, M. Luštrek, M. Gams, "Telehealth using ECG sensor and accelerometer," Proceedings of MIPRO 2014, 37th International Convention, May 26-30, 2014, Opatija, Croatia, pp. 270–274.

[9] K. Bregar and V. Avbelj, "Multi-Functional Wireless Body Sensor Analysis of Autonomy," Proceedings of MIPRO 2013, 36th International Convention, May 20-24, 2013, Opatija, Croatia, pp. 322–325.

[10] R. Trobec and I. Tomašić, "Synthesis of the 12-lead electrocardiogram from differential leads," IEEE Transactions on Information Technology in Biomedicine, vol. 15, pp. 615–621, July 2011.

[11] J. M. Kališnik et al., "Mobile health monitoring pilot systems," Proceedings of IS 2015, 18th International Multiconference Information Society, October 9-12, 2015, Ljubljana, Slovenia, pp. 62–65.

[12] E. Fung et al., "Electrocardiographic patch devices and contemporary wireless cardiac monitoring," Front Physiol., vol. 6, May 2016, doi: 10.3389/fphys.2015.00149.

[13] M. Hernandez-Silveira et al. "Assessment of the feasibility of an ultra-low power, wireless digital patch for the continuous ambulatory monitoring of vital signs," BMJ Open, vol. 5, May 2015, doi:10.1136/bmjopen-2014-006606.

[14] M. Etemadi et al., "A Wearable Patch to Enable Long-Term Monitoring of Environmental, Activity and Hemodynamics Variables," IEEE Transactions on Biomedical Circuits and Systems, in press.

[15] I. H. Hansen, K. Hoppe, A. Gjerde, J. K. Kanters, and H. B. Sorensen, "Comparing Twelve-lead Electrocardiography with Close-To-Heart Patch Based Electrocardiography," Proceedings of IEEE EMBC 2015, 37th Annual International Conference, August 25-29, 2015, Milan, Italy, pp. 330–333.

[16] P. Lynn, "Recursive digital filters for biological signals," Medical and Biological Engineering and Computing, vol. 9, pp. 37–43, 1979.

[17] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," IEEE Transactions on Biomedical Engineering, vol. 32, pp. 230–236, March 1985.

[18] V. X. Afonso, W. J. Tompkins, T. Q. Nguyen, and S. Luo, "ECG beat detection using filter banks," IEEE Transactions on Biomedical Engineering, vol. 46, pp. 192–202, February 1999.

[19] . C. Yeh and W. J. Wang, "QRS complexes detection for ECG signal: The difference operation method," Computer Methods and Programs in Biomedicine, vol. 91, pp. 245–254, September 2008.

[20] O. Sayadi and M. B. Shamsollahi, "A model-based bayesian framework for ECG beat segmentation," Physiological Measurement, vol. 30, pp. 335–352, March 2009.

[21] V. Avbelj, R. Trobec, and B. Gersak, "Beat-to-beat repolarisation variability in body surface electrocardiograms," Med. Biol. Eng. Comput., vol. 41, pp. 556–560, September 2003.

[22] L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis," 1st Edition, Wiley-Interscience, 1990.

[23] L. Rokach and O. Maimon, "Clustering methods," in Data mining and knowledge discovery handbook, Springer US, 2005, pp. 321–352.

[24] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition, " in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no. 1, pp. 43-49, 1978.

[25] C. A. Ratanamahatana and E. Keogh, "Three Myths about Dynamic Time Warping," Proceedings of SDM '05, SIAM International Conference on Data Mining, April 21-23, 2005, Newport Beach, CA, pp. 506–510.

[26] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," Intelligent Data Analysis, vol. 11, pp. 561–580, October 2007.