ELSEVIER

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Discovering dangerous patterns in long-term ambulatory ECG recordings using a fast QRS detection algorithm and explorative data analysis

*Matteo Paoletti**, *Carlo Marchesi*

Department of Systems and Computer Science, BIM Laboratory, University of Florence, Via S. Marta 3, 50100 Florence, Italy

ARTICLE INFO

Article history:
Received 26 April 2005
Received in revised form 16 January 2006
Accepted 17 January 2006

Keywords:
ECG analysis
QRS detection
Long-term ECG monitoring

ABSTRACT

The next two decades will see dramatic changes in the health needs of the world's populations with chronic diseases as the leading causes of disability, according to recent World Health Organization reports. Increases in the senior population living "confined" in domestic area are also expected producing a steep increase in the need for long-term monitoring and home care services. Independently of the particular features and specific architectures, long-term monitoring systems usually produce a large amount of data to be analyzed and inspected by the practitioners and in particular by the cardiologists dealing with ECG recordings analysis. This problem is well known and regards also the traditional holter-based practice. In this paper we present a program for discovering patterns in ECG recordings, to be considered as a medical decision-making support. Computational methods are based on a QRS detector especially designed for noisy applications followed by a parameters space reduction operated by the KL transform modified on a "user-fit" basis. Events characterization is based on a recently introduced clustering method, called KHM (K-harmonic means). The most representative beat families and the corresponding prototypes (physiological and pathological) are then presented to the user through appropriate graphics to facilitate an easy and fast interpretation. We tested the QRS detection algorithm using the MIT-BIH arrhythmia database. Our method produced 565 false positive beats and 379 false negative beats and a total detection failure of 0.85% considering all the 109.809 annotated beats in the database. While a clinical experimentation of our program is on the way, we used the VALE Database to perform a preliminary evaluation of the methods used for data exploration (PCA, KHM). Considering the entire database, we succeeded in identifying pathological clusters in 97% of the cases.

1. Introduction

According to recent WHO reports, [1] the next two decades will see dramatic changes in the health needs of the world's populations with chronic diseases, mental illness and injuries as the leading causes of disability. Increases in the senior population are also expected in many countries. Those factors, together with social and economical reasons, will increase the population of aged people living "confined" in domestic area producing a steep increase in the need for long-term monitoring and home care services [2–4] on demand.

* *Corresponding author*. Tel.: +39 055 4796255.
E-mail address: paoletti@dsi.unifi.it (M. Paoletti).

0169-2607/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2006.01.005

We think that development of personal devices, especially designed for data acquisition, elaboration and network communication [5–7], will play a pivotal role in order to build a bridge among chronic patient home, family practitioner and hospital specialist.

According to many researchers [8,9], long-term monitoring units designed for home care have not yet reached a technical level that is widely accepted by patients, their families and practitioners. Such long-term, home care units must be compact, comfortable to wear and they must be designed for low power consumption. Furthermore, they must be able to detect signals reliably and stably in the face of motion artifacts and noise, to process those signals and to transmit data to the assistance network when necessary. Recently many projects have been launched in order to design and develop home-based, long-term monitoring systems [10].

Independently of the particular features and specific architectures, those systems usually produce a large amount of data to be analyzed and inspected by the practitioners and in particular by cardiologists.

Long-term monitoring of vital signals is a technique well established to assess patient conditions in a number of diseases. It provides critical information for long-term assessment and preventive diagnosis for which long-term trends and signal patterns are of special importance. Such trends and patterns are often difficult to identify using traditional examinations. Those cardiac problems that occur frequently during normal daily activities may disappear the moment the patient is hospitalized, causing high costs, diagnostic difficulties, and also possible errors [8].

In this paper we present a method for discovering and exploring patterns in ECG recordings, to be considered as a medical decision-making support.

First, we analyze the recordings and classify heart-beat on a morphological basis. Next, the representative beat families (physiological and pathological) are presented to the user through appropriate graphics to facilitate an easy and fast interpretation.

Signal segmentation is based on a QRS detector especially designed for noisy applications (ambulatory recordings). The Karhunen–Loève (KL) transform provides an efficient parameters space dimension reduction while clustering is performed exploiting a recently introduced algorithm called KHM [11], designed for data mining applications, which is essentially insensitive to the initialization of the class centers and shows a rapid convergence.

Finally, to test the algorithms we analyzed the MIT-BIH arrhythmia Database and the ST-T/VALE DB European Society of Cardiology ST-T database [12,13]. Experiments and tests were done using *Mathcad* programming facilities while, after validation, the algorithms were implemented using C++ and standard mathematic libraries.

2. Computational methods and algorithms

Biological signal processing is usually developed through an appropriate DSP algorithms cascade (see Fig. 1). For ECG signals the main stage often consists in a segmentation algorithm, usually preceded by a pre-processing stage (noise reduction, signal quality enhancement, baseline subtraction), such as to include characteristic wave shapes to be detected [14]. Wave shapes detection generally is based on enhancing slope variations with respect to a reference level. For instance the elements to be identified in ECG raw data are the wave shapes of the QRS complex.

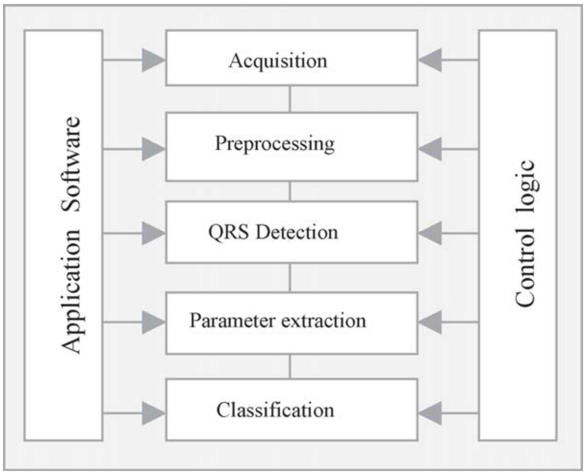


Fig. 1 – The DSP algorithms cascade.

Such detection is traditionally done with digital filter design techniques, both in time and frequency domain [14,15].

In this paper we present a time-domain algorithm for QRS recognition, characterized by a low computational cost and showing good performances also in presence of low-SNR recordings.

2.1. *QRS detection*

QRS detection algorithms are usually based on enhancing some features characterizing the main wave shapes and allowing a threshold detection [14,15]. In this study we were interested in a fast and reliable algorithm to analyze long-term and noisy ambulatory records. We thought that some indicators of the *length* of the ideal curve, representing the QRS complex, could be useful in order to design a time domain detector with the desired features.

2.1.1. *Algorithm overview*

In this section we propose a QRS detection method especially designed for noisy applications by exploiting the simple *curve-length concept* [16]. In Fig. 2, two examples of ECG signal segments (series “a” and “b”) are displayed. The figure shows how the lengths L1 and L2 characterize the local shape of the signal, given a certain time interval T_r. Our operator works in the time domain calculating, as output, a quantity derived by the length of successive overlapping ECG signal segments. In the discrete time domain we can calculate the length of the i-th segment as:

$$U_i = \sum_{j=1}^{N} \sqrt{T_x^2 + (y_{i+j} - y_{i+j-1})^2} \tag{1}$$

where T_x is the sampling interval, $(y_{i+j} - y_{i+j-1})$ represents the j-th ECG signal increment, and N is the number of samples in the width of the i-th calculus window.

Being T_x a constant in each segment we approximate the successive segments lengths U_i with the quantity:

$$U1_i = \sum_{j=1}^{N} \sqrt{(y_{i+j} - y_{i+j-1})^2} = \sum_{j=1}^{N} |y_{i+j} - y_{i+j-1}| \tag{2}$$

Since the calculus window duration is constant for all the overlapping windows, the quantity $U1$ is larger (considering the absolute value) for ECG segments containing strong slope variations (i.e. QRS complexes) and tends to be smaller in presence of flat segments or slower waves (P, T waves). To enhance this effect and to make all the output values positive (this feature is useful for the following thresholds-based detection stage) we introduce a nonlinear amplification of the output, squaring the differences between consecutive points:

$$U2_i = \sum_{j=1}^{N} (y_{i+j} - y_{i+j-1})^2 \tag{3}$$

Finally, in order to make the operator robust against high frequency noise we consider alternating point's differences instead of consecutive point's differences; this corresponds to operate a sort of smoothing task on the signal derivative before the squaring process. The final expression of this simple operator is:

$$U3_i = \sum_{j=2}^{N} (y_{i+j} - y_{i+j-2})^2 \tag{4}$$

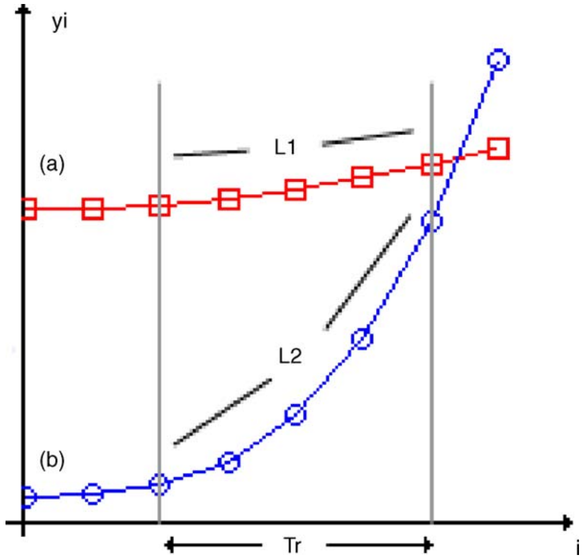


Fig. 2 – The simple *curve-length* concept. The figure shows two examples of ECG signal segments (series "a" and "b"). The lengths of the segments L1 and L2 characterize the local shape of the two signals, given a certain time interval T_r.

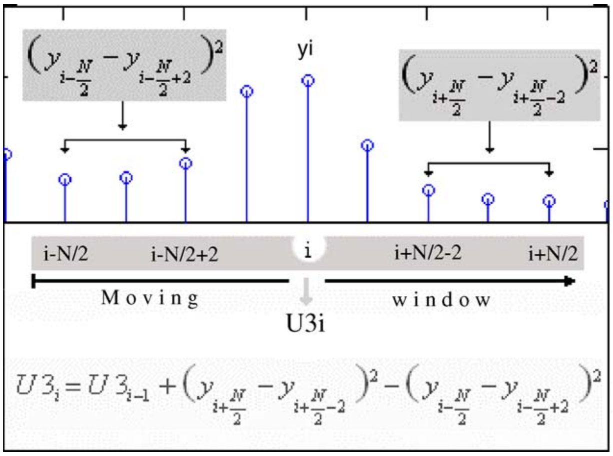

$\left(y_{i-\frac{N}{2}} - y_{i-\frac{N}{2}+2}\right)^2$
yi
$\left(y_{i+\frac{N}{2}} - y_{i+\frac{N}{2}-2}\right)^2$
i-N/2
i-N/2+2
i
i+N/2-2
i+N/2
Moving window
U3i
$U3_i = U3_{i-1} + \left(y_{i+\frac{N}{2}} - y_{i+\frac{N}{2}-2}\right)^2 - \left(y_{i-\frac{N}{2}} - y_{i-\frac{N}{2}+2}\right)^2$

Fig. 3 – The *U3* operator is implemented with a low computational cost and recursive expression. Each time the output is calculated, the new term $\left(y_{i+(N/2)} - y_{i+(N/2)-2}\right)^2$ is added to the old *U3* value while the term $\left(y_{i-(N/2)} - y_{i-(N/2)+2}\right)^2$ is subtracted. Each new output value is calculated performing six operations. Generally the moving window width *N* should be the same as the estimated duration of the episode to be detected, in this case the QRS complex. We used a window 100 ms wide.

In practice, the operator can be implemented with a low computational cost recursive expression (see Fig. 3 for details):

$$U3_0 = \sum_{j=2}^{N} (y_j - y_{j-2})^2 \tag{5}$$

$$U3_i = U3_{i-1} + (y_{i+(N/2)} - y_{i+(N/2)-2})^2 - (y_{i-(N/2)} - y_{i-(N/2)+2})^2 \tag{6}$$

Generally the moving window width should be the same as the estimated duration of the episode to be detected, in this case the QRS complex. We determined the duration of the window empirically. We used a window 100 ms wide.

In Fig. 4 we can observe some examples of the operator responses applying different ECG signals as inputs. Data records are provided by European ST-T/Vale databases [12].

As we can see in the figure, the operator $U3$ heavily attenuates all the perturbations different from QRS waveforms. Since the operator itself improves SNR also in presence of baseline shifts and high frequency noise, we didn't apply any filter during the preprocessing phase. We simply normalized [0, 1] the input signal, using non-overlapping windows (60 s wide), to limit the magnitude of the output.

2.1.2. *Comparison with other QRS detection algorithms*

A well known procedure producing similar output waveforms, but obtained with a different approach, was published by Pan and Tompkins [17] and consisted of a cascade of four filters: a band pass, a differentiator, squaring operation, and finally a moving window integrator. The authors implemented the following detection stages using a dual-threshold algorithm to find QRS complexes indexes and a control-logic to detect the eventually missed beats [17]. In order to val-

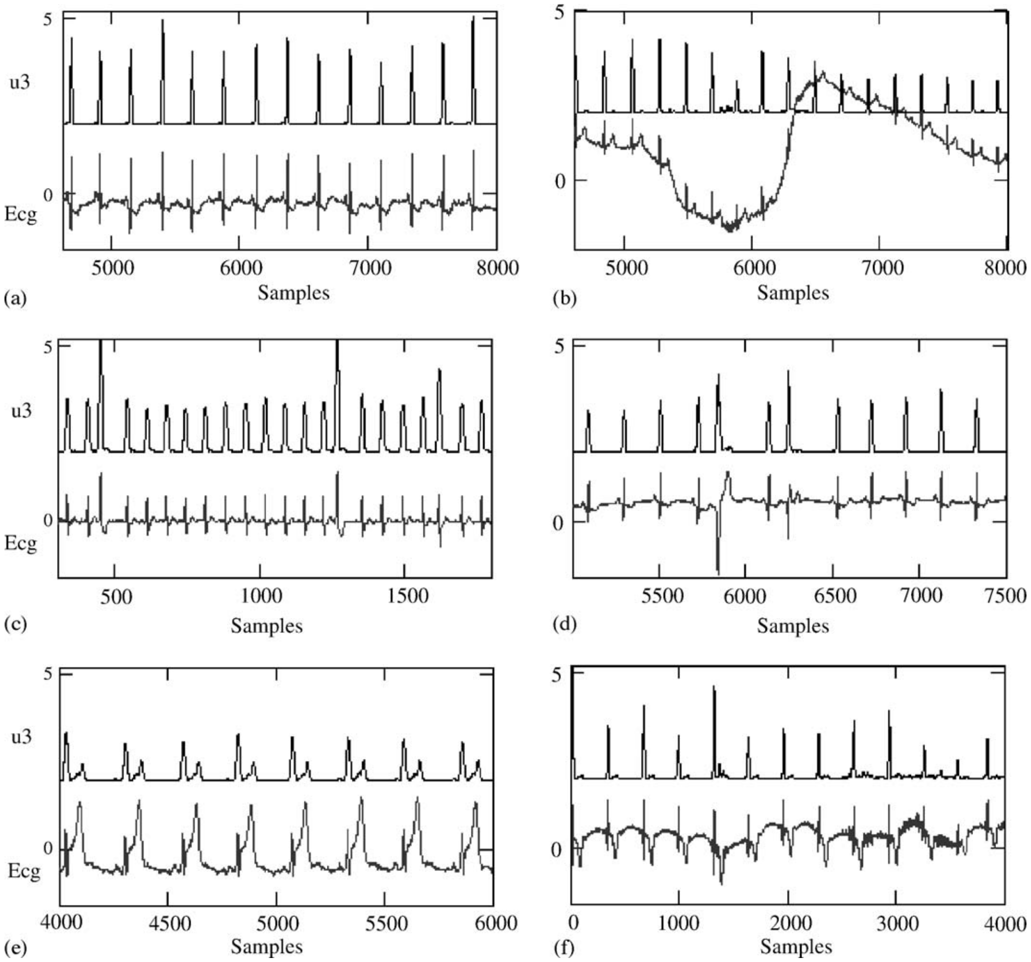


Fig. 4 – The *U*3 operator is able to operate also in presence of abnormal beat morphologies. The figure shows the ECG input signals and *U*3 responses in various conditions: (a) Normal sinusal rhythm, absence of noise; (b) normal sinusal rhythm, baseline shifts; (c and d) ectopic beats (PVC); (e) ischemic attack; (f) high frequency noise. Data were provided by Physionet [18], European ST-T/Vale Database [12,27].

idate the operator *U*3, we analyzed the MIT-BIH arrhythmia database [13] comparing the performance of our method with the results obtained by Pan and Tompkins with the same database [17]. The MIT-BIH arrhythmia database [13], was provided by Physionet [18] and consists of 48 half-hour excerpts of two-channel ambulatory ECG recordings, obtained from 47 subjects studied by the BIH Arrhythmia Laboratory between 1975 and 1979. The recordings were digitized at 360 samples per second per channel with 11-bit resolution over a 10 mV range. Two or more cardiologists independently annotated each record; disagreements were resolved to obtain the computer-readable reference annotations for each beat (approximately 109.000 annotations in all) included with the database

In particular, we applied the *U*3 operator to the first channel of each record in the MIT-BIH database. To extract the QRS complex indexes from the obtained output signals we used the same dual-threshold detection technique used by Pan and Tompkins [17] and the calculated markers were compared with the annotations in the database. Table 1 summarizes the results obtained with *U*3. A similar table showing the results obtained by Pan and Tompkins were presented in [17]. Our algorithm produced 565 false positive beats and 379 false negative beats and a total detection failure of 0.85% considering all the 109.809 annotated beats in the database. The algorithm proposed by Pan and Tompkins produced 507 false positive beats and 277 false negative beats for a total failure of 0.67%. They calculated the total failure percentage considering 116.137 annotated beats. An accurate check of the results published by Pan–Tompkins has demonstrated that the table published in the original paper apparently contains an error, in fact the total beats sum up to 109.809 instead of 116.137 declared by the authors. This difference should not influence very much the performance figures, but in any case considering 109.809 annotated beats, the total failure rate produced by Pan–Tompkins method is of 0.71%.

Although both algorithms show similar performances (both failure rates are <1%) in order to detect QRS complexes, *U*3 has the advantage to be faster, in fact, as we can see from Eq. (6), each sample of the output is obtained with six operations, while a larger number of operations are needed to calculate the output with the cascade of four filters proposed

Table 1 – Results obtained evaluating the *U*3 operator with the MIT-BIH arrhythmia database

Record no.	Total beats	FP (beats)	FN (beats)	Failed detection (beats)	Failed detection (%)
100	2273	0	0	0	0
101	1865	7	5	12	0.64
102	2187	0	0	0	0
103	2084	0	0	0	0
104	2230	2	2	4	0.17
105	2572	73	31	104	4.04
106	2027	7	2	9	0.44
107	2137	2	2	4	0.18
108	1763	207	19	226	12.81
109	2532	0	2	2	0.07
111	2124	2	0	2	0.09
112	2539	0	3	3	0.11
113	1795	0	0	0	0
114	1879	3	21	24	1.27
115	1953	0	0	0	0
116	2412	4	22	26	1.07
117	1535	1	3	4	0.26
118	2275	1	1	2	0.08
119	1987	3	1	4	0.20
121	1863	5	7	12	0.64
122	2476	1	3	4	0.16
123	1518	0	0	0	0
124	1619	0	0	0	0
200	2601	6	7	13	0.49
201	1963	2	17	19	0.96
202	2136	0	5	5	0.23
203	2982	57	41	98	3.28
205	2656	0	7	7	0.26
207	1862	5	5	10	0.53
208	2956	5	21	26	0.87
209	3004	4	2	6	0.19
210	2647	3	10	13	0.49
212	2748	0	0	0	0
213	3251	5	7	12	0.36
214	2262	3	7	10	0.44
215	3363	0	5	5	0.14
217	2208	4	8	12	0.54
219	2154	0	0	0	0
220	2048	0	0	0	0
221	2427	3	1	4	0.16
222	2484	112	92	204	8.21
223	2605	2	2	4	0.15
228	2053	26	9	35	1.70
230	2256	3	1	4	0.17
231	1886	0	0	0	0
232	1780	7	5	12	0.67
233	3079	0	3	3	0.09
234	2753	0	0	0	0
Total	109809	565	379	944	0.85

by Pan and Tompkins (see Table 2 for details). Since in our application we needed a fast algorithm to analyze long-term recordings, we preferred to include *U*3 for QRS detection in the final version of our system prototype instead of the well-known procedure introduced by Pan and Tompkins.

2.2. *Data exploration*

After applying the *U*3 operator and the same dual-threshold detection technique used by Pan and Tompkins [17], an intersection point X_p is calculated for each event (see Fig. 5). Then, to operate a proper segmentation, we proceed calculating the "initial" point (I_{on}), the end point (I_{off}), the *R*-peak index (R_x) and the barycentre index (B_x) of each complex. Those temporal references can be calculated exploiting directly the *U*3 output, as we can see in Fig. 5.

After QRS detection and signal segmentation we proceed calculating the heart rate as the time elapsed between consecutive ventricular contractions (the program calculates the differences between consecutive R_x point indexes). Since B_x points can also provide a suitable reference for *RR* series calculus, we let this choice to the user as a program option. Also the ST-level series are calculated with well-known techniques [19,20].

Table 2 – Number of operations needed to calculate one output value: a comparison between the solution proposed by Pan–Tompkins and the *U*3 quadratic operator

*U*3	Pan–Tompkins
4 + 2	Low-pass stage 4 + 2 High-pass stage 3 + 1 Derivative stage 4 + 1 Squaring 1 Integration *N* + 1
Total 6 Op.	Total 17 + *N* Op.

N represents the number of samples in the width of the moving integration window of Pan–Tompkins procedure [17].

Next the program performs the following steps:

(a) First, a subset of raw data (at least a 60-s wide ECG segment) not containing pathologic events (reference condition) is selected by the user from the record.

(b) The program extracts a segment from each heart-beat using the I_{on} and I_{off} indexes as references. The user can select the specific feature to be analyzed (QRS complex, ST-T segment, whole beat). The segments samples are then inserted into a data matrix M: each row represents one segment. Fig. 6 shows an example of this matrix M, obtained extracting 50 normal QRS complexes from a reference ECG segment.
If a QRS or a whole beat analysis is selected the segments are aligned into the matrix according to the R_x points, otherwise the ST-T segments are simply inserted into the matrix rows considering I_{off} as the starting point.

(c) The principal components [21] (PC) of the data matrix are calculated. Next, using the first three components all the beats in the record are projected in a 3D space performing the Karhunen–Loève (KL) transform. Only the first three vectors (PC) are considered because, in our experiments, they capture a sufficient percentage (>90%) [21] of the total original variance. To calculate the absorbed variance by the first three components we used the following equation [21]:

$$t_3 = 100\frac{\sum_{j=1}^{3}\lambda_j}{\sum_{j=1}^{N}\lambda_j} \qquad (7)$$

where t_3 is the ratio of variance represented by the first three components, λ_j the eigenvalues of the covariance matrix, and N is the total number of the eigenvalues. We found that, for all the analyzed recordings, the first three components represented a sufficient percentage of the original variance ($t_3 \geq 90\%$), however if the program finds t_3 <90% a warning is given to the user. The calculus of the orthonormal base (principal components) is performed only one time (with reference data), in fact, the remaining ECG beats/beat-segments (M rows) are transformed simply multiplying the samples by the three base vectors.

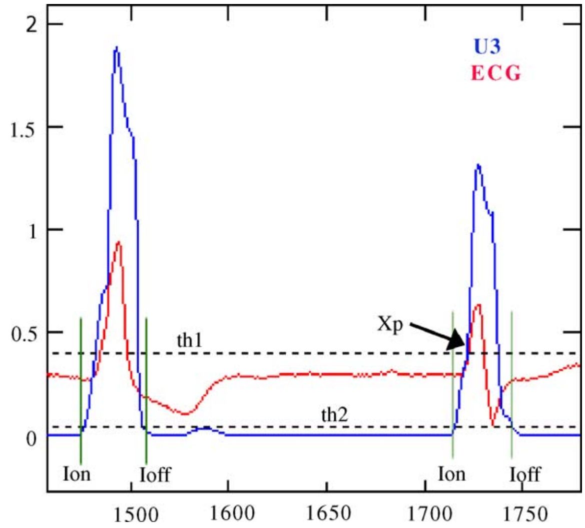


Fig. 5 – The figure shows the ECG signal, *U*3 output, and the detection thresholds. The primary threshold (th1) is used to detect QRS complexes, while a secondary threshold (th2) is used to calculate I_{on} and I_{off} indexes.

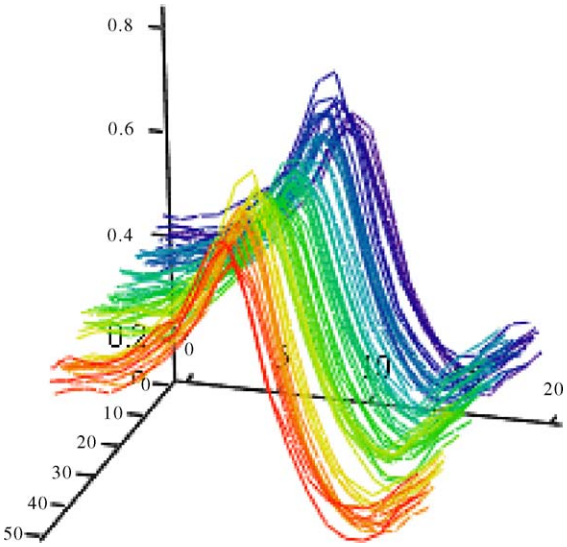
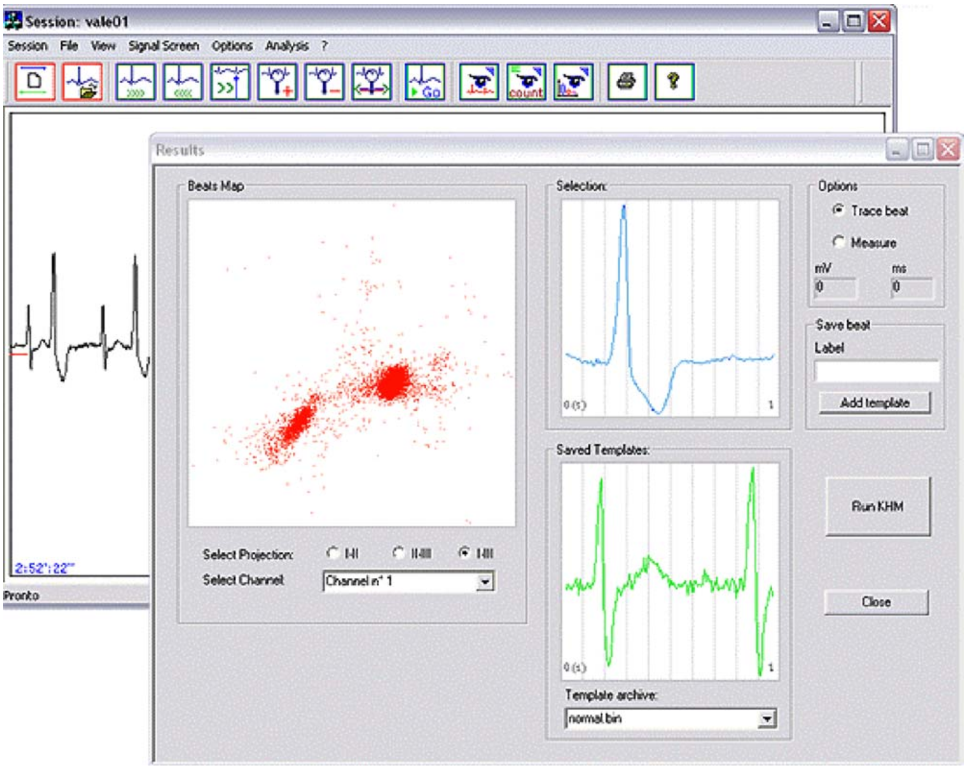


Fig. 6 – After a subset of raw data (at least a 60-s wide ECG segment) not containing pathologic events (reference condition) is selected by the user from the record, the program extracts a segment from each heart-beat using the I_{on} and I_{off} indexes as references. The user can also select the specific feature to be analyzed (QRS complex, ST-T segment, whole beat). The segments samples are then inserted into a data matrix M where each row represents one segment. In this example, the matrix M is obtained extracting 50 normal QRS complexes from the reference ECG segment indicated by the user.

Finally, a visualization of data points, each representing one beat in the principal component space, is provided to the user through appropriate 2D graphics. Fig. 7 shows a program window that allows the user to trace the various beats, selecting the desired projection (i.e. Components I–III in the figure). When the user clicks over one point in the *beats map* (left panel) the program selects the corresponding beat in the original ECG



Select Projection: ○ I-II ○ II-III ● I-III

Select Channel: Channel n° 1

Options: ● Trace beat ○ Measure

mV: 0

ms: 0

Template archive: normal.bin

Fig. 7 – The figure shows the main window of the program (in background) and the *exploration window* (foreground) that allows the user to trace the various beats in the reduced PC space displayed in the *beats map panel*. Through the radio buttons below it's possible to select the desired projection (i.e. Components I–III in the figure). When the user clicks one point in the *beats map panel* (left panel), the program selects the corresponding beat in the original ECG record and visualizes it in the *Selection Panel* (on the right), while the main window (in background) automatically shows the 30 s window containing the selected beat. The user can also assign a label to the selected ECG beats and store them as *reference templates* for further investigations. The saved templates can also be useful to make a comparison with new selected beats and represent a useful collection of waveforms to characterize the specific patient's profile. Through the *Saved Templates Panel,* the user visualizes all the stored beats in the current work-session. Selecting the various points in the main clouds and looking at the corresponding waveforms in the *Selection Panel*, the user can visually explore the most important morphological families that characterize the whole record without having to scroll it sequentially.

record and visualizes it in the *Selection Panel,* while the main window (in background) automatically shows the 30 s window containing the selected beat. The program allows the user to save and label the selected ECG beats for further investigations or to make a comparison with new selected beats. The saved templates can be viewed in the right panel below the *Selection Panel*. In Fig. 7 we can see the program window for visual exploration showing the *beats map* obtained analyzing the record "Vale01.dat" (European/ValeDB database [12]). In this case the ECG beats seem to be grouped in two different clusters, some outliers are also visible. Selecting the various points in the main clouds and looking at the corresponding waveforms in the *Selection Panel,* the user can visually explore the most important morphological families that characterize the whole record without having to scroll it sequentially. In particular, considering the "Vale01" record, we have two main groups, one corresponding to ectopic beats (premature ventricular contractions) cluster and the other representing the normal beats family. Fig. 8 shows some beats maps obtained with different records. As we can see, the applied transformation, based on patient specific reference signs, results very efficient in order to enhance pathological events and dangerous trends. Also small absolute morphologic variations appear, projected in this "user fit" space, as large relative perturbations respect to the physiological cluster. The physiological and pathological morphology-features families seems to be well separated, trends due to ischemic episodes are well visible and a "summary graphics", representing hours of recordings and thousands of heart-beats, is available for an easiest interpretation. After this kind of visual exploration the user can also proceed with a quantitative analysis performing the cluster analysis.

2.3. *Cluster analysis*

To characterize the different data groups (physiological and eventually pathological clusters) observed in the reduced space, a recently introduced algorithm, called KHM [11], was exploited to calculate clusters centers. According to many authors, the well-known *K*-Means method stands out, among the many clustering algorithms developed in the last years, as one of the most famous methods accepted by many appli-

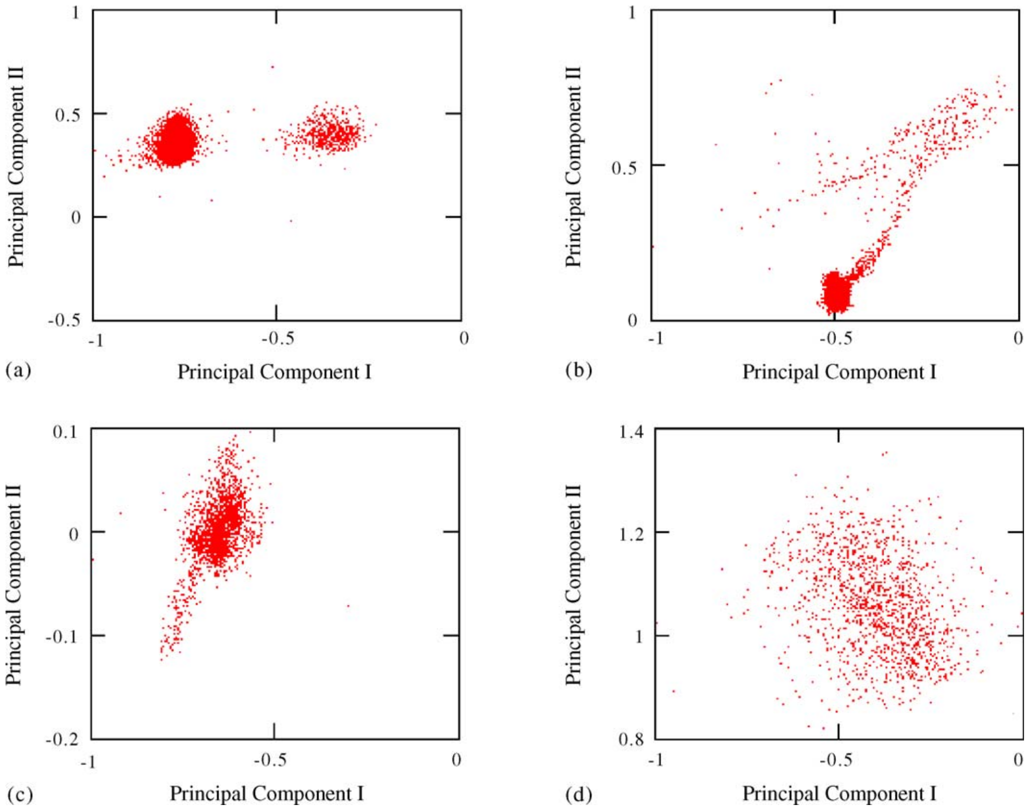


Fig. 8 – The graphs represent the ECG beats projections in the reduced space obtained analyzing recordings by different patients. Each point represents one beat. (a) This graph is obtained analyzing the record "Vale03" (ESC/VALE Database [12,27]). Two well-separated clusters are visible. Tracing the various beats with the *exploration window* of the program, we found that the left cloud represented the physiologic beat family while the cluster on the right was generated by aberrant QRS morphologies (ectopic beats, PVC). (b and c) In this cases we cannot see a nice separation between different clusters. The graphs are obtained analyzing, respectively the records "Vale 05" (ESC/VALE Database) and "E0106" (ESC/ST-T Database [12]) recorded by patients suffering with ischemic attacks. The figures show pathologic pathways that gradually depart from the main clusters. Those deviations are generated by the progressive changes in the ST-T segments due to ischemia. Tracing the various beats in the exploration window of the program, the user is able to visualize the various steps of the ischemic episodes and characterize their evolution. The projections in (d) are obtained analyzing a control ECG record not containing pathological events. In this case we can see one physiologic cluster. Data provided by Physionet [18].

cation domains, also for classification of biomedical data. A major problem with this algorithm is that it is sensitive to the selection of the initial partition and may converge to a local minimum of the objective function if the initial centers are not properly chosen [22]. The KHM algorithm, recently introduced by Zhang et al. [11,23], solves this problem by replacing the minimum distance from a data point to the centers, used in *K*-Means, by the harmonic averages of the distances from the data point to all centers. The harmonic average of *N* numbers is defined as:

$$H(a_1, \ldots, a_N) = \frac{N}{\sum_{k=1}^{N} 1/a_k} \tag{8}$$

This quantity is small also if at least one of the numbers is small. KHM shows better performances respect to *K*-Means, in fact the method uses the association provided by the harmonic means function, to replace the winner-takes-all strategy of *K*-Means.

The association of data points with the centers is distributed and the transition becomes continuous during convergence. According to the authors [11,23] KHM has a "built-in" dynamic weighting function, which boosts the data that are not close to any center by giving them a higher weight in the next iteration. The recursive expression used to calculate centers coordinates at each iteration is:

$$c_k = \frac{\sum_{i=1}^{N} \frac{1}{d_{i,k}^3 \left(\sum_{l=1}^{K} 1/d_{i,l}^1\right)^2} x_i}{\sum_{i=1}^{N} \frac{1}{d_{i,k}^3 \left(\sum_{l=1}^{K} 1/d_{i,l}^2\right)^2}} \tag{9}$$

where c_k, $k = 0, \ldots, K$ are the center positions, x_i are the coordinates of the *i*-th data point, the terms $d_{i,l} = ||x_i - c_l||$ represent the distances of data points from centers. The KHM algorithm starts with a set of initial positions of the centers, the terms $d_{i,l}$ are calculated and then the new positions of the centers are obtained using (9).

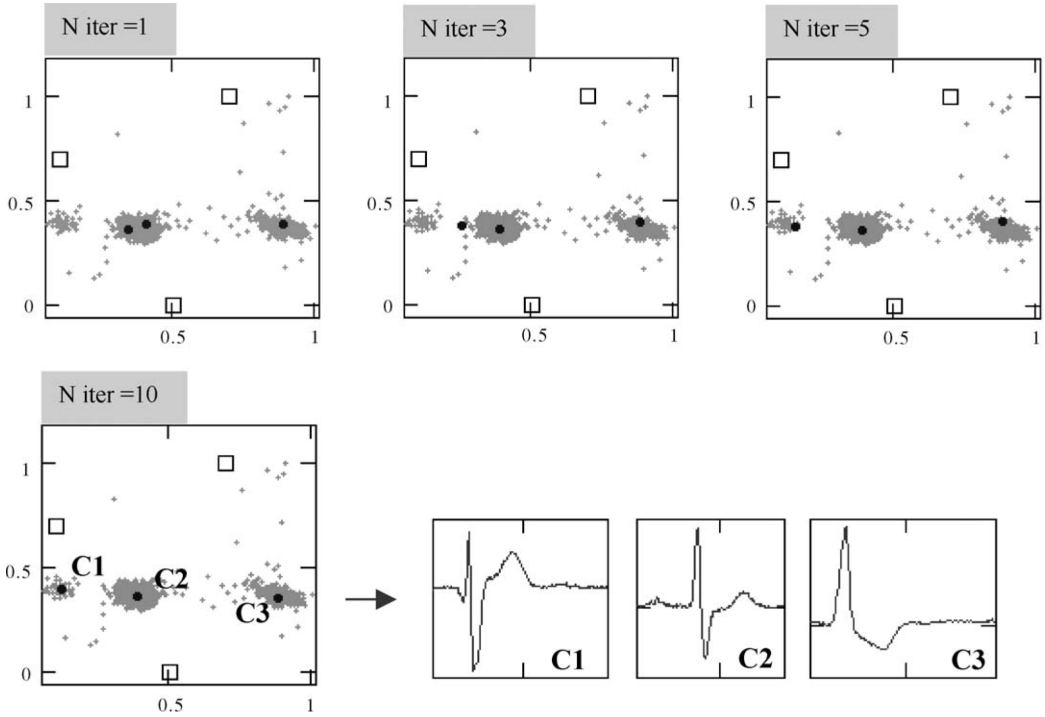


Fig. 9 – KHM algorithm convergence and identification of centers (C1, C2, and C3). The ambulatory ECG signal was recorded, during a holter session, by a patient suffering with rhythm disorder. Boxes represent the initial center positions; black circles indicate the center positions during the iterative process. Points are plotted in the I–II principal components plane. The algorithm calculates the center positions and the various beats memberships. The latter are useful in order to provide an automatic estimate of the aberrant beats number. Data are provided by Physionet [18].

The algorithm has been extensively tested [24–26] and compared with the well-known *K*-Means (KM) and expectation maximization (EM) methods. With regard to the computational cost, Zhang showed in [24] that the asymptotic computational complexity per iteration for KM, KHM and EM are all $O(NKD)$, where N represents the number of data points, K indicates the number of clusters, and D indicates the dimension of the data vectors. For all three algorithms, since the costs are dominated (especially for high dimensional data) by the distance calculations, and there is exactly the same number of distances to be calculated, the coefficients of the cost term NKD of all three algorithms are very close.

It is the convergence rate and the convergence quality that differentiate them in real world applications. Space complexity of KHM is ND for data points, KD for the K centers and $KD+2K$ for temporary storage. The temporary storage requirement tends to be lower than KM because the later needs a $O(N)$ temporary storage to keep the membership information and $N \gg K$ in real problems. It has been demonstrated [24], using “benchmark” datasets, that KHM outperforms KM and EM in terms of convergence rate and quality. In [25], the author compares KM, EM and KHM performances statistically, running the algorithms on 3.600 pairs of dataset/initialization to compare the statistical average and variation of the performance of these algorithms. The results are that, KHM performs consistently better than KM and EM. Also in [26] the authors compared the three clustering algorithms, showing the superiority of the *k*-harmonic means algorithm (KHM) for finding clusterings of high quality in experimental dataset.

For those reasons we preferred KHM to classic KM/EM methods in our applications.

The user, after the visual inspection of the reduced space, selects the number of clusters to be characterized and runs the computational routines. Fig. 9 shows an example of centers identification performed by KHM algorithm in a data matrix containing ectopic beats (figure shows data points in the I–II PC plane).

As figure shows, three families, representing different morphological patterns are visible in the reduced space. The various clusters are characterized by the centers C1, C2, and C3. The figure also shows the iterative process performed to reach convergence. The vision of the reduced space together with the calculation of the cluster centers provides to the user a valid instrument for a rapid and effective characterization of the entire record, allowing the practitioner to discover dangerous patterns without inspecting “manually” thousands of heart-beat. The center coordinates and the corresponding ECG waveforms are saved together with data points coordinates, centers coordinates, clusters density and *RR*–ST time series, in a session file to disk for further investigations. Also references (I_{on}, I_{off}, B_x, R_x) are saved to disk.

We used the VALE Database [27] to perform a preliminary evaluation of the methods used for data exploration (PCA module and KHM clustering algorithm).

The VALE (VALidation-Ecg) Database is intended to be used for evaluation of algorithms for arrhythmia detection as well as ST-T analysis. This database consists of 35 annotated records, selected from ambulatory ECG recordings of 35 subjects. The database includes about 100 episodes of ST segment change, about 200 episodes of T-wave change, as well as a variety of ventricular and supraventricular arrhythmias. Each record is 3 h in duration and contains one signal, sampled at 200 samples per second with 12-bit resolution over a 25–40 mV input range. The VALE database was developed by a joint effort of the Departments of Cardiology of the CNR Institute of Clinical Physiology in Pisa and of the University of Pisa Institute of Medical Pathology. Both groups contributed recordings and participated in the beat-by-beat annotation of the VALE Database [27].

3. Results

While a clinical experimentation of our program is on the way, we validated the proposed methods with standard ECG databases, obtaining encouraging results.

In particular, the segmentation algorithm performances are strictly related with the main $U3$ algorithm performances, so a wide validation has been carried out in order to test the quadratic operator. In order to validate $U3$, we analyzed the MIT-BIH arrhythmia database [13] comparing the performance of our method with the results obtained by Pan and Tompkins with the same database [17].

In particular, we applied the $U3$ operator to the first channel of each record in the MIT-BIH database. The obtained signals were analyzed with the same dual-threshold detection technique used by Pan and Tompkins [17] and the calculated QRS indexes were compared with the QRS-annotation indexes of cardiologists in the database. Table 1 summarizes the results obtained with this database. Our algorithm produced 565 false positive beats and 379 false negative beats and a total detection failure of 0.85% considering all the 109.809 annotated beats in the database. A similar table showing the results obtained by Pan and Tompkins were presented in [17]. Their algorithm produced 507 false positive beats and 277 false negative beats for a total failure of 0.71%.

Both algorithms show similar performances (failure rates are <1%) in order to detect QRS complexes. $U3$ has the advantage to be faster, in fact, as we can see from Eq. (6), each sample of the output is obtained with six operations, while a larger number of operations are needed to calculate the output with the cascade of four filters proposed by Pan and Tompkins (see Table 2 for details). Since in our application we needed a fast and reliable algorithm to analyze long-term recordings, we preferred to include $U3$ for QRS detection in the final version of our system prototype we used the VALE Database [27] to perform a preliminary evaluation of the methods used for data exploration (PCA module and KHM clustering algorithm). First, we explored each record in the database using the program interface (Fig. 7) presented in Section 2.2. Next, we characterized the observed clusters using the KHM method. Finally, we compared the waveforms corresponding to the cluster centers with the annotations of the cardiologists. Considering the entire database, we succeeded in identifying pathological clusters in 97% of the cases. Some episodes, annotated as *supraventricular ectopic beats*, were not detected because they didn't produce significant changes to the beat morphology, producing false negatives with the PCA methods. However those episodes could be easily detected simply analyzing the RR series calculated using $U3$.

4. Conclusions and perspectives

This study is part of a wide research project, called AMICUS [28], launched in 2003 at the BIM Lab, Department of Systems and computer science, University of Florence. The objective of this project is to design a personalized communication system that can improve the quality of daily life and medical care for the chronically ill, at the same time helping to reduce the number and duration of hospital recoveries through long-term monitoring of patients vital signs at home. AMICUS aims at contributing to the reduction of health costs and at providing easy access to medical assistance. This is in line with a recent position statement from WHO that strongly encourages the development of devices able to give patients an active, conscious role in the management of their disease: actions are needed that emphasize collaborative goal setting, patient skill building to overcome barriers, self-monitoring, personalized feedback, and systematic links to community resources. Briefly, the system has a split architecture and consists of two units: one is a wearable Blue Tooth biosignal transmitter whose range covers the patient home area and the other is a PC work-station, which receives data from the patients, submitting the information to the health care network. The system produces a large amount of data to be analyzed and inspected by the practitioners involved in this project. According to them, our methods represent an important aid for an easy and fast exploration of the ECG records transmitted by PC stations at patients' home to the health care network. We are going to begin a clinical experimentation of the system with a very well defined group of patients. In this respect we are currently talking to the Department of Critical Care, Section of Respiratory Medicine, University of Florence with reference to the population affected by COPD (Chronic Obstructive Pulmonary Disease). On-line long-term monitoring of vital signs could play an important role in the well being of these chronic patients. It could provide critical information for long-term assessment and preventive diagnosis for which trends over time and signal patterns are of special importance. Such trends and patterns are often ignored or scarcely identified by traditional examinations.

For those reasons, in this study we developed a procedure to analyze and discover patters in ECG ambulatory recordings to be considered as a medical decision-making support. The program is based on a beat detection algorithm especially designed for low SNR applications while events exploration and characterization is performed in a reduced dimension space through a recently introduced clustering method. The algorithms have been validated with standard ECG databases. Results are presented to the user with appropriate graphics to facilitate a visual and immediate interpretation of long recordings. Results are also saved to disk for further investigations.

We think that those methods may be useful also to analyse standard holter tapes and X-ECG records, especially to characterise the beat morphology modifications that can occur during stress test exercises due to ischemic episodes or to rhythm disorders. Now, we are working on the second version of the C++ program, in particular a new 3D interactive visualization panel will be added exploiting the Open-GL library.

REFERENCES

[1] Observatory on Health Care for Chronic Conditions, Home-Based and Long-Term Care, report who/hsc/lth/99.2, World Health Organization, 1999.

[2] J. Cleland, A. Amala, A. Rigby, U. Janssens, H.M.M. Aggie, et al., Noninvasive home telemonitoring for patients with heart failure at high risk of recurrent admission and death: the trans-European network-home-care management system (TEN-HMS) study, J. Am. Coll. Cardiol. 45 (10) (2005) 1654–1664.

[3] S. Guillen, M.T. Arredondo, V. Traver, J.M. Garcia, C. Fernandez, Multimedia tele-homecare system using standard TV set, IEEE Trans. Biomed. Eng. 49 (12) (2002) 1431–1437.

[4] V. Rialle, J.B. Lamya, N. Nourya, L. Bajollea, Telemonitoring of patients at home: a software agent approach., Comput. Methods Prog. Biomed. 72 (3) (2003) 257–268.

[5] T. Tamura, T. Togawa, M. Ogawa, M. Yoda, Fully automated health monitoring system in the home, Med. Eng. Phys. 20 (8) (1998) 573–579.

[6] M. Ogawa, T. Togawa, The concept of the home health monitoring, in: Proceedings of the 5th International Workshop on Enterprise Networking and Computing in Healthcare Industry, 2003, pp. 71–73.

[7] F. Magrabi, N. Lovell, B. Celler, A web-based approach for electrocardiogram monitoring in the home, Int. J. Med. Infor. 54 (2) (1999) 145–153.

[8] H. Asada, P. Shaltis, A. Reisner, S. Rhee, R. Hutchinson, Mobile monitoring with wearable photoplethysmographic biosensors, IEEE Eng. Med. Biol. Mag. 22 (3) (2003) 28–40.

[9] A. Lymberis, D. De Rossi, Wearable eHealth systems for personalised health management: state of the art and future challenges Studies in Health Technology and Informatics, vol. 108, IOS Press, 2004, pp. 162–171.

[10] R. Bellazzi, S. Montani, A. Riva, M. Stefanelli, Web-based telemedicine systems for home-care: technical issues and experiences, Comput. Methods Prog. Biomed. 64 (3) (2001) 175–187.

[11] B. Zhang, M. Hsu, U. Dayal, *K*-harmonic means – a data clustering algorithm, Technical Report HPL-1999-124, Hewlett-Packard Labs, 1999.

[12] A. Taddei, G. Distante, M. Emdin, P. Pisani, G.B. Moody, C. Zeelenberg, C. Marchesi, The European ST-T database: standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography, Eur. Heart J. 13 (1992) 1164–1172.

[13] G.B. Moody, R.G. Mark, The impact of MIT-BIH database, Eng. Med. Biol. Mag. IEEE 20 (3) (2001) 45–50.

[14] E.N. Bruce, Biomedical Signal Processing and Signal Modeling, Proakis Editor, 2001.

[15] L.J. Hadjileontiadis, K.I. Panoulas, T. Penzel, S.M. Panas, Performance of three QRS detection algorithms during sleep: a comparative study, in: Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 2, 2001, pp. 1954–1957.

[16] M. Paoletti, C. Marchesi, Model based signal characterisation for long-term personal monitoring, IEEE Comput. Cardiol. 28 (2001) 413–416.

[17] J. Pan, W.J. Tompkins, A real time QRS detection algorithm, IEEE Trans. Biomed. Eng. 32 (1985) 230–236.

[18] A. Goldberger, L. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R. Mark, J. Mietus, G. Moody, C. Peng, H. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) 215–220.

[19] F. Jager, Automated detection of transient ST-segment changes during ambulatory ECG-monitoring, Ph.D. Thesis, University of Ljubljana, 1994.

[20] F. Jager, R.G. Mark, G.B. Moody, S. Divjak, Analysis of transient ST segment changes during ambulatory ECG monitoring using the Karhunen–Loève transform, IEEE Comput. Cardiol. 19 (1992) 691–694.

[21] I.T. Jolliffe, Principal Components Analysis, Sprinter, 1986.

[22] M. Kantardzic, Data Mining: Concepts, Models, Methods and Algorithms, John Wiley and Sons, 2003.

[23] B. Zhang, M. Hsu, U. Dayal, *K*-harmonic means – a spatial clustering algorithm with boosting, in: Proceedings of the International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining, TSDM2000, Lyon, 2000, pp. 31–45.

[24] B. Zhang, Generalized *k*-harmonic means – boosting in unsupervised learning, Technical Report HPL-2000-137, Hewlett-Packard Labs, 2000.

[25] B. Zhang, Comparison of the performance of center-based clustering algorithms, in: Advances in Knowledge Discovery and Data Mining: 7th Pacific-Asia Conference, vol. 2637, PAKDD, 2003, pp. 63–74.

[26] G. Hamerly, C. Elkan, Alternatives to the *k*-Means algorithm that find better clusterings, in: Proceedings of the 11th International Conference on Information and Knowledge Management, ACM Press, 2002, pp. 600–607.

[27] A. Taddei, M. Varanini, A. Macerata, C. Marchesi, C. Contini, A. Biagini, M.G. Bongiorni, M. Mazzei, G.F. Mazzocca, M. Baratto, An annotated database for the evaluation of algorithms for the analysis of arrhythmias and ischemic events, IEEE Comput. Cardiol. 10 (1983) 191–194.

[28] M. Paoletti, L. Galeotti, C. Marchesi, Building a bridge for communication between patients, family doctors and specialist, Eur. Res. Consort. Infor. Math. 61 (2005) 45–46.