Outlier Detection in Non-intrusive ECG Biometric System

André Lourenço[1,2,⋆], Hugo Silva[2,⋆⋆], Carlos Carreiras[2], and And Fred[2]

[1] Instituto Superior de Engenharia de Lisboa
[2] Instituto de Telecomunicações and Instituto Superior Técnico
{arlourenco,hugo.silva,carlos.carreiras,afred}@lx.it.pt

Abstract. A recent trend in the field of biometrics is the use of Electrocardiographic (ECG) signals. One of the main challenges of this new paradigm is the development of non-intrusive and highly usable setups. Fingers and hand palms for example, allow the ECG acquisition at much more convenient locations than the chest, commonly used for clinical scenarios. These new locations lead to an ECG signal with lower signal to noise ratio, and more prone to noise artifacts. In this paper, we propose an outlier removal system to eliminate noisy segments, and enhance the performance of non-intrusive ECG biometric systems. Preliminary results show that this system leads to an improvement on the recognition rates, helping to further validate the potential of ECG signals as complementary biometric modality.

Keywords: Biometrics, ECG, Outlier Detection, Clustering.

1 Introduction

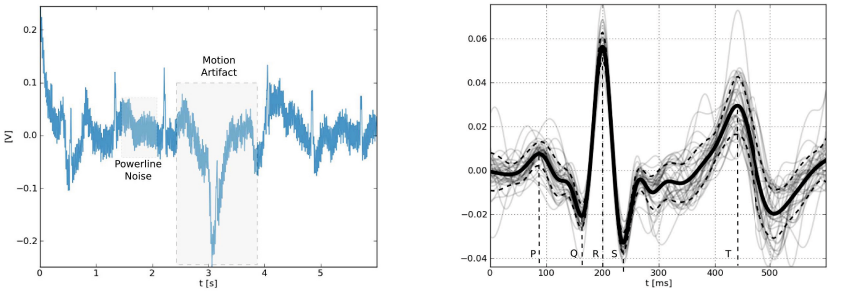
In the begining of the century it was shown that Electrocardiographic (ECG) signals hold relevant subject-dependent information [1,2]. Due to their characteristics, ECG signals are emerging as complementary biometric trait, as they are: a) originated by a vital and living organ; b) permanently available; c) more difficult to mimic since there is no association to reproducible external physical landmarks.

In the literature, ECG-based biometric methods can be classified as either fiducial or non-fiducial. The fiducial methods, use anchor points (called fiducia) as references [3,4]. Typically the ECG R-peak is used as reference, since it is the easiest to distinguish, but any other complex can be used [1,5,6,7]. Non-fiducial methods extract information from the ECG signals without having any reference point [8,9]. New approaches (partially fiducial) are starting to appear, where fiducial information is only used for ECG segmentation [10,11,3].

⋆ FCT grant SFRH/PROTEC/49512/2009, and Área Departamental de Engenharia Electrónica e Telecomunicações e de Computadores-ISEL.
⋆⋆ FCT grants SFRH/BD/65248/2009 and PTDC/EEI-SII/2312/2012.

M. Kamel and A. Campilho (Eds.): ICIAR 2013, LNCS 7950, pp. 43–52, 2013.

On clinical applications, these signals are acquired using 12 or more leads mounted on the chest and limbs, using gel to improve conductivity. The acceptance of ECG as biometric requires more practical sensing apparatus, namely, through the design of non-intrusive, high usability acquisition methods, making them more suitable for regular use and widespread applications. Recent advances in biomedical instrumentation have enabled the acquisition of ECG signals on a plethora of setups, some examples include: on the chest using a T-shirt with textile embedded electronics [12]; at the neck using a necklace with a pendant [13]; at the fingers and hand palms with a one lead sensor using dry Ag/AgCl electrodes [11] or using textile electrodes [14]. The latter case has been entitled off-the-person approach, since it does not require the signal acquisition devices to be placed on the body of the person as previous approaches.



(a) Noise Artifacts (b) Segmented Individual Heartbeats

Fig. 1. Example of ECG signals acquired at the fingers: a) motion artifacts and powerline noise; b) Segmented heartbeat waveforms with annotated complexes (P-QRS-T). The dark wave represents the mean, and dashed line the standard deviation; in light gray we provide an overlay with the segmented heartbeats.

These new locations lead to an ECG signal with different characteristics: lower signal to noise ratio, and more prone to noise artifacts. Figure 1 shows an example of a signal aquired at the fingers using dry Ag/AgCl electrodes, and the difficulties arising in this context. Figure 1(a) shows examples of motion artifacts and powerline interference, which lead to additive noise level almost close to the maximum amplitude of the R-peak. Figure 1(b) illustrates the typical ECG heartbeat waveform (with annotated complexes) and the segmented individual heartbeats (after digital filtering), illustrating the intra-subject variability induced by this new acquisition setup.

In this paper we focus on the development of an outlier removal system, to rule out noisy segments from the decision process, and enhance the performance of non- or minimaly-intrusive ECG biometric systems, which by their characteristics are much more prone to artifact interference. We provide a review on the state-of-the-art in outlier detection algorithms, and study their application to ECG signals.

The remainder of the paper is organized as follows. In Sections 2 and 3, we present an overview of ECG biometric systems and outlier removal algorithms, respectively. In Section 4 we propose solutions for outlier detection in the context of ECG signal processing. Finally, in Sections 5 and 6 we present the experimental setup, a summary of the results and outline the main conclusions.

2 ECG Biometric Systems

The typical block diagram of a fiducial, or partially fiducial, biometric system is depicted in Figure 2. These systems rely on the detection of notable ECG complexes for segmentation and extraction of a sequence of individual heartbeats. Typically, the QRS complex is used for such purpose.

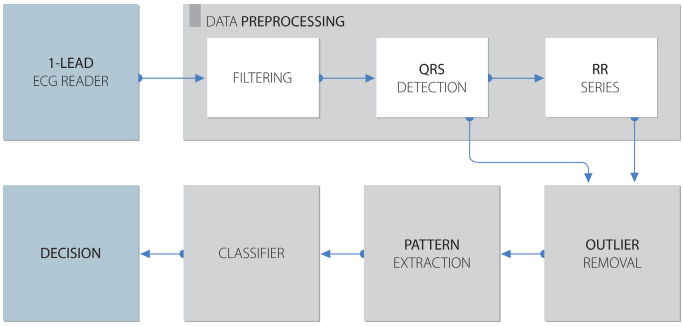


Fig. 2. Block diagram of a typical ECG biometric system

Our ECG biometric system, starts with the acquisition of raw data, using a custom one lead sensor with virtual ground and dry Ag/AgCl electrodes [14]. The acquired signal is then converted from analog to digital, and submitted to a data preprocessing block, which performs a digital filtering step (band pass FIR filter order 300, and cutoff frequencies [5-20]Hz) and the QRS complex detection [15]. The outputs of this block are segmented individual heartbeats, and a RR interval time series. Figure 1(b) illustrates a typical ECG waveform, representing the mean of the segmented individual heartbeat waveforms depicted in dark gray.

Not so common is the outlier detection block, which is the focus of our work, and that uses all the available information for the detection and removal of anomalous ECG heartbeats. Alternatively, in some biometric systems, instead of discarding outliers, the paradigm can be focused on selecting the most representative instances from the set of available templates [16,17].

The pattern extraction block takes the preprocessed input signals, and starts by aligning all the heartbeat waveforms by their R-peak instants, and by clipping them in the interval [-200;400] miliseconds around that instant. For the scope of this work we consider the features to be all the amplitudes within this interval.

The templates are formed based on the mean (or median) of 5 consecutive heartbeat waveforms. In the end, a k-NN classifier (with k=3) is used together with a distance metric, to produce a decision on the recognition of the individual (either in authentication or identification), as it was found to be a good compromise between performance and computational cost [10]. In this work we compare the Euclidean ($D_{Euclidean}$) with cosine (D_{cos}) distances.

3 Outlier Detection and Removal

Traditional outlier detection systems rely on models of "normal" data and on the detection of deviations from this model in the observed data. This approach belongs to the supervised learning paradigm, since models for "normal" data are being learned from the data itself. One example are the one-class classification systems, which use only one class (the target class), usually well characterized by instances in the training data and not requiring instances from any other class as in conventional multi-class classification algorithms [18].

In a different paradigm, unsupervised outlier detection algorithms, determine the abnormal instances with no prior knowledge from the data [19,20,21], not requiring any pre-labeled instances. Usually these systems are based on clustering techniques; the objective of clustering is to map a partition of unlabeled instances, into a set of K classes or groups [22].

Given a predefined outlier criteria, the outlier detection can be a by-product of a clustering algorithm [23]. As an example, let us consider the CLARANS (Clustering Large Applications based on RANdomized search) algorithm [24], where outliers are instances belonging to clusters with silhouette value widths below 0.5, and discarded if the total number of objects removed until a given point was not higher than predetermined threshold (25% of the total number of objects was suggested). In a similar way, in DBSCAN (Density Based Spatial Clustering of Applications with Noise) [25], outliers are considered instances not belonging to any cluster, but following a density-based notion of cluster. These techniques implicitly define outliers as the background noise in which the clusters are embedded [23].

Other classes of algorithms define outliers as instances, which are neither part of a cluster, nor part of the background noise, just instances which behave very differently from the norm [23]. Eskin *et al.*[26] presented three algorithms for anomaly detection, where anomalies are detected by determining which instances lie in sparse regions of the feature space. The first algorithm, cluster-based estimation, relies on counting the number of instances that are within a sphere of radius w around a point. Instances that are in dense regions are considered "normal", while instances that are in a sparse region are considered anomalies. The second algorithm, k-Nearest Neighbor (k-NN), computes the sum of the distances to the k-NN of a point, to determine whether or not it is in a sparse region. The last algorithm, One Class SVM, uses an unsupervised variant of SVM, which tries to separate the entire dataset from the origin.

4 Proposed Approach

Let $X = \{x_1, \ldots, x_n\}$ be a set of n individual heartbeats, x_i, which can be described by a set of features $x_i = (x_i(1), \ldots, x_i(m)) \in \mathbb{R}^m$. Our approach for outlier detection can be characterized by applying a function $f: \ x_i \rightarrow \{0, 1\}$, from the feature space representation of the heartbeats, x_i, to the set of "normal" (0) and "outlier"(1) heartbeats. Different methodologies lead to several definitions of f. We define as hypothesis, that the majority of the heartbeats is "normal" and that outliers lie in sparse regions of the feature space. We present two different approaches for outlier detection on ECG heart beat sequences based on clustering, namely: a) Distance-based detection and b) Clustering with integrated outlier detection criterion.

4.1 Distance-Based Detection

Eskin *et al.*[26], detected outliers based on the number of instances within a sphere of radius w around each instance. This approach requires the computation of the pairwise distances between all the instances in X and the definition of the threshold w. We followed a different approach computing the distances, $D(,)$ to a single reference, the mean template of each recording session, being given by $\mu_x = \frac{1}{N}\sum_1^n x_i$. Additionaly, instead of fixing a threshold, we defined an adaptative one based on the mean and standard deviation of $D(,)$. Furthermore, we checked empirical consistency rules of the ECG, namely the amplitude of the maximum and minimum values and the position of the R peak. For that let's define x^{min} and x^{max} as the minimum and the maximum value of each template, and i_{min} and i_{max} as its corresponding indexes.

We call this method "DMEAN", and it consists of the following steps:

1. For each $x_i \in X$ compute its distance to the mean μ_x: $D(x_i, \mu_x)$;
2. Compute the 1^{st} and 2^{nd} order statistical moments of the distances $D(\cdot, \mu_x)$; $\mu_{D(\cdot,\mu_x)}$ corresponds to the mean value, and $\sigma_{D(\cdot,\mu_x)}$ to the standard deviation;
3. Compute the median of the minimum and maximum values of each template (x^{min} and x^{max}), denoted as $\tilde{x}_{min}$ and $\tilde{x}_{max}$, respectively;
4. For each for a given $x_i \in X$ verify each of the conditions below (if any is met x_i will be considered an outlier):
   (a) i_{max} does not correspond to R-peak position (determined from the segmentation step);
   (b) $x^{max} > 1.5 * \tilde{x}_{max}$ and $x^{min} < 1.5 * \tilde{x}_{min}$;
   (c) $D(x_i, \mu_x) > (\mu_{D(x_i,\mu_x)} + 0.5 * \sigma_{D(x_i,\mu_x)})$.

This algorithm can be used with any distance metric; in this work we choose the cosine distance:

$$D_{cos}(x_u, \mu_x) = 1 - \frac{\sum_{k=1}^m x_u[k]\mu_x[k]}{\sqrt{\sum_{k=1}^m x_u[k]^2 \sum_{k=1}^m \mu_x[k]^2}}, \quad (1)$$

4.2 Clustering with Integrated Outlier Detection Criterion

Several clustering algorithms integrate criteria for outlier detection, an example of which being the DBSCAN [25]. In this algorithm, a cluster is defined based on the concept of density reachability. A given point x_i is directly density-reachable from a point x_j (and can be considered as part of the same cluster), if it is not farther away than a given distance ε, and is surrounded by a sufficient number of points, $MinPts$. This algorithm considers as noise (or outliers) those points not belonging to any cluster. In our tests we considered $\varepsilon = 0.95$ and $MinPts = 10$.

5 Results and Discussion

For the evaluation of the proposed approach, we used a dataset consisting of the data from 63 subjects (49 males and 14 females) with an average age of 20.68 ± 2.83 years, in an experimental setting where subjects were only asked to sit for 2 minutes in a resting position with two fingers, one from the left and another from the right hand, placed in each of the dry electrodes depicted in Figure 3. The signals were digitalized using a bioPLUX research acquisition unit (12-bit resolution and 1kHz sampling frequency). The data consists of two independent acquisition sessions separated by a 3-month interval, entitled "T1" and "T2". None of the participants reported any health problems, reason for which we consider the collected data to be representative of the normal population. All of the participants signed an informed consent form in order to participate in the experiment.

The evaluation of our outlier detection system is based on: a) a descriptive statistic and inspection of the detected outliers; b) a quantitative analysis on the recognition performance.

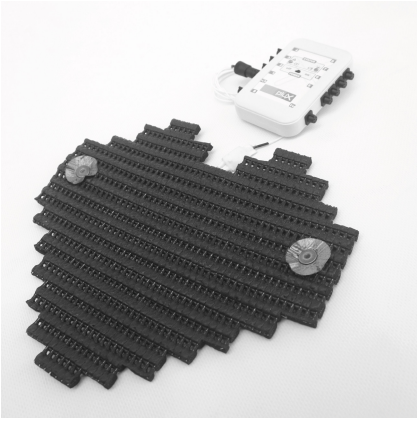
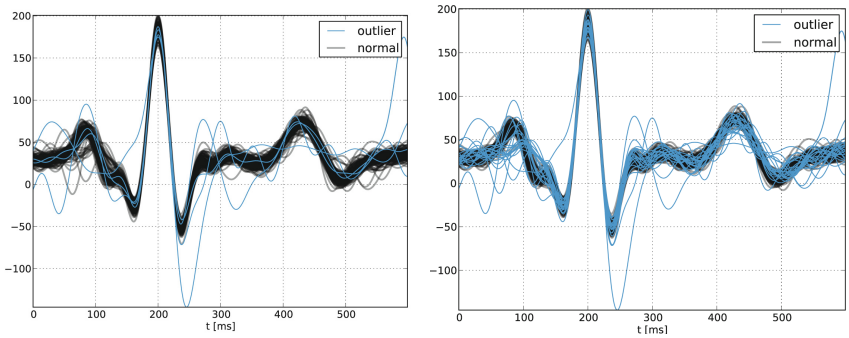


Fig. 3. Experimental apparatus consisting of the sensor pad with a Heart shaped form-factor where a pair of Ag/AgCl electrodes are integrated, and a biosignal acquisition device

5.1 Statistics and Visual Inspection

Figure 4 presents an illustrative examples of the proposed algorithms, showing examples of "normal" (black lines) and "outlier" (light blue) templates, for the same subject. The heartbeat waveforms that are clearly distinct from the majority are detected and removed by both approaches. The DBSCAN is more selective, considering as outliers only a small proportion of the heartbeats, and in several situations misses some heartbeats that could be considered by experts as outliers. On the other hand, DMEAN tends to detect much more heartbeats than DBSCAN, considering as "normal" only those that are close to the mean template.



(a) DBSCAN (b) DMEAN

Fig. 4. Examples of outlier detection using the proposed approaches

Table 1 summarizes the descriptive statistics of both methods over the T1 and T2 acquisition sessions; our results reinforce that sistematically DBSCAN tends to detect less outliers than DMEAN. DBSCAN removes approximatelly 6% of the heartbeats per session, while DMEAN removes approximatelly 25%.

Table 1. Descriptive Statistics ($\mu \pm \sigma$) of the proposed Outlier Detection algorithms over both acquisition sessions (T1 and T2). The *total* column describes the total number of analysed segments, and the *removed* column the number of segments considered as "outliers".

Session	Algorithm	Total	Removed
T1	DBSCAN	168.7 ± 23.5	11.3 ± 9.1
	DMEAN		42.0 ± 19.2
T2	DBSCAN	172.2 ± 30.0	11.3 ± 14.3
	DMEAN		45.0 ± 21.5

5.2 Biometric Recognition Rates

The biometric recognition rates with and without outlier detection were assessed using a 3-NN classifier, using as training information the templates from T1, and as testing information the templates from T2. We compared the performance of the system using as templates the means ($\overline{x}$) and medians ($\tilde{x}$) of 5 consecutive

heartbeats [5,10]. Furthermore we compared Euclidean ($D_{Euclidean}$) with cosine (D_{cos}) distance. Table 2 summarizes the results, showing that the use of outlier removal leads to a significant improvement on the Equal Error Rates (EER). The DMEAN approach outperforms DBSCAN in every situation. Regarding the different classifier configurations, two present the best result: a) D_{cos} using as templates $\overline{x}$; b) $D_{Euclidean}$ using as templates $\tilde{x}$. Regarding the mean and median templates, when we are not using an outlier removal step, the median ($\tilde{x}$) shows better results, but when this step is included, marginal differences are found..

Table 2. Equal Error Rates (EER) of the proposed Outlier Detection Approaches: None - without outlier removal; DBSCAN - using the DBSCAN algorithm, which has an integrated outlier detection criterion; DMEAN - using the distance-based detection.

Outlier Detection	Distance Metric			
	D_{cos}		$D_{Euclidean}$	
	$\overline{x}$	$\tilde{x}$	$\overline{x}$	$\tilde{x}$
None	16.8	15.9	15.7	14.4
DBSCAN	15.7	15.0	14.8	13.4
DMEAN	**12.4**	12.8	12.5	**12.4**

6 Conclusions

The use of finger and hand palms as signal acquisition locations, is setting a new paradigm on the ECG biometric systems landscape: the off-the-person approach. The ECG signal acquired at these new locations is more prone to noise artifacts and has a much lower signal to noise ratio. The development of an outlier detection system, that removes anomalous heartbeat waveforms from the collected signals is of critical importance. In this paper we propose two different approaches for the detection and removal of outlier on ECG heartbeats based on clustering. The first, DMEAN, is an adaptative distance-based approach which uses a single reference for this computation, the mean of the recording session. The later, is the DBSCAN, which has an integrated criterion for outlier detection based on the concept of density-reachable instance. We show that the use of these outlier detection steps leads to an improvement of the system performance. Future work will be focused on extending the clustering step for the selection of the representative template of each user.

References

1. Biel, L., Petterson, O., Phillipson, L., Wide, P.: ECG analysis: A new approach in human identification. IEEE Trans. Inst. and Measurement 50(3), 808–812 (2001)
2. Kyoso, M., Uchiyama, A.: Development of an ECG identification system. In: Proc. 23rd Int. Conf. IEEE Eng. Medicine and Biology Society, vol. 4 (2001)

3. Wang, Y., Agrafioti, F., Hatzinakos, D., Plataniotis, K.N.: Analysis of human electrocardiogram for biometric recognition. EURASIP J. Adv. S. Processing (2008)
4. Odinaka, I., Lai, P.H., Kaplan, A., O'Sullivan, J., Sirevaag, E., Rohrbaugh, J.: ECG biometric recognition: A comparative analysis. IEEE Trans. on Information Forensics and Security 7(6), 1812–1824 (2012)
5. Silva, H., Gamboa, H., Fred, A.: One lead ECG based personal identification with feature subspace ensembles. In: Perner, P. (ed.) MLDM 2007. LNCS (LNAI), vol. 4571, pp. 770–783. Springer, Heidelberg (2007)
6. Shen, T.: Biometric Identity Verification Based on Electrocardiogram. PhD thesis, University of Wisconsin (2005)
7. Israel, S., Irvine, J., Cheng, A., Wiederhold, M., Wiederhold, B.: ECG to identify individuals. Pattern Recognition 38(1), 133–142 (2005)
8. Coutinho, D.P., Fred, A.L.N., Figueiredo, M.A.T.: String matching approach for ECG-based biometrics. In: 16th Portuguese Conf. Pattern Recognition (2010)
9. Chan, A.D.C., Hamdy, M.M., Badre, A., Badee, V.: Wavelet distance measure for person identification using electrocardiograms. IEEE Trans. on Instrumentation and Measurement 57(2), 248–253 (2008)
10. Silva, H., Lourenço, A., Fred, A.: In-vehicle driver recognition based on hand ECG signals. In: Proc. of the IUI 2012 Conf. IUI (2012)
11. Lourenço, A., Silva, H., Fred, A.: Unveiling the biometric potential of Finger-Based ECG signals. Computational Intelligence and Neuroscience (2011)
12. Cunha, J., Cunha, B., Xavier, W., Ferreira, N., Pereira, A.: Vital-Jacket: A wearable wireless vital signs monitor for patients' mobility. In: Proceedings of the Avantex Symposium (2007)
13. Gamboa, H., Silva, F., Silva, H.: Patient tracking system. In: 4th Int Conf on Pervasive Computing Technologies for Healthcare, pp. 1–2 (March 2010)
14. Silva, H., Lourenço, A., Lourenço, R., Leite, P., Coutinho, D., Fred, A.: Study and evaluation of a single differential sensor design based on electro-textile electrodes for ECG biometrics applications. In: Proc. IEEE Sensors, Ireland (2011)
15. Lourenço, A., Silva, H., Leite, P., Lourenço, R., Fred, A.: Real time electrocardiogram segmentation for finger based ECG biometric. In: Proc. of BIOSIGNALS 2012, pp. 49–54 (February 2012)
16. Uludag, U., Ross, A., Jain, A.: Biometric template selection and update: a case study in fingerprints. Pattern Recognition 37(7), 1533–1542 (2004)
17. Lumini, A., Nanni, L.: A clustering method for automatic biometric template selection. Pattern Recognition 39(3), 495–497 (2006)
18. Tax, D., Duin, R.: Outliers and data descriptions. In: Proc. 7th Annual Conf. Advanced School for Computing and Imaging, ASCI (2001)
19. Zhang, Y., Meratnia, N., Havinga, P.J.M.: A taxonomy framework for unsupervised outlier detection techniques for multi-type data sets. Technical Report TR-CTIT-07-79, Centre Tel. and Inf. Tec. Univ. of Twente (November 2007)
20. Chandola, V., Banerjee, A., Kumar, V.: Outlier detection: A survey. ACM Computing Surveys (2007)
21. Hodge, V.J., Austin, J.: A survey of outlier detection methodologies. Artificial Intelligence Review 22 (2004)
22. Jain, A.K., Dubes, R.: Algorithms for Clustering Data. Prentice Hall (1988)
23. Aggarwal, C.C., Yu, P.S.: Outlier detection for high dimensional data. ACM SIGMOD, 37–46 (2001)

24. Ng, R.T., Han, J.: Efficient and effective clustering methods for spatial data mining. In: Proc. 20th Int. Conf. on Very Large Data Bases, VLDB 1994, pp. 144–155. Morgan Kaufmann Publishers Inc., San Francisco (1994)
25. Ester, M., Kriegel, H., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. of 2nd Int. Conf. on Knowledge Discovery and Data Mining, pp. 226–231. AAAI Press (1996)
26. Eskin, E., Arnold, A., Prerau, M., Portnoy, L., Stolfo, S.: A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In: Applications of Data Mining in Computer Security. Kluwer (2002)