ECG ANALYSIS USING CONSENSUS CLUSTERING

André Lourenço[*†], *Carlos Carreiras*[†], *Samuel Rota Bulò* [‡], *Ana Fred*[†]

[*] Instituto Superior de Engenharia de Lisboa
R. Cons. Emídio Navarro, 1
1959-007 Lisboa, Portugal

[†] Instituto de Telecomunicações
Av. Rovisco Pais, 1
1049-001 Lisboa, Portugal

[‡] FBK-irst
via Sommarive, 18, I-38123
Trento, Italy

ABSTRACT

Biosignals analysis has become widespread, upstaging their typical use in clinical settings. Electrocardiography (ECG) plays a central role in patient monitoring as a diagnosis tool in today's medicine and as an emerging biometric trait. In this paper we adopt a consensus clustering approach for the unsupervised analysis of an ECG-based biometric records. This type of analysis highlights natural groups within the population under investigation, which can be correlated with ground truth information in order to gain more insights about the data. Preliminary results are promising, for meaningful clusters are extracted from the population under analysis.

Index Terms— ECG analysis, ECG-based biometrics, consensus clustering, evidence accumulation

1. INTRODUCTION

Biosignals can be generally defined as observations of electrophysiological, biomechanical, or chemical processes of a living organism, ranging from protein and gene sequences, neural or cardiac rhythms, to tissue and organ images. Today, awareness and monitoring of biosignals have become widespread, upstaging their typical use in clinical settings. Novel trends towards biosignal-based well-being and quality-of-life products for end users are nowadays rapidly growing multi-million dollar market. Some examples of recent applications include: brain-computer interfaces [1], sports [2], physiotherapy [3], ergonomics [4], and biometrics [5–7].

The measurement and recording of the electrical activity of the heart, using electrocardiography (ECG), has come a long way since its introduction at the end of the 19th century. The ECG plays a central role in patient monitoring and as a diagnosis tool in today's medicine. It has also been used in sports, by monitoring athletes' performance, or in the affective computing domain for the extraction of features (*e.g.* heart rate variability), which are used for emotional state assessment.

A number of situations require the acquisition of long-term ECG recordings (several hours). Visual analysis of such records is tedious and prone to error. As a result, computer-based methods for the automatic analysis and interpretation of the ECG records have been developed, making this task easier, as well as minimizing inter- and intra-observer variation in human interpretation [8]. In this regard, clustering turns out to be a natural tool, for it allows to group ECG segments having a similar morphology, thus simplifying the analysis task.

In this paper we apply a state-of-the-art clustering methodology, namely consensus clustering, for the analysis of ECG signals in the context of ECG-based biometrics. Consensus clustering summarizes a set of input clusterings obtained for a particular dataset into a single consensus partition. Several authors have shown that these methods tend to reveal more robust and stable cluster structures than the individual clusterings in the ensemble [9, 10], and in the context of ECG clustering shown to perform well [11, 12]. This is particularly appealing as it allows to exploit different representations for the EGC-signals in the base clustering algorithms to better capture the cluster structures within the ECG data at the consensus clustering level. In the specific, we employ the consensus clustering algorithm introduced in [13, 14], which is based on the Evidence Accumulation Clustering (EAC) framework [9]. Accordingly, the consensus clustering problem is addressed by summarizing the information of the ensemble into a pairwise *co-association matrix*, where each entry holds the fraction of clusterings in which a given pair of objects is placed in the same cluster, thus subsuming the problem of associating the labels coming from different clusterings. The advantages of this approach are twofold: (i) it can deal with incomplete partitions in the ensemble, *i.e.* partitions comprising a subset of the data points, which is particularly helpful when the dataset is too large, or if the baseline clustering algorithm do not scale well; (ii) it can deal with *partial*, sparse observations of the co-association matrix entries (*a.k.a.* partial evidence accumulation), thus overcoming its intrinsic bad scalability issue.

Preliminary experiments highlights natural groups within the population under investigation, which can be correlated with ground truth information in order to gain more insights about the data.

This work was partially funded by Fundação para a Ciência e Tecnologia (FCT) under grant PTDC/EEI-SII/2312/2012

[*] gratefully acknowledges the grant SFRH/PROTEC/49512/2009

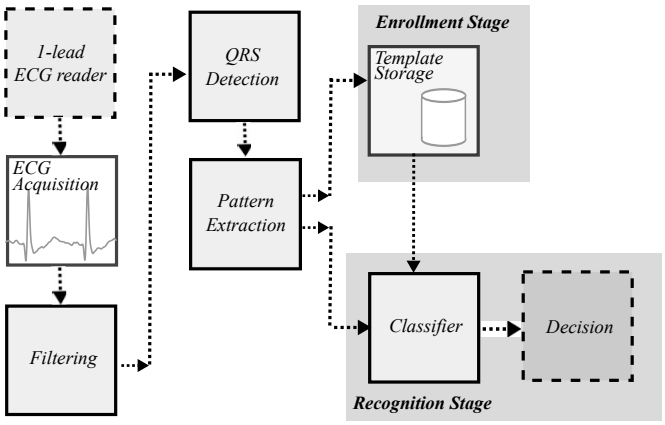


Fig. 1. General block diagram of an ECG-based biometric system.

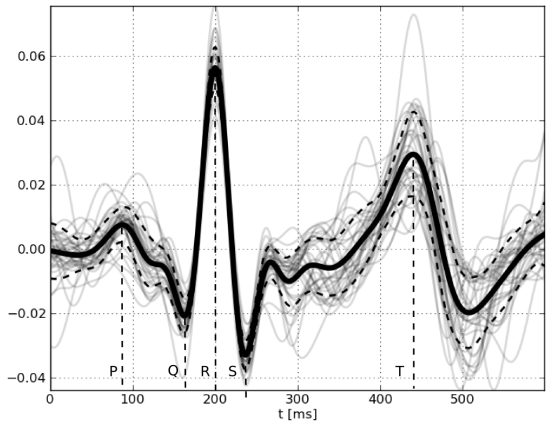


Fig. 2. Example of ECG acquired at the fingers: Segmented heartbeat waveforms with annotated complexes (P-QRS-T); the black line represents the mean, and dashed lines the standard deviation, while in dark gray we provide an overlay with all the segmented heartbeats.

2. ECG-BASED BIOMETRICS

ECG-based biometrics approaches can be classified as fiducial, non-fiducial or partially fiducial [5–7]. Fiducial approaches are based on reference points in the signals and/or specific features derived from them (such as the P-QRS-T complexes illustrated in Fig. 2) [15, 16]. Non-fiducial approaches refer to techniques which rely on intrinsic information from the ECG signals, without having any particular cues within the signal as a reference [6, 17, 18]. Partially fiducial approaches use fiducial information only for ECG segmentation [6, 19, 20]. We refer the reader to [5–7] for a comprehensive literature review.

The general block diagram of a typical ECG-based biometric system is depicted in Fig. 1. Raw data is acquired through an ECG recording device, which, in our case, uses a one-lead sensor with the acquisition being performed at the hands [5]. The raw data is submitted to a preprocessing block that filters the signal. The pattern extraction block takes the preprocessed input signals and, depending on the approach, extracts fiducial information from them, or features using some other method. In this work we follow a partially-fiducial framework by extracting features from segments of the individual heartbeat waveforms. Specifically, we abstract each individual heartbeat in terms of a feature vector, comprising amplitude samples of the heartbeat waveform in the interval $[t_R - 200; t_R + 400]\, ms$, t_R being the time instant of the R-peak reference complex, see Fig. 2.

During the enrollment stage, the system extracts heartbeat waveforms (or corresponding features), which will then be stored in a database, and used as representative templates during the recognition phase. In the recognition phase, a classifier is used to automatically assign an entity to the user (in identification scenarios), or to verify if the user is who he claims to be (in authentication scenarios). In a previous work we used a multi-class support vector machine (SVM) classifier at this stage [19]. In this paper, we depart from the supervised setting and consider an unsupervised learning scenario, aimed at highlighting natural groups within the population under analysis. The obtained groups could then be correlated with ground-truth information to better understand the sources of variability, or categorize subjects.

3. CONSENSUS CLUSTERING

In this section we present a brief review the consensus clustering approach based on evidence accumulation proposed in [13, 14]

The goal of consensus clustering is to partition a set of $\mathtt{n}$ data points $\mathcal{X} = \{\boldsymbol{x}_i\}_{i \in \mathcal{I}}$, indexed by $\mathcal{I} = \{1, \ldots, \mathtt{n}\}$, starting from a set of clusterings, *a.k.a.* ensemble of clusterings, obtained by running different algorithms (or different parametrizations/initializations)on possibly sub-sampled versions of the data set $\mathcal{X}$. This *ensemble of clusterings* is denoted as $\mathcal{E} = \{\phi_u\}_{u \in \mathcal{U}}$, where $\mathcal{U} = \{1, \ldots, \mathtt{m}\}$, each function $\phi_u : \mathcal{J}_u \to \{1, ..., \mathtt{k}_u\}$ encoding a partition of a subset of data points indexed by $\mathcal{J}_u \subseteq \mathcal{I}$ into $\mathtt{k}_u$ clusters. Partitions not comprising all data points might arise if one works with sub-sampled versions of the dataset, *e.g.* in the presence of a large amount of data points. The set $\mathcal{U}_{ij} \subseteq \mathcal{U}$ tracks the partition indices where both data points $i, j \in \mathcal{I}$ have been clustered, *i.e.* $(u \in \mathcal{U}_{ij}) \iff (i, j \in \mathcal{J}_u)$. Matrix $\mathtt{N}$ counts the number of times two distinct data points appeared in a partition of the ensemble and has a zero diagonal, *i.e.* $\mathtt{N}_{ij} = |\mathcal{U}_{ij}|$ if $i \neq j$ and 0 otherwise.

Given an ensemble $\mathcal{E}$, a *consensus clustering* is a partition minimizing its divergence from the other partitions in the

ensemble:

$$\phi^* \in \underset{\phi':\mathcal{I}\to\{1,\ldots,\mathsf{k}\}}{\arg\min} \sum_{u\in\mathcal{U}} d(\phi', \phi_u)\,, \tag{1}$$

where $d(\cdot,\cdot)$ is a function providing the divergence between the given partitions. To sidestep the problem of cluster correspondences and adhere to EAC principles, the following divergence is adopted

$$d(\phi', \phi_u) = \sum_{i,j\in\mathcal{J}_u} \left[\mathbb{1}_{\phi'(i)=\phi'(j)} - \mathbb{1}_{\phi_u(i)=\phi_u(j)}\right]^2, \tag{2}$$

which counts the number of times two data points are assigned the same cluster in ϕ_u, but different ones in ϕ', and vice versa. Here, $\mathbb{1}_P$ is the indicator function giving 1 if proposition P is true, 0 otherwise.

The so-called *co-association matrix* [9] is the matrix $\mathtt{C}$ holding the fraction of times two distinct data points have been assigned the same cluster, *i.e.*

$$\mathtt{C}_{ij} = \begin{cases} \frac{1}{N_{ij}} \sum_{u\in\mathcal{U}_{ij}} \mathbb{1}_{\phi_u(i)=\phi_u(j)} & \text{if } \mathtt{N}_{ij} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

By encoding the consensus clustering in terms of a binary matrix $\mathtt{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_\mathsf{n})$, each column $\boldsymbol{z}_i$ being an indicator vector of the cluster assignment of data point $i \in \mathcal{I}$, it is possible to link (1) with divergence (2) to the co-association matrix explicitly by means of the following matrix factorization:

$$\mathtt{Z}^* \in \underset{\mathtt{Z}\in\mathcal{S}_{01}^{\mathsf{k}\times\mathsf{n}}}{\arg\min} \|\mathtt{N} \circ (\mathtt{C} - \mathtt{Z}^\top\mathtt{Z})\|^2\,, \tag{3}$$

where $\circ$ is the Hadamard (*a.k.a.* element-wise) matrix product, $\|\cdot\|$ is the Frobenious matrix norm, and $\mathcal{S}_{01}^{\mathsf{k}\times\mathsf{n}}$ denotes the set of binary, left-stochastic matrices. The solution $\mathtt{Z}^*$ is equivalent to ϕ^*, *i.e.* $(\mathtt{Z}^*_{ki} = 1) \iff (\phi^*(i) = k)$. We omit the detailed proof of this relation due to lack of space.

In order to render the formulation (3) more scalable, it is possible to take into account only a subset $\mathcal{P} \subseteq \mathcal{I}\times\mathcal{I}$ of data point pairs. This is equivalent to forcing $\mathtt{N}_{ij} = 0$ for all $(i,j) \in \mathcal{P}$. By doing so, one relies only on a *partial* view of the co-occurrence statistics stored in the co-association matrix $\mathtt{C}$ (*a.k.a.* partial evidence) but, on the other hand, there is no need to compute and store all entries of $\mathtt{C}$, thus rendering the approach scalable. In general, the set $\mathcal{P}$ is sparse, *i.e.* $|\mathcal{P}| \ll \mathsf{n}^2$, and can be sampled randomly. Moreover we assume it to be symmetric, *i.e.* $(i,j) \in \mathcal{P} \iff (j,i) \in \mathcal{P}$.

Finding the global solution (3) is in general a hard, non-convex problem. We employ the efficient algorithm proposed in [13] to recover a local solution with a multi-start strategy to be more robust to bad local minima. The algorithm solves a relaxation of (3) and delivers probabilistic assignments of data points to clusters in place of the hard assignment matrix $\mathtt{Z}$.

4. EXPERIMENTS

We perform our cluster analysis on a dataset consisting of ECG data collected from 63 subjects (49 males and 14 females) with an average age of 20.68 ± 2.83 years. The subjects were 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 (more details in [5]). The signals were acquired using a bioPLUX research acquisition unit (12-bit resolution and 1kHz sampling frequency) shown in Fig. 3. The data consists of two independent acquisition sessions separated by a 3-month interval, entitled "T1" and "T2" [21]. We focused our analysis on "T1", which is composed by more than 10000 individual heartbeat waveforms.

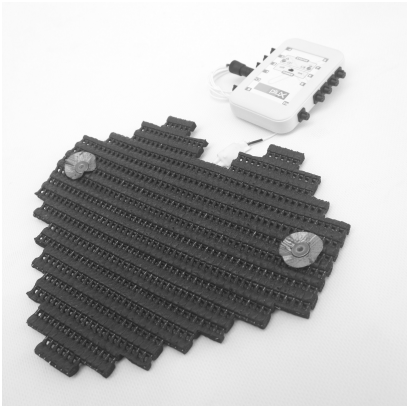


Fig. 3. Experimental apparatus consisting of the sensor pad with a heart shaped form factor, where a pair of Ag/AgCl electrodes are integrated, and a biosignal acquisition device.

We created two clustering ensembles $\mathcal{E}_1$ and $\mathcal{E}_2$, each with $\mathsf{m} = 200$ partitions, obtained by running the classical k-means algorithm [22] on the ECG dataset with different number of clusters, different initializations. Each ensemble is characterized by the use of a different distances: $\mathcal{E}_1$ with the euclidean distance, and $\mathcal{E}_2$ with one minus the cosine.The ensemble construction follows a split-and-merge strategy [9], applying a clustering algorithm with different number of clusters K randomly chosen from $\{K_{min}, \ldots, K_{max}\}$. A large value for K_{min} is used, which leads to partitions with high granularity and prevents the occurrence of clusters comprising different "natural" clusters. Splitting natural clusters into smaller clusters induces micro-blocks (smaller than the perfect block diagonal structures) in the $\mathtt{C}$ matrix, resulting in increased sparseness (lower density). The merging step is performed by the combination of the ensemble in the co-association matrix.

Fig. 4 (left) presents the co-association matrix obtained by combining the $\mathcal{E}_1$. Each line of the matrix represents an individual heartbeat, and the heartbeats of each individual are adjacent in the matrix, leading to a block-diagonal structure. There is some inter-subject noise, which is expressed by the

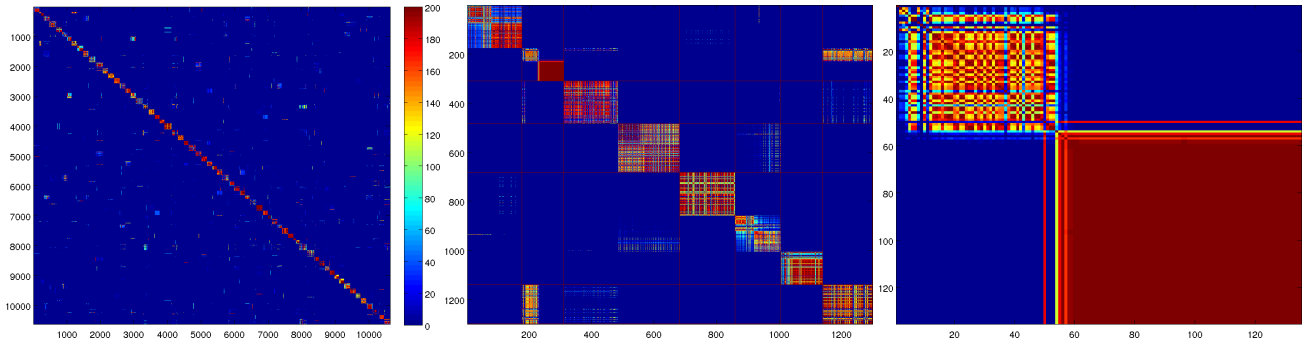
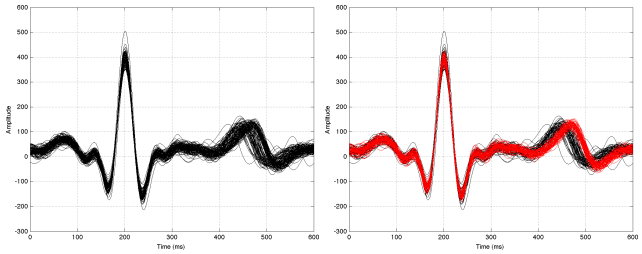


Fig. 4. Left: co-association matrix for ensemble $\mathcal{E}_1$. Middle: co-association sub-matrix of subjects 1-8. Right: co-association sub-matrix of subject 2.

of the diagonal elements in the co-association matrix. Moreover the intra-subject similarity is not constant as can be seen in Fig. 4 (right), where two different blocks are perfectly delineated for subject 2.

Fig. 5 shows the ECG of subject 2, showing why there are two distinctive clusters. These clusters are characterized by the different position of the T-wave, due to change on the heart-rate from the beginning to the middle of the acquisition. Note in the co-association matrix, where the lines are sorted by temporal order of acquisition it corresponds to the line where the block diagonal structures are delineated.



(a) ECGs (b) Clusters

Fig. 5. ECGs of subject 2: two distinctive clusters are easily perceived, characterized by the different position of the T-wave.

In Tab. 1, we summarized the performance of our algorithms after several runs, for ensemble $\mathcal{E}_1$ and $\mathcal{E}_2$, accounting for possible different solutions due to initialization, in terms of $\mathcal{H}$ index, a clustering validation index that gives the probability of agreement when the number of clusters is equal to the ground truth [23]. We fixed the final number of clusters $\mathsf{k} = 63$, the total number of subjects (equal to the ground truth). There is a significant variation in performance as the density $d = |\mathcal{P}|/\mathsf{n}^2$ of the co-association matrix changes. The best results are obtained with $d = 0.08$.

Fig. 6 illustrates the cluster assignments of heartbeat to clusters obtained from the consensus clustering algorithm with $d = 0.08$. Remind that the algorithm solves a relaxation

	1%	5%	8%	10%
$\mathcal{E}_1$	0.3502	0.8118	0.8226	0.8095
$\mathcal{E}_2$	0.3983	0.8318	0.8652	0.8657

Table 1. Accuracy as a function of the density and on the ensemble.

of (3), thus providing probabilistic cluster assignments in place of hard ones, as mentioned at the end of Sec. 3. For the sake of visualization, the matrix has been reordered to best fit the ground-truth solution. This matrix has 63 lines, each corresponding to one subject, and on the columns we have the heartbeats. It has a block-diagonal structure, where each block corresponds to each user's heartbeats. This indicates a good correlation with the ground-truth solution, as also highlighted by the quantitative result in Tab.1. There are, however, also cases where subjects are completely confused with others, like subject 20 (corresponding to line 20), where indeed the diagonal block is missing.

5. CONCLUSIONS AND FUTURE WORK

In this paper we apply a consensus clustering approach to the unsupervised analysis of an ECG-based biometric database. The results highlighted the natural groups within the population. Preliminary analysis enabled the discovery of groups characterized by different characteristics, as multiple heart rates. On going work focus the correlation of groups with user performance, in the biometric menagerie perspective.

REFERENCES

[1] V. Stanford, "Biosignals offer potential for direct interfaces and health monitoring," *Pervasive Computing, IEEE*, vol. 3, no. 1, pp. 99 – 103, jan.-march 2004.

[2] J. P. Clarys and J. Cabri, "Electromyography and the study of sports movements: A review.," *J Sports Sci*, vol. 11, pp. 379–448+, 1993.

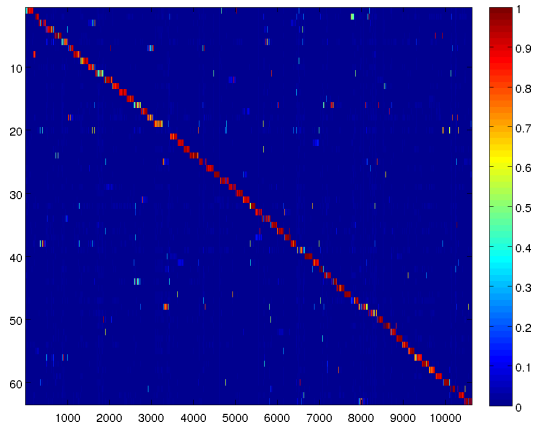


Fig. 6. Probability assignments obtained from the consensus clustering algorithm with $d = 0.08$ and $\mathsf{k} = 63$. Each line corresponds to a subject, and each column to a heartbeat.For the sake of visualization, the matrix has been reordered to best match the ground-truth solution.

[3] M. Schwartz and F. Andrasik, *Biofeedback: A Practitioner's Guide*, The Guilford Press, 3rd edition, June 2005.

[4] A. Freivalds, *Biomechanics of the Upper Limbs: Mechanics, Modeling and Musculoskeletal Injuries, Second Edition*, CRC Press, 2 edition, Feb. 2011.

[5] H. Silva, A. Lourenço, F. Canento, A. Fred, and N. Raposo, "ECG biometrics: Principles and applications," in *Int. Conf. on Bio-inspired Systems and Signal Processing - Biosignals*, Feb 2013, pp. –.

[6] A. Foteini, J. Gao, and D. Hatzinakos, *Biometrics*, chapter Heart Biometrics: Theory, Methods and Applications, Biometrics, InTech, 2011.

[7] I. Odinaka, P.-H. Lai, A. Kaplan, J. O'Sullivan, E. Sirevaag, and J. Rohrbaugh, "ECG biometric recognition: A comparative analysis," *IEEE Trans. on Information Forensics and Security*, vol. 7, no. 6, pp. 1812 –1824, Dec. 2012.

[8] D. Cuesta-Frau, J. Prez-Corts, and G. Andreu-Garcia, "Clustering of electrocardiograph signals in computer-aided holter analysis," *Computer Methods and Programs in Biomedicine*, vol. 72, no. 3, pp. 179 – 196, 2003.

[9] A. Fred and A.K. Jain, "Combining multiple clustering using evidence accumulation," *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 835–850, June 2005.

[10] A. Strehl and J. Ghosh, "Cluster ensembles — a knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, Mar. 2003.

[11] A. Kelarev, A. Stranieri, J. Yearwood, and H. Jelinek, "Empirical investigation of consensus clustering for large ecg data sets," in *Computer-Based Medical Systems (CBMS), 2012 25th International Symposium on*, June 2012, pp. 1–4.

[12] J.H. Abawajy, A.V. Kelarev, and M. Chowdhury, "Multistage approach for clustering and classification of {ECG} data," *Computer Methods and Programs in Biomedicine*, vol. 112, no. 3, pp. 720 – 730, 2013.

[13] A. Lourenço, S. Rota Bulò, N. Rebagliati, A. L. N. Fred, M. A. T. Figueiredo, and M. Pelillo, "Consensus clustering using partial evidence accumulation," in *Iberian Conf. on Pattern Recognition and Image Analysis*, June 2013.

[14] A. Lourenço, S. Rota Bulò, N. Rebagliati, A. L. N. Fred, M. A. T. Figueiredo, and M. Pelillo, "Probabilistic consensus clustering using evidence accumulation," *Machine Learning*, April 2013.

[15] S. Israel, J. Irvine, A. Cheng, M. Wiederhold, and B. Wiederhold, "ECG to identify individuals," *Pattern Recognition*, vol. 38, no. 1, pp. 133–142, 2005.

[16] T. W. Shen, W. J. Tompkins, and Y. H. Hu, "One-lead ECG for identity verification," *Proc. of the Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, vol. 1, pp. 62–63, October 2002.

[17] A. D. C. Chan, M. M. Hamdy, A. Badre, and V. Badee, "Wavelet distance measure for person identification using electrocardiograms," *IEEE Trans. on Instrum. and Meas.*, vol. 57, no. 2, pp. 248–253, Feb. 2008.

[18] D.P. Coutinho, H. Silva, H. Gamboa, A. Fred, and M. Figueiredo, "Novel fiducial and non-fiducial approaches to electrocardiogram-based biometric systems," *Biometrics, IET*, vol. 2, no. 2, pp. 64–75, 2013.

[19] A. Lourenço, H. Silva, and A. L. N. Fred, "Ecg-based biometrics: A real time classification approach," in *IEEE International Workshop on Machine Learning for Signal Processing*, September 2012, pp. –.

[20] A. Lourenço, H. Silva, and A. Fred, "Unveiling the biometric potential of Finger-Based ECG signals," *Computational Intelligence and Neuroscience*, 2011.

[21] H. Silva, A. Lourenço, A. L. N. Fred, and A. K. Jain, "Finger ecg signal for user authentication: Usability and performance," in *IEEE Int. Conf. on Biometrics: Theory, Applications and Systems - BTAS*, Sep 2013.

[22] A. K. Jain and R.C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, 1988.

[23] M. Meila, "Comparing clusterings by the variation of information.," in *Proc. of the Sixteenth Annual Conf. of Computational Learning Theory (COLT)*, Springer, Ed., 2003.