ECG Biometrics: A Dissimilarity Representation Approach

Francisco Marques

1Instituto Superior Técnico, Avenida Rovisco Pais, 1 1049-001 Lisboa

Electrocardiogram (ECG) biometrics are a relatively novel trend in the field of biometric recognition, comprising 13 years of development in peer-reviewed literature. This thesis studies all the inherent processes to an ECG biometric system: pre-processing, heartbeat segmentation, feature extraction and classification. For pre-processing, two filters are proposed. Three state-of-the-art segmentation algorithms are analysed and one chosen for consecutive use. This thesis subsequently introduces a novel ECG representation, based on a dissimilarity space, formed by taking pairwise dissimilarities between templates and input subjects signals. This representation can additionally exploit the potential extra information sourced from multiple ECG leads. For a 12-lead ECG database the hereby proposed dissimilarity-based method is compared to a single-lead state-of-the-art fiducial technique. Both these methods are, moreover, contrasted with another non-fiducial published method, for a single-lead, finger-obtained ECG database. Using the same k-NN classifier, results over the 12-lead database favour a dissimilarity-based representation. However, a multi-lead configuration did not prove itself advantageous. Contrastingly, over the noisier finger-obtained ECGs, the fiducial technique presented better results.

Index Terms—Electrocardiogram, ECG, Biometric System, Authentication, Identification, Fiducial, Non-fiducial, Heartbeat, R-peak, Segmentation, Autocorrelation, Linear-Discriminant Analysis, Multi-lead, Dissimilarity Representation, Baseline Noise, Power-line Noise.

I. INTRODUCTION

A. ECG Basics

Humans possess an intrinsic ability to recognize patterns and observations as part of a certain class. Often the class-assigning process is immediate and deemed obvious. However, when asked to explain how that conclusion was achieved, the human observer is unable to detail the iterations that led him to it.

A biometric system attempts to replicate this behaviour and pursue an even more demanding goal – subject identification. For that purpose a specific procedure must be designed which will employ sensors to measure the data requested by it.

In what concerns the field of Electrocardiogram (ECG) Biometrics the sensors record the heart’s electrical activity – the ECG itself – and build a procedure for feature extraction and classification. An ECG heartbeat is composed by three main components: P wave, QRS complex, and T wave, illustrated in Figure 1.

The measurement of an ECG is obtained via correct placement of a number of electrodes in specific parts of the body and by tracing voltages between them. Specific linear combinations between electrode potentials result in different leads. For instance, lead I corresponds to the voltage between the electrodes in the right and left arms, as stated by Einthoven’s Triangle [8] shown in Figure 2.

Fig. 2: Figure representing Einthoven’s Triangle and respective leads. The lead voltages are taken from the plus to the minus signs.

An ECG biometric system works just like any other biometric system or any supervised learning algorithm. Its functioning can be divided into two modes of operation – enrolment and classification.

During enrolment the system is given a user’s ID and extracts his/her information – the training data. From this data the classifier generates the templates which best represent the user and build the system’s database. More enrolled subjects imply greater difficulty for identification as the number of possible choices increases (and the probability of having an individual with similar data also increases). This effect can be countered by storing more templates for a given subject.

The second mode, classification extracts a current user’s data and compares it to the registered training data. It then outputs the best match, i.e. the correct user in the case of a biometric system. Classification can be divided into two procedures, authentication and identification.
**Authentication**
The system is provided an ID and will confirm or deny that a given user is who he/she claims to be. For that purpose the classifier performs only comparisons with respect to the given ID’s stored templates.

**Identification**
In this case, the classifier is not given any ID by the user. As a result, it must search its entire database for the subject whose templates best match the input data.

**B. ECG Biometrics**
All the studies on ECG biometrics are based on one-lead, two-lead, three-lead, or 12-lead ECG signals, mirroring what is most commonly employed in clinical situations [24]. The most widespread lead configuration used in ECG biometrics is by far the one-lead configuration as [3] demonstrated that a signal from one lead contains enough information to form the basis for a biometric system. Nonetheless, some studies have resorted to multiple-lead ECG signals in their biometric strategy. [34] utilized two leads [14, 33] opted for three leads and [3, 2] used 12 leads. The technique proposed in this thesis will be tested in a single and triple-lead scenario, employing the limb-lead I or all the limb-leads I, II and III respectively.

With an ECG signal in hand, regardless of the lead, there are several existing techniques for carrying out the feature extraction procedure. These fall into three different categories in what concerns the type of features they employ [24].

**Fiducial Features**
These methods use characteristic points of an ECG heartbeat and/or relationships between them as features – the fiducial features. Characteristic points include the peak of the R-wave, while a relationship between points can be the temporal duration of a QRS impulse. Several combinations of the four types of fiducial features have been used in the literature [30] – temporal, amplitude, angle and dynamic (R-R intervals) [24].

**Non-Fiducial Features**
Techniques based on non-fiducial features do not use characteristic points as features. Despite that, most of them rely on some characteristic points for heartbeat segmentation [5, 15] while others simply create windows from the ECG recording [25, 2, 18, 1]. Afterwards these segments/windows are transformed into another domain so as to extract features from the resulting signals. Also, heartbeat characteristic points might be used to aid in outlier removal. However, they will never be used as features or during classification.

**Hybrid Features**
Algorithms in this group resort to both fiducial and non-fiducial features for their biometric system. Some use a combination of them as features [32, 28]. Others design two classifiers where the first uses non-fiducial features to reduce the match set and the second outputs the classification while being fed with the fiducial features [27, 26].

Having assembled the feature space some algorithms additionally perform its dimensionality reduction. Afterwards, the reduced or raw feature space is then directly used for the training of classifiers. Regardless, it must be emphasised that the ideal feature representation is yet to be found – all the existing ones present pros and cons in relation to one another.

It is on these grounds that the current study introduces the concept of a feature space based on a Dissimilarity Representation for ECG biometrics by resorting to comparisons between signals as inspired by the study in [12]. The current study explores this possibility by designing a multi-lead configuration and comparing it with the single-lead version. Dissimilarity Representations will be explained in Section III.

Moreover Section II describes the notation used in this paper. Section IV then outlines the pre-processing feature extraction and template generation carried out on the ECG signals, as well as the utilized classifier. Section V outlines the executed experiments and their results. This paper is concluded in Section VI which draws the main findings and conclusions.

**II. Notation**
In order to better understand the methodology hereby proposed, as well as the process of building the dissimilarity representation, the notation that will be employed throughout this work is here presented.

The basic element behind the proposed method’s biometric system is the ECG heartbeat. Heartbeats will be employed as templates so as to create the dissimilarity representations. With this in mind we shall consider:

- A population of $S$ existing subjects;
- A percentage $p$ which is considered sufficient to represent the whole population variation. From this percentage, a number of subjects $S_p$ will be randomly chosen from $S$. For this work $p$ is set to 15%, which was considered sufficient to represent the ECG variation within a large population.
- A set $C$ of leads. For the scope of this work, $C$’s elements will be either the single-lead $[l]$ or the 3 limb-leads $[I, II, III]$. The variable $L$ is defined as $L = |C|$.
- $N_i$ as the number of extracted heartbeats from subject $i$’s ECG, $i = 1, \ldots, S$.
- $N = \sum_{i=1}^{S} N_i$ as the number of all extracted heartbeats over all subjects.
- A heartbeat belonging to subject $i$ is denoted by $h_{ij}^l$ with $j = 1, \ldots, N_i$ and $l \in C$. The variable $h_{ij}$ implies a heartbeat originating from lead $l$. Beats are represented by a 600 ms window which is built having the R-peak as a reference at position 200 ms.
- A reference lead, $R$, which for the scope of this work will be lead $I$. Note that $h_{ij}^R$ will correspond to signals taken from lead $R$.
- A Feature Space $F$, represented by an $N \times M$ matrix. For example, $M$ consists of the 300 samples present in a heartbeat, if the used sampling rate is 500 Hz.

Two metrics have been employed in various steps of the proposed methodology – the euclidean distance and the cosine similarity. For the sake of clarity they are respectively defined here in Equations 1 and 2.

$$ D(h_{ij}, h_{ik}) = \sqrt{(h_{ij} - h_{ik})^T (h_{ij} - h_{ik})} $$  

$$ D(h_{ij}, h_{ik}) = \frac{h_{ij} - h_{ik}}{\sqrt{h_{ij}^T h_{ij}} \sqrt{h_{ik}^T h_{ik}}} $$
\[ D(h_{ij}, h_{ik}) = 1 - \frac{h_{ij} \cdot h_{ik}}{\|h_{ij}\| \|h_{ik}\|} \]  

where \(\|\|\) represents the euclidean norm.

III. DISSIMILARITY REPRESENTATION

A Dissimilarity Representation is built on the fact that similarity between objects plays a crucial role in class formation, i.e. a class is a set of similar objects [13]. A universal object similarity, however, does not exist and always depends on the classification context, procedure and/or the domain of study. Moreover, the presence of other classes will influence the degree to which an object should or should not be assigned to a particular class.

This study puts forward the notion of dissimilarity between ECG elements. Calculating a dissimilarity is simply comparing elements, pairwise, according to some pre-defined rules [13]. Metrics, for instance, fit this criteria. The present study will explore two different ones – euclidean distance and cosine similarity.

A dissimilarity-based representation can be constructed from any type of elements which can be any kind of feature array possible. Also, they can be built from as many comparisons between elements as one wishes. Consequently, note that this process is easily extensible to using more than one-lead by simply comparing one lead’s elements with another lead’s elements.

A dissimilarity space intends to take the original feature space \(F\) and output another one, \(F_D\), by taking pairwise distances between ECG elements \(i\), where \(i = 1, \ldots, N\). This thesis proposes two different approaches for defining the dissimilarity space. Subsection IV-D details their development.

IV. PROPOSED METHODOLOGY

The whole methodology for the proposed technique is summarized in Figure 3.

A. Pre-processing

Due to the presence of several noise sources during measurement – power line interference, electrode contact loss, baseline drift due to respiration and motion artefacts, for example [16]– it is imperative to filter the signal to facilitate the procedure of feature extraction.

In order to remove baseline drift from raw ECG signals we apply the method by [10]. It applies two median filters (of 200 ms and 600 ms respectively) to the raw ECG signal. The first removes the effect of the QRS waveforms, while the second deletes the T-waves from the signal. Without these waveforms, the resulting signal represents the estimation of the baseline of the raw ECG, which is subtracted to it.

After the application of this step, an additional filtering stage eliminates the effect of other noise sources through bandpass filtering. Two FIR filters have been designed, both with order 150 for a sampling rate of 500 Hz but different bandwidths. Other sampling rates will be used for other ECG databases, in which the filter order changes accordingly. Both filters’ influence over classification will be analysed in Section V.

1) Method A

Method A employs a FIR band-filter of bandwidth \(2 \rightarrow 40\) Hz. After FIR filtering the ECG, the effect of EMG is still visible. As such, a simple moving-average filter proposed in [6] is applied to the signal. It acts in 28 ms intervals with a first zero at about 35 Hz. Some unwanted distortion, such as R-peak height reduction, is caused by this step of the filter. Nonetheless, this filter preserves ECG information very well.

2) Method B

The FIR band-filter in Method B has a bandwidth of \(5 \rightarrow 20\) Hz. This restricted bandwidth loses some valuable ECG information, However, this filter very efficiently reduces both EMG and motion artefact influence. It is here hoped this noise reduction compensates the information loss.

B. Segmentation

In order to segment the ECG signal into heartbeats, R-peak detection is first carried out. Versions of three state-of-the-art segmentation algorithms were developed [6, 17, 23]. These techniques were compared in terms of execution time, and performance over the MIT-BIH arrhythmia database. The algorithm in [17] was chosen as the most suitable for this work, and will be outlined in the current Sub-section. In [17], beat detection is performed over a transformed signal which is obtained according to the diagram in Figure 4. Its basic detection rules (represented by the last block in Figure 4) are as follows:

![Fig. 4: Transformation process applied to the ECG signal so as to perform beat detection.](image)
1) Ignore all peaks that precede or follow larger peaks by less than 200 ms.
2) If the peak occurred within 360 ms of a previous detection and had a maximum slope less than 0.7 times the maximum slope of the previous detection assume it is a T-wave.
3) If the peak is larger than the detection threshold – DT – call it a QRS complex; otherwise call it noise.
4) If an interval equal to 1.5 times the average R-to-R interval has elapsed since the most recent detection, check for a peak larger than DT within that interval. If the peak followed the preceding detection by at least 360 ms then classify that peak as a QRS complex.

The detection threshold – DT mentioned in rules 3 and 4 – resorts to the following data structures:

**QRS-peak buffer**
Stores the 8 most recent R-peak values. Its entries are used in the detection threshold – DT – calculation. It is initialized with the highest-valued peaks in one second intervals for 8 seconds.

**Noise-peak buffer**
Behaves like the previous structure but stores the 8 most recent noise-peak values instead. It is, however, initialized at 0.

**RR-interval buffer**
Stores the 8 most recent interval between R-peaks. These are initialized at a value corresponding to a 1 s interval.

The DT threshold is computed as:

$$DT = Noise\_Peak\_Buff_{med} + TH(QRS\_Peak\_Buff_{med} - Noise\_Peak\_Buff_{med})$$  \hspace{1cm} (3)

where Noise\_Peak\_Buff_{med} and QRS\_Peak\_Buff_{med} are the median of the Noise-peak buffer and QRS-peak buffer arrays respectively. \(TH = 0.45\) was empirically found most suitable for the database used in testing.

The peaks detected on the transformed signal do not map directly to the R-peaks in the ECG signal. The points found are mapped to the exact R-peak location by searching within ±150 ms and choosing the highest sloped peak of the two largest ones found in that interval. This process is executed only for lead I – the beats are then matched on the other utilized leads by once again looking for the highest sloped peak of the two largest ones, but this time in a smaller interval consisting of ±60 ms.

Having the location of the R-peaks for all leads, the segments are constructed simply by taking the ECG window from 0 ms to 600 ms, where 200 ms correspond to the R-peak. As mentioned in Section II, all segments have a fixed length of 600 ms.

1) **Outlier Detection**
Outlier removal is performed as in [21]. It will only be executed for lead I – the beats here discarded will be discarded for the other employed leads as well.

This algorithm receives as input the \(N_i\) heartbeats for subject \(i\) and begins by calculating the average beat via:

$$h_{i}^{\text{av}} = \frac{1}{N_i} \sum_{j=1}^{N_i} h_{ij}, \hspace{1cm} (4)$$

Heartbeats which stray away from the average beat are discarded according to the following procedure:
1) For each \(h_{ij}\), compute its distance \(D(h_{ij}, h_{i}^{\text{av}})\) to the mean waveform \(h_{i}^{\text{av}}\).
2) Compute the 1st and 2nd order statistical moments of the distances \(D(h_{ij}, h_{i}^{\text{av}})\): \(\mu_{D_i(h_{i}^{\text{av}})}\) corresponds to the mean value and \(\sigma_{D_i(h_{i}^{\text{av}})}\) to the standard deviation.
3) Compute the median of the minimum and maximum values over all templates \(h_{ij}\), denoted as \(h_{i}^{\text{med}}\) and \(h_{i}^{\text{med}}\) respectively.
4) Verify the conditions below for every \(h_{ij}\). If any is confirmed, then \(h_{ij}\) is discarded as an outlier:
   a) \(h_{ij_{\text{min}}} < 1.5 \times h_{i}^{\text{med}}\), \(h_{ij_{\text{min}}}\) is the minimum value for beat \(h_{ij}\);
   b) \(h_{ij_{\text{max}}} > 1.5 \times h_{i}^{\text{med}}\), \(h_{ij_{\text{max}}}\) is the maximum value for beat \(h_{ij}\);
   c) \(D(h_{ij}, h_{i}^{\text{av}}) > \mu_{D_i(h_{i}^{\text{av}})} + 0.5 \times \sigma_{D_i(h_{i}^{\text{av}})}\);

Note that \(D(h_{ij}, h_{i}^{\text{av}})\) can refer to any distance metric. The present work utilizes the cosine distance in 2.

**C. Template Generation**
The problem of template selection may be posed as follows: given a set of \(N\) heartbeats, select \(K\) templates that “best” represent the variability as well as the typically observed patterns according to a given similarity criterion [20].

Clustering methods are especially adequate for this task, and have already been used for template selection in other modalities [31, 7, 19, 22]. In this thesis the K-means algorithm was used, with \(K\) empirically set to 5, the cluster’s centroids being used as templates [20]. Dissimilarity representations are calculated by taking comparisons between input beats and these reference templates.

**D. Dissimilarity Computation**
Dissimilarities can be calculated by measuring a distance between beats according to a metric. Two such metrics have been explored in this study – euclidean distance and cosine similarity referred respectively in Equations 1 and 2. In Section V their effects can be contrasted.

Two different dissimilarity extraction techniques are hereby proposed.

**Subject based**
This first and simplest approach computes the distance \(D(h_{ij}^R, h_{ij}^l)\) between each segment \(h_{ij}^R\) and the set of \(h_{ij}^l\) template beats for each lead in set \(C\). The process is repeated for all subjects \(S\). It is presented in pseudo-code below.

```plaintext
for each subject \(i\) in \(S\) do
    for each beat \(j\) in \(N_i\) do
        \(d_{ij} = []\)
        for each lead \(l\) in \(C\) do
            for each template \(t\) do
```

```plaintext
                      ... (Pseudo-code continues)
```

The above formula is used to compute the dissimilarity between the template and the beats within each subject. The dissimilarity measure is then used to classify the beats into clusters. This process is repeated for all subjects, resulting in a set of templates for each subject.
Considering $T$ templates per subject, the resulting dissimilarity representation is then an $N \times T^*L$ matrix composed by $N$ dissimilarity arrays with $T \times L$ components.

**Inter-subject based**

A second strategy computes the distance $D(h^R_{ij}, h^L_{it})$ between each segment $h^R_{ij}$ and the set of $h^L_{it}$ template beats of the randomly chosen $S_p$ subjects for each lead in set $C$. The following pseudo-code outlines these steps.

```plaintext
for all $S$ subjects do
  for each beat $j$ in $N$ do
    for each lead $l$ in $C$ do
      for each subject $s$ in $S_p$ do
        for each template $t$ do
          $d_{ij,l} = \text{append}(D(h^R_{ij}, h^L_{it}))$
```

Once again, considering $T$ templates per subject, the obtained dissimilarity representation is an $N \times T^*S_p^*L$ matrix consisting of $N$ dissimilarity arrays composed by $T \times S_p \times L$ elements.

**E. Classification**

In classification, both authentication and identification follow the same principles for matching. A feature space comprising dissimilarities supports a large variety of classifiers [11]. The current work employs a $k$-Nearest Neighbours ($k$-NN) model applied to dissimilarity arrays, with $k$ set to 3.

The steps taken during the classification process depend on the dissimilarity representation approach taken in Sub-section IV-D. During enrolment, for a Subject based approach, the classifier will store all the $S$ users’ template heartbeats $h^L_{it}$ for all $C$ leads.

As for the Inter-subject method the algorithm saves only the template beats belonging to the $S_p$ subjects (which have been recorded prior to enrolment) over the $C$ used leads.

```plaintext
The obtained dissimilarity arrays are called $v_{id}$. Then, the template dissimilarity arrays $v_{t}$ chosen for the determination of the 3-NN are either sourced from the input subject – for authentication – or obtained from all subjects – for identification.

Distances between dissimilarity vectors $D(v_{id}, v_{t})$ are once again measured according to both euclidean distance and cosine similarity referred in Equations 1 and 2. In Section V their effect in classification is compared as well. See Figures 6 and 7 for the classification procedure for the Subject based and Inter-subject based dissimilarity approaches respectively.

After calculating all the $D(v_{id}, v_{t})$, the 3 smallest distances are taken as the 3-Nearest Neighbours. They will then be compared with a threshold, $th_{auth}$ for authentication or $th_{id}$ for identification. This threshold will validate the distances’ votes according to:

$$d_k <= th$$

where $d_k$ is one of the 3 resulting distances and $th$ is either $th_{auth}$ or $th_{id}$ according to the chosen mode. Distances $d_k$
not respecting Equation 5 are not considered in the final classification. If at least 2 $d_k$ distances have been validated, then for authentication the input user is confirmed as valid. In identification the most voted user corresponding to those $d_k$ is provided as output. However, for this mode, if no majority exists for the valid $d_k$, then identification fails.

V. RESULTS

Two ECG databases were employed for testing the developed ECG biometric systems. They are described below and from here on they will be addressed according to the label here assigned.

Hospital de Santa Marta (HSM)

The HSM database consists of records from a local hospital, Hospital de Santa Marta, specialized in cardiac issues. All signals were acquired using Philips PageWriter Trim III devices, following the standard 12-lead placement, with a sampling rate of 500 Hz and 16 bit resolution. These records were obtained during normal hospital operation and some originate from ill patients. This study focuses only on signals originating from healthy individuals. Each record is composed by ECGs from all 12 leads, with a duration of little more than 10 s. 832 records were then employed belonging to 618 subjects.

Cruz Vermelha Portuguesa (CVP) [9]

CVP’s ECG data was acquired at the fingers with dry Ag/AgCl electrodes. ECG acquisition was performed using a custom two lead differential sensor design with virtual ground, found in [29]. Raw ECGs were recorded via a bioPLUX research, Bluetooth wireless biosignal acquisition unit; the device was configured to employ a 12-bit resolution and 1 kHz sampling frequency. Signal acquisition was made in two sessions separated by a 3-month interval, comprising a total of 63 subjects. This population is composed by 14 healthy males and 49 healthy females, with ages ranging between 18 and 50 years (20.68 ± 2.83). In each session, the subjects sat in a resting position for 2 min, with two fingers, one from each hand, placed on each of the electrodes. The first session will be called T1 & Sitting while the second T2 & Sitting.

Over both databases, the various results are compared through Receiver Operating Characteristic (ROC) and corresponding Equal Error Rates (EER) for the authentication mode of operation while the identification mode is characterized by its Identification Error (EID) as in:

$$E_{ID} = \frac{F_{ID}}{T_{ID} + F_{ID}}$$  \hspace{1cm} (6)

where $F_{ID}$ and $T_{ID}$ correspond to the number of incorrect and correct identifications respectively.

The hereby developed dissimilarity based approaches are tested on both databases. The Subject based approach will be tagged with letter $U$ while an Inter-subject based will be referred by $I$.

Both databases will also test a fiducial technique proposed by Carreiras et al. [4], the Simple Heartbeat Comparison (SHC). This system is similar to the one presented in this thesis but ditches the dissimilarity computation phase. Instead, it uses heartbeats as feature arrays and then applies $k$-NN, with a cosine similarity metric, directly to them. Also this method is exclusively single-lead. It shall be tagged with the expression $O$.

For the CVP database, an additional method will be tested – the Auto-Correlation/Linear-Discriminant Analysis (AC/LDA) algorithm, designed by Agrafioti [1]. This is a non-fiducial method which ditches R-peak detection altogether. Instead, it segments the ECG record into 5 s non-overlapping windows. Auto-correlation is computed for these windows over a set of 100 lags. Linear-Discriminant Analysis (LDA) is executed on them so as to reduce their dimensionality. An identical $k$-NN algorithm is then applied over these transformed ACs. This method is tagged with the expression $AC$.

A. HSM

All of the designed experiments were executed according to a simple random approach. The existing valid heartbeats were randomly partitioned into two sets, a training set containing 75% of those beats and a test set composed by the remaining 25%, whose beats are individually evaluated. This procedure was repeated 10 times and the average EER and EID were calculated from all these 10 runs. The minimum number of training heartbeats is 5, an amount demanded by the template generation block. Some subjects did not satisfy this requirement and were thus not considered for testing. The employed number of subjects thus becomes 503.

The following experimental set-ups were designed. The effects of the metric choice over both dissimilarity computation in Section IV-D and classification in Section IV-E were tested. The chosen metrics for both steps were the cosine similarity metric, directly to them. Also this method is exclusively single-lead. It shall be tagged with the expression $U$.
Due to time constraints, the set of metrics corresponding to the best results, for each dissimilarity method, were employed for $C = [I]$ and for $C$ consisting of all 12 leads. The total number of dissimilarity-based experiments is then 12. Tags are given to these experiments so as to facilitate the observation of results.

**Subject based ($U$)**

The employed metric scenario for a given experiment will be described by two letters, $C$ or $E$, whether the employed metric is a cosine similarity or euclidean distance respectively. The first letter will correspond to the metric utilized in the dissimilarity computation process, while the second refers to the metric used in the classification procedure. The utilized set $C$ is shown subscripted. As a result, a tag of $U_{II,III} - CE$ pinpoints the use of cosine similarity for dissimilarity calculation and the use of euclidean distance to get the distances between dissimilarity arrays, for a set $C = [I,II,III]$. As a result, for approach $U$ there are:

- $U_{II,III} - CC$
- $U_{II,III} - CE$
- $U_{II,III} - EC$
- $U_{II,III} - EE$
- $U_I - CE$
- $U_{All} - CE$

**Inter-subject based ($I$)**

The tag $I$ was attributed to experiments following this approach. The notation associated to them follows the same as in the previous method. As such:

- $I_{II,III} - CC$
- $I_{II,III} - CE$
- $I_{II,III} - EC$
- $I_{II,III} - EE$
- $I_I - CC$
- $I_{All} - CC$

Also, a tag of $O_I$ was given to the SHC method, due to it being single-lead. Table I summarizes results obtained for all experiments in the form of averaged EERs and EIDs and respective standard deviations.

<table>
<thead>
<tr>
<th></th>
<th>EER [%]</th>
<th>EID [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O_I$</td>
<td>8.51 ± 0.30</td>
<td>12.04 ± 0.68</td>
</tr>
<tr>
<td>$U_{II,III}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CC$</td>
<td>21.55 ± 0.36</td>
<td>87.97 ± 2.21</td>
</tr>
<tr>
<td>$CE$</td>
<td>4.45 ± 0.13</td>
<td>8.89 ± 1.11</td>
</tr>
<tr>
<td>$EC$</td>
<td>30.25 ± 0.34</td>
<td>92.59 ± 1.78</td>
</tr>
<tr>
<td>$EE$</td>
<td>1.53 ± 0.09</td>
<td>23.46 ± 1.96</td>
</tr>
<tr>
<td>$U_I$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CE$</td>
<td>4.45 ± 0.09</td>
<td>9.92 ± 0.76</td>
</tr>
<tr>
<td>$U_{All}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CE$</td>
<td>4.26 ± 0.16</td>
<td>9.22 ± 1.29</td>
</tr>
<tr>
<td>$I_{II,III}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CC$</td>
<td>2.46 ± 0.09</td>
<td>5.48 ± 0.33</td>
</tr>
<tr>
<td>$CE$</td>
<td>4.76 ± 0.24</td>
<td>6.90 ± 0.38</td>
</tr>
<tr>
<td>$EC$</td>
<td>10.75 ± 0.06</td>
<td>13.02 ± 0.68</td>
</tr>
<tr>
<td>$EE$</td>
<td>3.83 ± 0.11</td>
<td>19.74 ± 0.50</td>
</tr>
<tr>
<td>$I_I$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CC$</td>
<td>2.46 ± 0.09</td>
<td>5.37 ± 0.48</td>
</tr>
<tr>
<td>$I_{All}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$CC$</td>
<td>2.50 ± 0.06</td>
<td>5.32 ± 0.57</td>
</tr>
</tbody>
</table>

**TABLE I: All experiments’ EER & EID rates over the HSM database.**

The ROC curve for the experiment $U_{II,III} - EE$ is shown in Figure 8(a). A ROC curve comparing all the metrics over both approaches, for authentication mode, is shown in Figure 8(b). This plot only shows experiments for set $C = [I,II,III]$, with the exception of $O_I$, of course.

Fig. 8: ROC curves outlining obtained results for authentication.

The results in Table I suggest the following observations.

- Both single-lead approaches $U_I$ and $I_I$ present better results than those obtained from $O_I$. From here, it is possible to conclude that a dissimilarity based representation can originate a more effective biometric system.
- Dissimilarity-based classifiers whose training computes dissimilarities based on a cosine metric proved much superior to their euclidean distance equivalents. This was expected due to the large initial feature size of $M = 300$, as a cosine metric better outlines shape and signal orientation. An euclidean metric focuses on distance which for an array of 300 samples is harder to measure. The exception concerns authentication results for $CE$ vs $EE$, in which the latter is better.
• For a Subject-based dissimilarity approach, experiments employing a cosine similarity metric during classification present very high EER and EID indices. This was expectable due to the very small dissimilarity array size \( (T \ast L = 15) \) rendering the cosine similarity metric unable to extract viable information. Precisely the opposite can be said for an Inter-subject approach which presents a dissimilarity array size of \( (T \ast S \ast L = 1125) \).

• Authentication EERs did not vary much by changing the lead-set \( C \), over the dissimilarity-based classifiers. EIDs changed slightly but neither consistently nor significantly. These deviations can be attributed to the randomness behind the template selection process, outlined in Section IV-C. Also, in case of identification, potential ties result in random ID choices, further triggering variability.

• Lastly, the lowest EER for authentication originated from \( U_{I,I,III} – EE \). However it also presents a high EID. The lowest EID is shown by \( I_{I,III} – CC \) which also gave a good and second lowest EER value.

B. CVP

There are two possible experimental settings for this database – within and between sessions. Both of them follow a simple random approach and are outlined below.

Within Session

In this experiment set, testing is only performed over signals from session T1 & Sitting. Valid heartbeats or AC segments (depending on the method) are partitioned into two sets, training and testing, each containing a random half of the ECG elements. This procedure was repeated 10 times, so as to account for the variability in set formation. Average EER and EID rates were calculated from all these 10 runs.

Between Session

For this set-up, training heartbeats or AC segments are taken from records in T1 & Sitting. Testing is performed over T2 & Sitting’s elements. Here, only 1 run is performed, as no variability amongst training and testing sets exist.

Furthermore, two filters were developed in Section IV-A, Method A and B. The outcome in classification these two filters originate will be contrasted for tests over this database.

With this in mind, four experiments were run for each biometric classifier, combining both sessions and both filters. For the dissimilarity-based classifiers, the only approach returning best results was employed – Inter-subject based \( I \). As for the metric set, the reduction in the number of subjects in the database suggested the change from metric set \( CC \) to \( CE \). Furthermore, due to the greater number of heartbeats per subject, 10 templates were used rather than 5, for the template generation block outlined in Section IV-C. Thus, this method is referred by \( I^{CE} \). Table II summarizes obtained results. Figures 9(a) and 9(b) show ROC curves for all authentication experiments carried over the filter in Method A.

Obtained results suggest the following observations:

• Choosing the filter in Method A for Pre-processing substantially improves results among identical experiences over the technique \( O \). For the other methods, \( AC \) and \( I – CE \), this distinction is not noticeable – note the high standard-deviation rates for these modes.

• The SHC technique presents the lowest EER and EID rates underlining its better robustness to both noise and intra-subject variability.

• The AC/LDA method seems to present highest EER and EID rates. This can be explained by its windowing process coupled with the characteristics of this database. This technique demands windows of 5 s. CVP’s records present a lot of noise and it will be difficult to have entire 5 s windows devoid of noise. This not only causes some ECG records to have very few valid AC segments, but also hinders the quality of the extracted features.

• A dissimilarity based approach also outputs high EER and EID rates. Once again, CVP’s signals are very noisy, which results in high intra-subject beat variability. Subject-representative templates will try to account for this variability, causing potential heartbeat overlap with other subjects. Thus, the distinction between subjects becomes blurred which raises EER and EID rates.

• Between session set-ups returned substantially higher EERs and EIDs than Within session experiments. This underlines one of the biggest problems of ECG biometrics – intra-individual variation over time.

• Lastly, these signals originate much larger EER and EID rates than the HSM database. They were acquired from a finger-electrode set-up which introduces much more noise, and severely hinders classification. Note the much smaller subject number for this database in relation to HSM (62 vs 503).

VI. Conclusions and Future Work

The number of possible ECG representations is endless and so far none has managed to stand out at the expense of all the others. In this thesis, a new ECG representation space is developed and integrated into an already existing biometric system. This feature space is built through dissimilarity computation, where the new features are a direct and pairwise comparison between those present in two ECG signals, which here were taken via metrics.

Moreover, the computation of this novel representation can be extended to various types of ECG configurations or signals, underlining its versatility. The current study extended its usage to multi-lead ECG signals, where an EER rate of 1.53% has
been achieved, for authentication over a database of 503 subjects, as well as an EID rate of 5.65%. It should be emphasized that for authentication/identification, a single heartbeat was used. The usage of a larger number of beats for classification will likely lead to better results. When contrasting with the original technique, which does not compute a dissimilarity representation, this feature space returns better results, proving the usefulness of such a representation. However, as in previous work, the usage of more than one lead did not significantly improve results.

On another note, the convenience offered by measuring ECGs through fingers is immense. Results from the CVP database show that it is not yet possible to perform adequate biometrics over these signals. Improving the ECG acquisition set-up in relation to the one employed for CVP’s records is of high importance. Also, testing should be performed over other finger-obtained ECG databases in order to find out if the trend here shown for CVP is maintained.

Lastly, temporal variation among ECGs from the same subject remains a big hindrance, if not the biggest, for ECG biometrics. This problem underlines the need to perform regular (probably monthly) re-enrolment of every subject’s ECGs, in order to maintain biometric efficiency. Future work should also focus on this problem and look for the potential heartbeat features offering more stability over time.

ACKNOWLEDGMENTS

My first thanks will go to Instituto Superior Técnico whose strict excellence throughout these 5 years of coursework allowed me to develop this work. Regardless, none of it would have been possible without Professor Ana Fred who always found time to support me and my setbacks, despite her great workload always pressing her.

Some very special thanks go to André Lourenço and Carlos Carreiras who always gave great suggestions to my workflow. They were also invaluable for briefing me into their software system and for result extraction.

REFERENCES


