Cancelable Biometrics: A dissimilarity based approach using hermite polynomial approximations

Afonso Alexandre Vicente Clara
Instituto Superior Técnico - Universidade de Lisboa

Abstract

Nowadays, the demand for robust and secure identity recognition systems is increasing. Cancelable biometric systems emerge as one of the possible answers in this context. The present work aims to develop a new system of this kind applied to electrocardiogram (ECG) signals, based on two major techniques: a) exploring hermite polynomial approximations of the ECG; b) a dissimilarity based representation. The hermite polynomial approximation is applied directly to the ECG records, leading to a new representation, avoiding original data storage. The dissimilarity based representation can be applied to any characteristic which fulfils the criteria established by biometrics and its recognition process (authentication and identification) is done in a different space from the original one, granting security and robustness to the proposed system.

This methodology was applied to a database with records regarding 612 subjects, evaluating the influence of several parameters related with both the hermite polynomial approximation and dissimilarity based representation. The system was compared with a classic biometric system as well as with a cancelable one using the electrocardiogram.

The system had a good performance regarding authentication. When the hermite polynomial approximations were presented as input to the system, a slightly decrease on performance was verified when compared with the system using the original records.

Using a biometric menagerie technique, it was possible to make conclusions regarding the quality of the records contained on the used database.

Keywords: Electrocardiogram, cancelable biometrics system, dissimilarity based representation, hermite polynomial approximation.

1. Introduction

Nowadays, there is an increasing demand to have robust and secure identification/authentication systems, hereafter referred globally as recognition systems. This happens because there is an increasing number of processes which are not directly controlled by human beings. Several examples can be found: having access to our banking account through an ATM machine or internet; having access to a certain restricted area; or a device being able to adapt itself according to a given user. During a long period of time, these systems were based on credentials, such as the use of identification documents or PIN's. Obviously, these kind of procedures were not able to meet the growing demands regarding security in many applications, like border crossing, access control, financial transactions or telecommunications [1].

In a first attempt, biometric systems emerged as a possible answer to the issues mentioned above. Some advantages of these systems are: biometric information cannot be lost or forgotten in contrast to cards or PIN's; it is not easy to forge this kind of information and from the user's perspective, it is not necessary to remember any password. However they present some flaws too: when the system is compromised and someone gains access to the database information, there is a permanent biometric data compromise, since this kind of information cannot be revoked [2].

To give an answer to the revocable issues, cancelable biometric systems have emerged. This approach has two main purposes: on the one hand, tries to develop recognition systems more difficult to forge or copy; on the other hand, it is able to protect the user's information since that information is not directly stored in the database, granting confidence to the system. From this perspective, it is possible to point some advantages of these systems. They are more secure and more difficult to crack than the traditional biometric systems. Besides that, it could be expected that these kind of systems raise more social acceptance regarding the traditional biometric systems, since the biometric data is not kept in the database, proving privacy to the users.

The main purpose of the present work is to develop a new cancelable biometric system able to give a response to the problems raised above. Thus, two main fields will be explored. On the one hand, a dissimilarity based approach is used as a process of recognition. On the other hand, hermite polynomial approximation are applied to the electrocardiogram signal (ECG) waveform. These two techniques are able, separately or set together, to guarantee original data protection. It is important to note that the dissimilarity based approach can be applied not only to the ECG signal, but to any other signal which may be used to perform subject recognition.

2. Founding Concepts

2.1. Biometric Systems

One can define biometrics as automated recognition of individuals based on their behavioural and/or physiological characteristics (ISO/IEC JTC1 SC37). Not all human characteristics are eligible to be considered as a biometric feature. To be applied to a biometric system, the behavioural or biological characteristic has to fulfil some properties, namely [2]:

- *Universality* - every subject has that characteristic.

- *Uniqueness* - there are not two persons with the same terms of a given characteristic.
- *Permanence* - the characteristic is invariant along time.
- *Collectability* - the characteristic can be measured quantitatively.
- *Performance* - the ability to, given a certain characteristic, the system performs accurately.

Examples of human characteristics which can be applied to a biometric system are: fingerprint, palm print, signature, face, iris, voice or hand geometry.

The architecture of a biometric system follows a typical design of a pattern recognition system. A typical biometric system has two different operation modes or phases: enrolment and recognition which can be an authentication or an identification process.

At the enrolment phase (see figure 1), the system has the ability to build a database from data belonging to a certain subject to whom identity is known and also provided to the system. For each subject, a certain number of templates can be generated as a way to best represent that individual [3].

In authentication mode (see figure 2), a subject provides to the system his/hers biometric measurement and a claimed ID, and the system tries to match that data only with the stored one referring to that subject, and come up with a decision [3]. This decision is based in a threshold, meaning if the resulting distance measure regarding the comparison between the provided data and the template(s) is below of a certain defined threshold, the subject is accepted by the system.

In identification mode (see figure 3), the system measures a certain biometric characteristic from the subject, tries to match the data with the one stored and eventually is able to identify the subject [3]. This operation mode is more susceptible to errors, since the system needs to compare the provided data whit the data belonging to all subjects from the database.

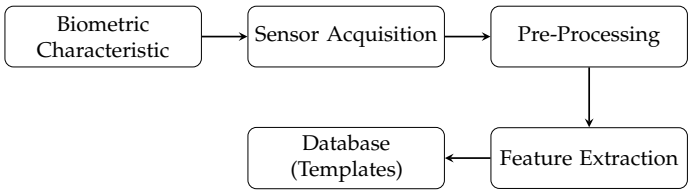

Figure 1: Biometric system architecture, when working on enrolment mode.

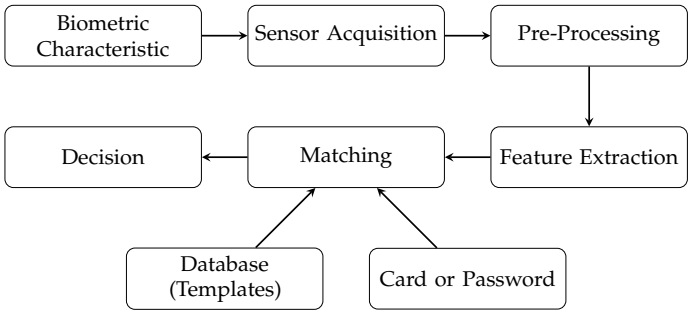

Figure 2: Biometric system architecture, when working on authentication mode.

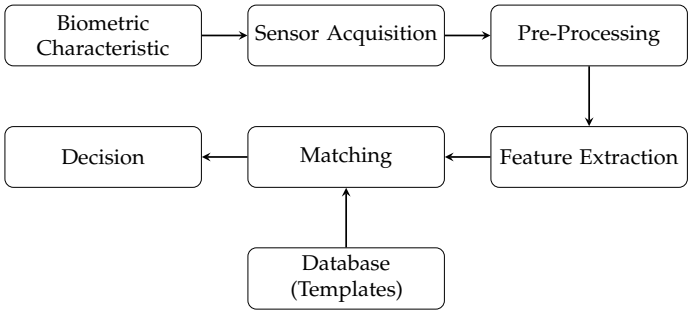

Figure 3: Biometric system architecture, when working on identification mode.

2.2. Performance

When one develops a biometric system, it is important to evaluate its performance. Its accuracy depends not only on the system's design but also on the data used to test the system [3]. Thus, when one is designing this kind of systems, it is important to keep in mind that there is not a perfect configuration and the design should reflect the environment for which the system is being developed. In order to test the system's performance, in the authentication mode, it is possible to compute some parameters. When a system is being tested, four possible scenarios can occur [3]: a genuine individual is accepted, meaning a true positive occurred (TP); a genuine individual is rejected, a false negative (FN); an impostor individual is accepted, a false positive (FP); or an impostor individual is rejected, a true negative (TN). With these four measures, it is possible to calculate some rates, namely the false acceptance rate (FAR) and the false rejection rate (FRR), which are defined respectively by:

$$FAR = \frac{FP}{TN + FP} \tag{1}$$

$$FRR = \frac{FN}{TP + FN} \tag{2}$$

Thus, it is possible to conclude that smaller threshold values will favour smaller FAR values while larger threshold values will favour smaller FRR values, meaning that a compromise between these two measures should be obtained. The FAR and FRR plots (known as ROC curves) give another important quantity, the equal error rate (EER), defined as the value where FAR and FRR are equal, which can be interpreted as the compromise point between them.

In order to evaluate a system when working in identification mode, two different scenarios can be evaluated: with or without threshold application. In the scope of the present work, performance regarding identification made use of a threshold set. When trying to identify a subject, the smallest distance between the provided data and the templates, d_i is compared to the threshold value, t. Three different situations can occur:

- $d_i < t$ and d_i is taken from a template belonging to the subject. Thus a t_{ID} (true identification) is generated.
- $d_i < t$ and d_i is taken from a template not belonging to the subject. Thus a f_{ID} (false identification) is generated.
- $d_i > t$, meaning the system does not attribute an identification to the subject, thus being rejected, r_{ID}.

In this way, it is possible to compute two different metrics in order to evaluate the identification process: the error probability, Pe_{ID} and the rejection rate, RR_{ID}, which are defined respectively by:

$$Pe_{ID} = \frac{F_{ID}}{R_{ID} + T_{ID} + F_{ID}} \quad (3)$$

$$RR_{ID} = \frac{R_{ID}}{R_{ID} + T_{ID} + F_{ID}} \quad (4)$$

where F_{ID} is the total number of incorrect identifications and T_{ID} is the total number of correct identifications and R_{ID} is the total number rejections. In the same way as for authentication, there is also a compromise between Pe_{ID} and RR_{ID}, the operating point such that $Pe_{ID} = RR_{ID}$, hereafter denoted as PC_{ID}.

2.3. Electrocardiogram

The ECG is the recording of the electrical activity of the heart. Through ECG it is possible to observe the phenomena related with the polarization/depolarization of the cardiac cells. These phenomena are represented in ECG by waves/complexes, namely: **P wave**, which is caused by the atria depolarization; **QRS complex**, which is caused by the ventricle depolarization; **T wave**, which is caused by the ventricular repolarization [4].

The measurement of heart's electrical activity is possible by obtaining the voltage between two different points where electrodes are placed, usually known as a *lead*.

The use of ECG in biometrics is based in three assumptions [5], covering the criteria mentioned above (see section 2.2):

- ECG is difficult to counterfeit in supervised conditions.
- It is present in all living individuals, during their entire life (*universality*).
- Easiness to collect (*collectability*).

2.4. Cancelable Biometrics

The concept of cancelable biometrics was first introduced by Ratha et al. [6], as a way of applying intentional and repeatable distortions/transformations to a biometric signal to ensure the privacy of biometric data. Ideally, these distortions/transformations are non-invertible. Therefore, one can store the transformed version of a biometric characteristic and hence providing higher privacy. According to [7], there are two main ways to generate cancelable templates: through a **non-invertible transformation** or through **biometric salting**.

When using a non-invertible transformation, all the recognition is performed not in the original space but in the space generated by that transformation. Biometric salting usually uses invertible transforms. Since the original information can be recovered by applying the inverse transformation, one must keep in secrecy the transform parameters. This can be done by giving to a certain user a key or password with the needed information.

Cancelable biometrics templates are designed according to two major criteria [7, 2]:

1. *Irreversibility* - Meaning it should be easy to generate a cancelable template, through a **transformation or distortion**, but should be computationally hard or impossible to reconstruct the original data, hence protecting it.
2. *Unlinkability* - Meaning that the system has the ability to generate different cancelable templates from the original data, ensuring **renewability** in case of an attack. Additionally, by crossing several cancelable templates originated by the same data, one cannot have information about that data.

The design of this kind of recognition systems is very similar to the one described above. Additionally, there is a module which is responsible by applying a non-invertible transformation or to perform salting.

2.5. Biometric Menagerie

Since it is important to guarantee that the subjects belonging to a given database have a consistent performance under the recognition process, biometric menagerie aims to observe the influence of each subject in that process by dividing them into several subgroups according to its behaviour under the system.

These groups are defined in terms of a relationship between genuine and impostor match scores. In order to define these scores, a few measures can be computed. The present work uses a simple approach, with the system working on identification mode. On the one hand, the genuine performance is based on the count of how many user's sample tests are indeed from an authorized user, meaning the samples which are correctly classified. On the other hand, the impostor performance is based on the number of sample tests regarding an user that system considers to be forged or unauthorized, meaning the samples which are incorrectly classified. The different animal groups are [8]:

1. **Sheep** - Theoretically, most of population belongs to this group, which is characterized by middle values of impostor and genuine scores. Thus, these subjects will tend to match well against themselves but poorly against others.
2. **Worms** - These individuals are characterized by an high impostor score and a low genuine score, meaning they will tend to match against other subjects easily and perform badly when matched against themselves. Thus, they will contribute to decrease the classification quality.
3. **Phantoms** - This group has both low genuine and impostor scores. Thus, it is hard to match them against themselves.
4. **Chameleons** - This group is characterized by both high genuine and impostor scores, meaning that they easily match against other subjects, besides the good behaviour when they match against themselves.
5. **Doves** - In an ideal scenario, all the subjects should belong to this animal group since it presents low impostor scores but high genuine scores. Therefore, these subjects perform very well when matched

against themselves and it is hard to match them against other subjects.

2.6. Hermite Polynomials

The use of hermite functions to represent heartbeats emerged as a way to provide a very compact representation of ECG.

According to [9], hermite polynomials will provide a better characterization of the beat if the point of maximum symmetry is selected as the center of the window of signal to be fitted. In a normal heartbeat this point is the peak of QRS complex, which corresponds to the R wave. Thus, hermite polynomials will approximate the QRS complex from a heartbeat.

A QRS complex, $x(t)$, which is approximated by the hermite polynomials, can be described as:

$$x(t) = \sum_{n=0}^{N-1} c_n(\sigma)\phi_n(t,\sigma) + e(t) \tag{5}$$

where N is the number of hermite polynomials used to do the approximation, $\phi_n(t,\sigma)$ is the n^{th} hermite function, $c_n(\sigma)$ are the coefficients or weights of the linear combination associated with each function, σ is the parameter which controls the width of the functions and $e(t)$ is the error associated with the approximation.
The n^{th} hermite function is defined as:

$$\phi_n(t,\sigma) = \frac{1}{\sqrt{\sigma 2^n n! \sqrt{\pi}}} e^{-t^2/2\sigma^2} H_n(t/\sigma) \tag{6}$$

$H_n(t/\sigma)$ is the n^{th} hermite polynomial. Hermite polynomials are recursively given by:

$$H_0(x) = 1 \tag{7}$$

$$H_1(x) = 2x \tag{8}$$

$$H_n(x) = 2xH_{n-1}(x) - 2(n-1)H_{n-2}(x) \tag{9}$$

Determine the coefficients $c_n(\sigma)$ is a problem of minimizing the summed square error

$$\sum_t |e(t,\sigma)|^2 = \sum_t |x_t - \sum_n c_n(\sigma)\phi_n(t,\sigma)|^2 \tag{10}$$

3. Dissimilarity Based Representation and Classifier

In order to make a decision about a given subject, all the biometric systems use a classifier.

Typically, classifiers use as base data representation a feature space. In this work, inspired by [10, 11], it is proposed the use of a dissimilarity based representation, to which the classifiers are built upon.

3.1. Dissimilarities

Before describing the system, it is important to define the dissimilarity measure concept and some of its properties. According to [10], a dissimilarity measure $d(o_i, o_j)$ between two objects o_i and o_j can be seen as the degree of difference between them and has several properties, such as:

- *Non-negativity*: $d(o_i, o_j) \geq 0$;
- *Identity of indiscernibles*: $d(o_i, o_j) = 0$ if and only if $o_i \equiv o_j$;
- *Symmetry*: $d(o_i, o_j) = d(o_j, o_i)$.

Thus, one way to extract dissimilarities is by using a similarity measure. In the scope of the present work, the measures that will be used are the **euclidean distance** and the **cosine similarity**. This choice is based on two main facts: on the one hand, the ECG shape suggests that the cosine similarity is a good shape similarity measure to apply at the extraction moment; on the other hand, one of the most used measures in k-NN classifiers is the euclidean distance. If two objects are represented by two vectors o_i and o_j respectively, the euclidean distance and cosine similarity can be defined respectively as

$$D(o_i, o_j) = \sqrt{(o_i - o_j)^T (o_i - o_j)} \tag{11}$$

$$D(o_i, o_j) = \frac{o_i . o_j}{\|o_i\| \, \|o_j\|} \tag{12}$$

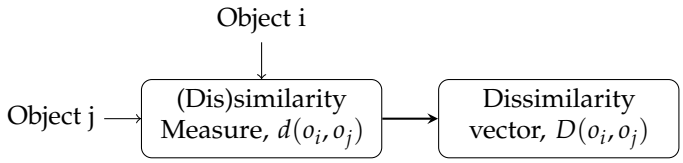

Figure 4: Dissimilarity vector generation.

3.2. Dissimilarity Representation

The approach followed to extract dissimilarities is explained in [11] and it is named **inter-subject approach**, meaning that dissimilarity based representation assumes the existence of n representative objects (in the limit, all the training data), which can be called *prototypes*. Prototypes are randomly selected from the training data, assuming that they can cover all the population variation. Thus, the representation of a single object is given by the vector of dissimilarity of this object to each of the prototypes. Finally, it is possible to define the dissimilarity space, D_S resulting from all the vectors of dissimilarity of all the objects to each of the prototypes.

In the present work, the objects are considered to be the ECG heartbeats, in the case of applying the system solely to records, and the hermite polynomial approximations of the QRS complexes.

4. State of the Art

4.1. Biometric Systems

An architecture for a biometric system using ECG follows, generally speaking, a typical architecture for a pattern recognition system. The majority of the ECG biometric studies are based on one-channel ECG [12]. These recordings are, often, contaminated with noise, both as in high as in low frequencies. To remove the noise, it is applied a band pass filter, usually, a finite impulse response (FIR) filter with cut-off frequencies between 5Hz and 20Hz [13]. Besides de-noising, the pre-processing stage includes, often, ECG segmentation into single heartbeats. ECG biometric

methods can be divided in three classes: fiducial, non-fiducial and hybrid or partially fiducial methods. Fiducial methods, like the one proposed by Coutinho et.al [14], are based on extracting features from characteristic points within an heartbeat, such as wave amplitudes, slope information or time between two waves. On the other hand, non-fiducial methods, such as the work of Coutinho et.al [14], do not use characteristic points to extract features. Instead, feature extraction is done on another space, for example in the frequency domain. Finally, hybrid methods combine both fiducial and non-fiducial methods, like in the work of Wang et.al [15].

As a way to provide a concrete example of a biometric system using ECG, the work of Carreiras et.al [13] is referenced here. Their purpose was to study the *uniqueness* question regarding ECG. The proposed biometric system follows the typical architecture described on section 2.1, using a k-NN classifier, with $k = 3$ and the cosine similarity as classification metric. Their approach was applied to the ECG's heartbeats and all the amplitudes were used as features. For each subject, three heartbeats were used as templates and one was used to perform recognition.

4.2. Cancelable Biometrics

In spite of being a recent topic, there are already some interesting studies in this field of study. These approaches are focused mainly in the fingerprint, iris and handwriting/signature.

Nagar et.al [16] proposed a method to generate non-invertible fingerprint templates and a new measure of non-invertibility which they called coverage-effort, able to compute a degree of how a template is non-invertible. In order to generate cancelable templates they followed the strategy proposed by Ratha et.al [17] based on an one-way transformation in the feature domain. As in [17], they applied three different transformations: cartesian transformation, polar transformation and surface folding transformation. All of them were applied to the minutiae which are the points that best give a distinctive representation of fingerprint since they are the points where the friction ridges end or bifurcate [16]. By computing the coverage-effort parameter, it is possible to chose between the several non-invertible templates, those which present the lowest risk security and the better matching performance.

The application of cancelable biometrics to ECG is residual. One of the already proposed methods was developed by Dey et.al [18]. Their approach was based in biohashing. By using a modified version of the Pan-Tompkins algorithm, they extracted several time features (R-R, S-S, Q-Q, T-T, P-R, Q-T, Q-Tc and QRS complex). After, they were able to compute the inner product between the extracted features and a previously randomly generated tokenized number. Finally, by applying a threshold, they digitized the inner product result. The resulting code, named biohash code, can be used to perform authentication of the subjects. According to them, this method is less susceptible to noise and more resistant to intra-class similarity, however the database used was small and it would be useful to apply this methodology to a larger set of records in order to prove its performance. Besides that, they do not present results regarding the authentication performance.

5. Proposed System

The main purpose of the present work is to develop a new cancelable biometric method by using ECG as biometric feature. On the one hand, hermite polynomial approximation was applied to heartbeats, modelling the QRS complexes. On the other hand, a dissimilarity based representation was applied to the heartbeats' hermite approximations and also to the original records.

The main system's blocks are explained below (see figure 5):

1. The system receives an ECG regarding lead I. The database used to test the developed system contains already filtered signals. Signals were filtered using a FIR (order 150) and with cut-off frequencies of 5Hz and 20Hz. More information regarding the database used can be found in section 6.

2. Filtered signals are then divided into heartbeats by a process of segmentation based on the work of Hamilton [19].

3. Hermite polynomial approximation is applied to the QRS complex from each heartbeat.

4. The previous result is then used to feed a dissimilarity based classifier. The classifier will be tested in both situations: authentication and identification.

Note that the steps described above concern the case when dissimilarity based classifier uses as input the hermite polynomial approximations. When classifier uses as input the original heartbeats, step 3 is omitted. The remaining of the process is the same.

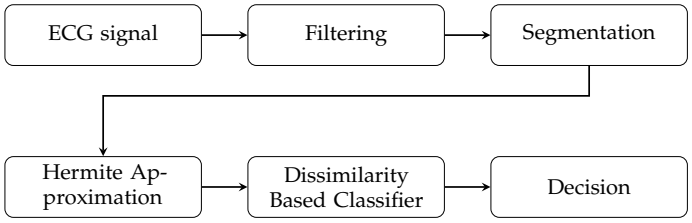

Figure 5: Block diagram regarding the proposed methodology, when classifier uses as input the hermite polynomial approximations.

5.1. Hermite Polynomials

This approximation is done by minimizing equation 10. Thus, it is not possible the recovery of the original QRS complex from its approximation, meaning that this process guarantees *irreversibility*. Additionally, since one can choose the number of polynomials to perform the approximation and the minimization can converge to a different minimum of eq. 10, this process ensures *unlikability*. Thus, in the event of the system being compromised, it is always possible to generate different approximations.

In the present work, a 200 ms window was centred around each R peak. The window chosen is wide enough to cover all the QRS complex, but narrow enough to not include P and T waves [9]. Since all the hermite functions converges to zero in $\pm\infty$, it was added 100ms on each side of the QRS window. Thus, the total window's length is 400ms.

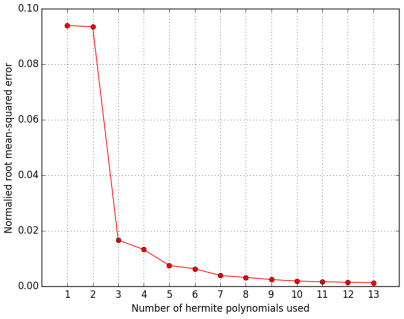

Figure 6: Normalized root mean-squared error

In order to observe the information loss in the hermite polynomial approximation, an error measure was introduced: the normalized root-mean square error (NRMSE). This error measure is defined as follows:

$$NRMSE = \frac{RMSE}{x_{max} - x_{min}} = \frac{\frac{\sqrt{\sum_t |e(t)|^2}}{N}}{x_{max} - x_{min}} \quad (13)$$

where N is the size of window in samples, the denominator represents the signals range of values and $|e(t)|$ is the difference between the original signal and its approximation. This quantity can be seen as the average error expressed as a percentage of the range of values in the signal fragment $(x_{max} - x_{min})$ [9].

As shown in Figure 6, NRMSE is higher if the approximation uses just one or two hermite polynomials and there is no significant improvement if the algorithm uses eight or more polynomials. Besides that, the computational time required to perform approximations grows proportionally with the number of polynomials used. Thus, in the present work, approximations from 4 to 7 polynomials are used across several experiments.

5.2. Dissimilarity Based Representation and Classification

A dissimilarity based representation computes the dissimilarity between an object and a set of prototypes, resulting in a vector which is used by a classifier to perform a decision. In the present work, the objects as input to the classifier are, on the one hand, the ECG heartbeats, and on the other hand, its approximations resulting of the QRS's hermite polynomial approximation.

5.3. Notation

To simplify the understanding of the classifier, some notation that will be used in the rest of this work is presented:

- A population of S subjects;
- A testing population denoted by S_{tst};
- A training population denoted by S_{tr};
- A percentage per representing a fraction of the population which will be denoted as population S_{per}, randomly chosen from S_{tr};
- h_i representing a heartbeat associated to the subject i with $i = 1, ..., S$;
- H_i representing the total number of heartbeats associated to each subject i with $i = 1, ..., S$;
- a_i^p is a hermite polynomial approximation associated to the subject i, with $i = 1, ..., S$ and p represents the number of polynomials used with $p \in [3, 7]$;
- A_i^p is the total number of hermite polynomial approximations from a subject i, with $i = 1, ..., S$ and p represents the number of polynomials used with $p \in [3, 7]$;
- The generated dissimilarity space, D_S either using H_i or A_i^p.

5.4. Dissimilarity Representation

The approach followed to extract dissimilarities is described in section 3.2. In contrast to the work developed by [11], the system will be applied only to one lead.

The dissimilarity space is built based on a randomly chosen population, S_{per}. S_{per} is composed by a certain number of samples, representative of the population, called *prototypes*. For the tests described on section 6, $per \in [10, 90]\%$ in 10% intervals.

Two different dissimilarity representations are built regarding the proposed system: one at the time of the enrolment and another regarding the moment of classification.

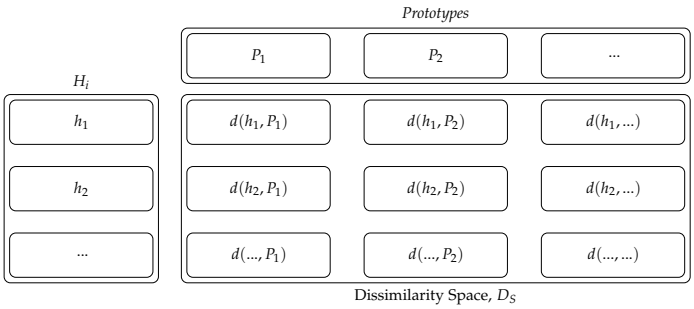

Figure 7: Dissimilarity space generation. In this example, the system uses as input the original records. If one considers H_i as being the templates regarding the i^{th} subject, thus D_S represents the dissimilarity space which is built at the enrolment phase.

5.5. Classifier Description

In order to better understand the proposed system, its modes are explained next:

- **Enrolment** (see figure 8)
  1. Following a simple random sampling approach, a S_{tr} is built by choosing five samples regarding each subject, for all the subjects. Hereafter the samples belonging to S_{tr} will be called *templates*;
  2. Based on S_{tr}, a percentage per of subjects is randomly chosen leading to a sub population S_{per}. As referred above, the samples composing S_{per} are named *prototypes*;

3. For all H_i or A_i^p from S_{tr}, a dissimilarity representation is built. In other words, a dissimilarity representation of each template is represented by a dissimilarity vector of that template to each of the prototypes. Thus, a D_S is obtained (see figure 7).
4. D_S and H_i or A_i^p from S_p are stored.

- **Authentication/Identification** (see figure 9)

1. The samples not belonging to S_{tr} are used as test samples, meaning they define the S_{tst};
2. For all H_i or A_i from D_{tst}, a dissimilarity representation is built. This step is similar to step 2 of the enrolment phase. Another D_S is built.
3. A decision is made, taking into account the evaluation performed by a k-NN classifier by comparing the two generated dissimilarity representations from S_{tr} and S_{tst}.

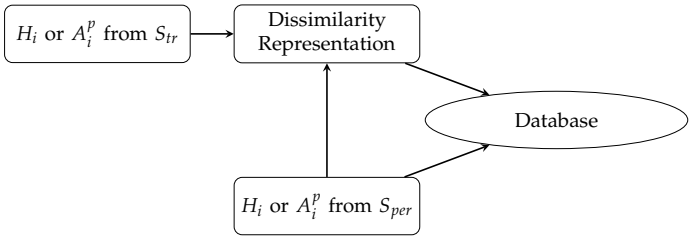

Figure 8: Enrolment regarding the dissimilarity based classifier.

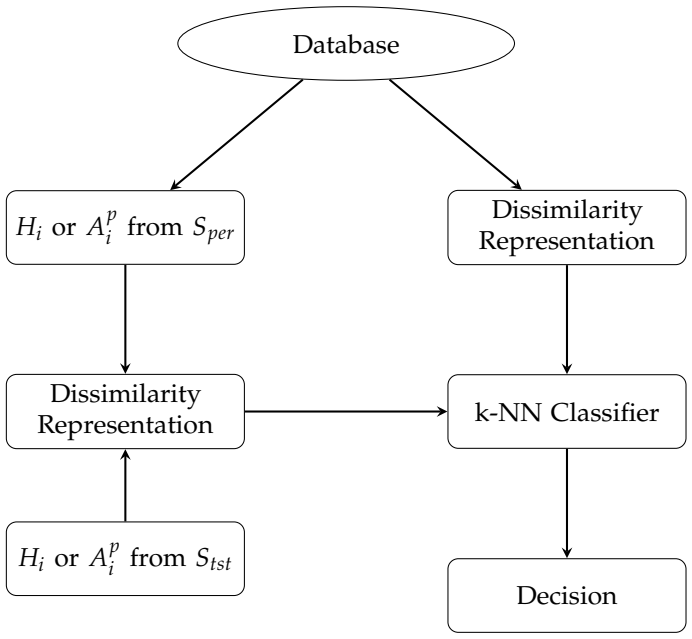

Figure 9: Recognition process regarding the dissimilarity based classifier.

Since it is possible to set *per* (note that this is a random process between the heartbeats or approximations of a given subject and dataset), in case of the system becoming compromised, it is possible to generate different templates, ensuring *renewability*. Besides that, the system is working on a dissimilarity space, meaning it is not possible to get back to the heartbeat/approximation space since the same dissimilarity representation can be obtained by several different pairs of objects, this process also guarantees *irreversibility*.

The proposed system uses two measures: one serves as dissimilarity measure to compute the dissimilarity representation while the another serves as classification metric. As seen above those measures are the euclidean distance and the cosine similarity. Those two measures allow the possibility of different performance tests:

- Cosine similarity used as both dissimilarity measure and classification metric, labelled as **CC**.
- Cosine similarity used as dissimilarity measure and euclidean distance used as classification metric, labelled as **CE**.
- Euclidean distance used as both dissimilarity measure and classification metric, labelled as **EE**.
- Euclidean distance used as dissimilarity measure and cosine similarity used as classification metric, labelled as **EC**.

In order to evaluate the system's performance, two different measures were calculated. One one hand, regarding authentication, EER was computed taking into account both FAR and FRR regarding each test. One the another hand, PC_{ID} was computed based on Pe_{ID} and RR_{ID}, in order to evaluate the system's performance in the identification mode. Both authentication and identification measures were computed based on a predefined set of thresholds, and all the values showed in section 6 are with respect to the performance regarding all the subjects of the used database.

6. Results

In order to test the proposed system, a series of tests were ran. Those tests aimed to study the influence of some parameters. Regarding the hermite polynomial approximations, the influence of the number of polynomials used to model the QRS complex. Regarding the dissimilarity based representation, the influence of the use of different dissimilarity measures as well as the influence of the *per* parameter. Besides that, the influence of the number of neighbours, k and the classification metric used by the k-NN classifier was also studied. Since those tests leaded to a great amount of information, only the more important results are showed.

The *Hospital de Santa Marta* database was employed to test the proposed methodology. The records belonging to this database were collected during scheduled appointments, emergency cases and bedridden patients. The signals were acquired using Philips PageWriter Trim III devices, with a sampling rate of 500Hz and 16bit resolution. Each record is composed by the records regarding the 12 leads and have a duration of about 10s. Since only the records belonging to healthy individuals (in respect to heart diseases) were considered, the following tests were applied to 832 records belonging to 612 subjects.

The database used has for several subjects different record sessions, thus, it was considered to be interesting to test the proposed methodology with four different approaches in order to observe the intra-variability degree

between different sessions belonging to the same subject, thus evaluating system's performance under different conditions. The four used approaches were:

- **Approach 1** - Records were grouped by the subjects they belong to. Thus, the number of tested subjects was 612. No distinction between sessions was taken into account at the moments of training or testing.
- **Approach 2** - Each of the records were considered to be a different subject, meaning that the total number of tested "subjects" was 832.
- **Approach 3** - Records were grouped by the subjects they belong to but only one record session was used at the training phase, while the remaining were used at the test phase.
- **Approach 4** - Records were grouped by the subjects they belong to but in contrast to approach 3, training was performed with several sessions and only one session was used to evaluate the system's performance.

All of performed tests present some properties in common, namely, all of them used 5 randomly chosen heartbeats or hermite polynomial approximations regarding each subject in order to produce the training set, S_{tr}. The remaining data related with a given individual was used to built the test set, S_{tst}. At the recognition phase, each heartbeat or approximation was tested individually. All of the tests presented below were ran 5 times each. For the sake of simplicity just EER and PC_{ID} average values were plotted (see figures from 10 to 15).

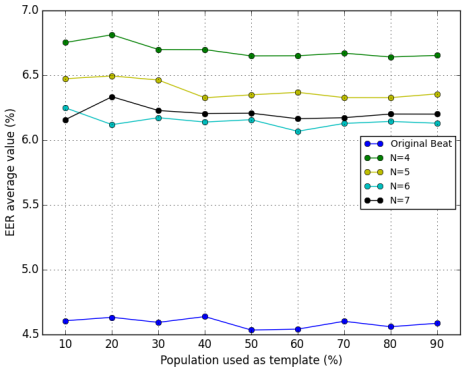

Figure 10: **EER** average values regarding **EC** combination, using **approach 1** and k=1.

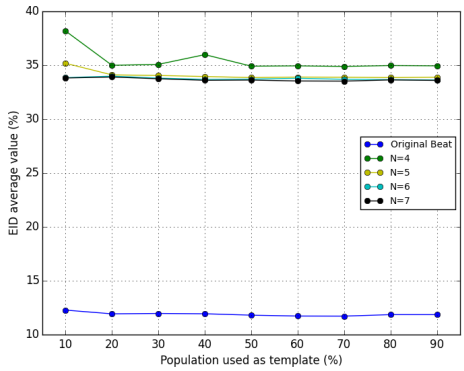

Figure 11: PC_{ID} average values regarding **EC** combination, using **approach 1** and k=1.

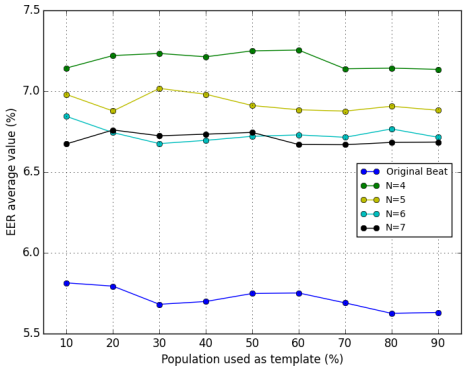

Figure 12: **EER** average values regarding **EE** combination, using **approach 1** and$k$=1.

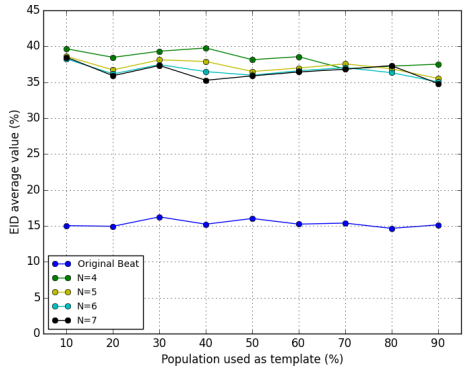

Figure 13: PC_{ID} average values regarding **EE** combination, using **approach 1** and k=1.

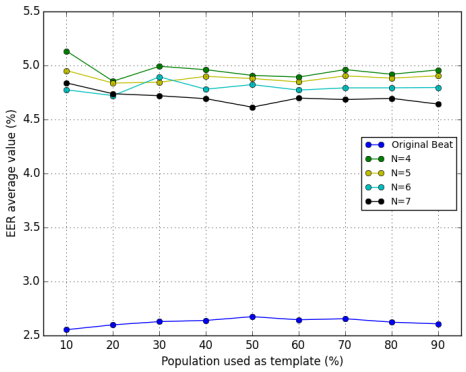

Figure 14: **EER** average values regarding **approach 2** and k=1, using **EC** combination.

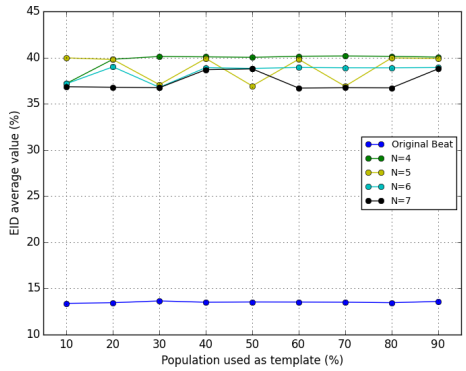

Figure 15: PC_{ID} average values regarding **approach 2** and k=1, using **EC** combination.

The work of Carreiras et.al [13], described on section 4.1, was used in order to obtain a comparison regarding the proposed methodology.

Not only the heartbeats were used to train the classifier, but also its hermite polynomial approximations as input to the classifier. A simple approach was done. The records were grouped by subject, similar to **approach 1**. For each subject, four heartbeats were randomly chosen, three of them were used as templates and the remaining one was used to test the classifier, according to [13]. This process was repeated 10 times for each classifier. In order to evaluate the system's performance, EER and PC_{ID} were obtained. Table 1 presents the average and standard deviation values related with this system.

Table 1: EER and PC_{ID} values regarding the system proposed by Carreiras et al. [13].

	EER (%)	PC_{ID}**(%)**
Original beat	9.83 ± 0.30	10.90 ± 1.47
N = 4	13.14 ± 0.32	42.39 ± 0.40
N = 5	13.43 ± 0.51	41.19 ± 0.48
N = 6	12.16 ± 0.38	40.05 ± 0.28
N = 7	12.53 ± 0.38	39.78 ± 0.34

6.1. Biometric Menagerie

The following steps were made, in order to apply this technique:

1. The technique was directly applied to the results regarding the system proposed by Carreiras et.al [13].

2. The system developed by Carreiras et.al [13] and the proposed method were then applied to each species group.

This technique was applied following **approach 1**, and k set to 1, meaning 612 subjects were taken into consideration.

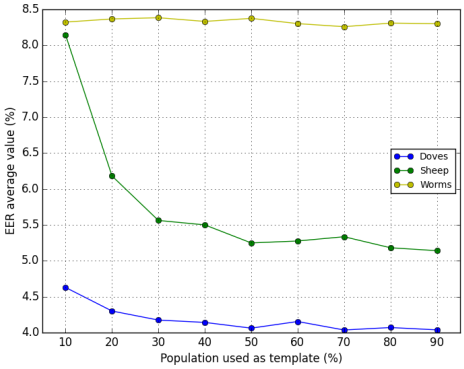

Figure 16: An example of the obtained results through the biometric menagerie, in this case, the **EER** average values regarding the renewed application of the proposed system, when **four polynomials** are used to model the QRS complex.

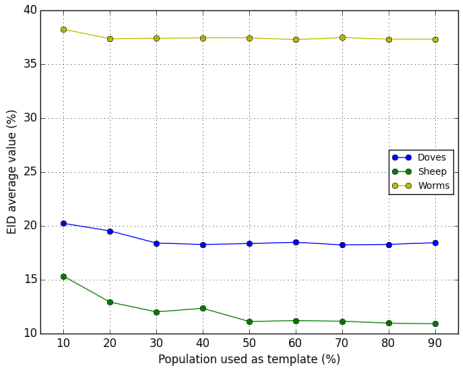

Figure 17: An example of the obtained results through the biometric menagerie, in this case, the PC_{ID} average values regarding the renewed application of the proposed system, when **four polynomials** are used to model the QRS complex.

7. Conclusions

Nowadays, security and robustness are two of the main key factors regarding the application of biometrics systems in real life situations. Therefore it is crucial to develop systems able to convey confidence to the users, maintaining a good recognition performance. The present work explored two main fields. On the one hand, hermite polynomial approximations modelling QRS complex and one the other hand, the dissimilarity based representation. These two fields are able, in separated or set together, to ensure a cancelable biometric system.

Regarding hermite polynomial approximations, used to modelling the QRS complex, it was possible to conclude that using it as input to the proposed system conducted to a slightly decrease in the recognition performance, when compared with the system working with the original records. Besides that, the influence of the number of used polynomials was studied. It was not completely clear that using a higher number of polynomials conducted to a better system's performance.

Regarding dissimilarity based representation and classifier, besides the influence of the dissimilarity measure and of the classification metric, the influence of the number of neighbours, k, used by the k-NN classifier and the percentage of population used as templates, *per*, was studied. Besides studying the influence of the parameters referred above, also the way of how database records were used was studied.

It was possible to conclude that the proposed method, globally performed better than the other mentioned biometric systems when using the original records as well as when using the hermite polynomial approximations. Note that all the tests were performed using the same dataset. It is a very important result, since besides performing well than a traditional biometric system like the one proposed by [13], it ensures the protection of the biometric data, which is a crucial point regarding the security needed in the present recognition systems, while maintaining or even improving the recognition performance.

Table 2: Percentage average and standard deviation values of HSM database assigned to each species.

	Doves	**Sheep**	**Worms**	**Chameleons**	**Phantoms**
Original beat	73.40 ± 1.05	0	18.46 ± 1.09	0	8.14 ± 0.60
N=4	26.94 ± 1.81	12.44 ± 2.52	60.61 ± 1.28	0	0
N=5	26.96 ± 0.91	18.14 ± 0.94	54.90 ± 0.90	0	0
N=6	28.82 ± 1.49	19.82 ± 2.85	51.36 ± 1.82	0	0
N=7	29.01 ± 1.27	19.85 ± 1.86	51.13 ± 1.27	0	0

References

[1] A. K. Jain and A. Kumar, "Biometrics of next generation: An overview," *Second Generation Biometrics*, vol. 12, no. 1, pp. 2–3, 2010.

[2] A. B. Teoh, Y. W. Kuan, and S. Lee, "Cancellable biometrics and annotations on biohash," *Pattern recognition*, vol. 41, no. 6, pp. 2034–2044, 2008.

[3] A. Jain, R. Bolle, and S. Pankanti, *Biometrics: personal identification in networked society*, ch. Introduction to Biometrics, pp. 1–41. Springer Science & Business Media, 2006.

[4] J. E. Hall, *Guyton and Hall textbook of medical physiology*. Elsevier Health Sciences, 2010.

[5] I. Odinaka, P.-H. Lai, A. D. Kaplan, J. O'Sullivan, E. J. Sirevaag, S. D. Kristjansson, A. K. Sheffield, J. W. Rohrbaugh, *et al.*, "Ecg biometrics: A robust short-time frequency analysis," in *Information Forensics and Security (WIFS), 2010 IEEE International Workshop on*, pp. 1–6, IEEE, 2010.

[6] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM systems Journal*, vol. 40, no. 3, pp. 614–634, 2001.

[7] C. Rathgeb and A. Uhl, "A survey on biometric cryptosystems and cancelable biometrics," *EURASIP Journal on Information Security*, vol. 2011, no. 1, pp. 1–25, 2011.

[8] N. Yager and T. Dunstone, "The biometric menagerie," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 2, pp. 220–230, 2010.

[9] D. G. Márquez, A. Otero, P. Félix, and C. a. García, "On the Accuracy of Representing Heartbeats with Hermite Basis Functions," *Biosignals*, pp. 338–341, 2013.

[10] R. P. Duin and E. Pekalska, "The dissimilarity representation for structural pattern recognition," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pp. 1–24, Springer, 2011.

[11] M. Francisco, C. Carlos, Lourenço, F. Ana, and F. Rui, "ECG Biometrics Using a Dissimilarity Space Representation," 2015.

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

[13] C. Carreiras, A. Lourenco, A. Fred, and R. Ferreira, "Ecg signals for biometric applications-are we there yet?," in *Informatics in Control, Automation and Robotics (ICINCO), 2014 11th International Conference on*, vol. 2, pp. 765–772, IEEE, 2014.

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

[15] Y. Wang, F. Agrafioti, D. Hatzinakos, and K. N. Plataniotis, "Analysis of human electrocardiogram for biometric recognition," *EURASIP journal on Advances in Signal Processing*, vol. 2008, p. 19, 2008.

[16] A. Nagar and A. K. Jain, "On the security of non-invertible fingerprint template transforms," in *Information Forensics and Security, 2009. WIFS 2009. First IEEE International Workshop on*, pp. 81–85, IEEE, 2009.

[17] N. K. Ratha, S. Chikkerur, J. H. Connell, and R. M. Bolle, "Generating cancelable fingerprint templates," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 4, pp. 561–572, 2007.

[18] N. Dey, B. Nandi, M. Dey, D. Biswas, A. Das, and S. S. Chaudhuri, "Biohash code generation from electrocardiogram features," in *Advance Computing Conference (IACC), 2013 IEEE 3rd International*, pp. 732–735, IEEE, 2013.

[19] P. Hamilton, "Open source ecg analysis," in *Computers in Cardiology, 2002*, pp. 101–104, IEEE, 2002.