Generic Biometry Algorithm Based on Signal Morphology Information: Application in the Electrocardiogram Signal

Tiago Araújo, Neuza Nunes, Hugo Gamboa and Ana Fred

Abstract This work presents the development, test, and implementation of a new biometric identification procedure based on electrocardiogram (ECG) signal morphology. ECG data were collected from 63 subjects during two data-recording sessions separated by six months (Time Instance 1, T1, and Time Instance 2, T2). Two tests were performed aiming at subject identification, using a distance-based method with the heartbeat patterns. In both tests, the enrollment template was composed by the averaging of all the T1 waves for each subject. Two testing datasets were created with five meanwaves per subject. While in the first test the meanwaves were composed with different T1 waves, in the second test T2 waves were used. The T2 waves belonged to the same subjects but were acquired in different time instances, simulating a real biometric identification problem. The classification was performed through the implementation of a kNN classifier, using the meanwave's Euclidean distances as the features for subject identification. The accuracy achieved was 95.2 % for the first test and 90.5 % for the second. These results were achieved with the optimization of some crucial parameters. In this work we determine the influence of those parameters, such as, the removal of signal outliers and the number of waves that compose the test meanwaves, in the overall algorithm performance. In a real time identification problem, this last parameter is related with the length of ECG signal needed to perform an accurate decision. Concerning the study here depicted, we

T. Araújo (✉) · N. Nunes · H. Gamboa · A. Fred
CEFITEC, New University of Lisbon, Caparica, Portugal
e-mail: taraujo87@gmail.com

T. Araújo · N. Nunes · H. Gamboa · A. Fred
Plux - Wireless Biosignals, Lisbon, Portugal
e-mail: nnunes@plux.info

T. Araújo · N. Nunes · H. Gamboa · A. Fred
Instituto de Telecomunicações, Scientific Area of Networks and Multimedia,
Lisbon, Portugal
e-mail: hgamboa@fct.unl.pt

A. Fred
e-mail: afred@lx.it.pt

A. Fred and M. De Marsico (eds.), *Pattern Recognition Applications and Methods*,
Advances in Intelligent Systems and Computing 318,
DOI 10.1007/978-3-319-12610-4_19

conclude that a distance-based method using the subject's ECG signal morphology is a valid parameter for classification in biometric applications.

Keywords Biometry · Classification · Electrocardiography · Meanwave · Signal processing

1 Introduction

Every day, large amounts of confidential data are stored and transferred through the internet. New concerns about security and authentication are arising; speed and efficiency in intruders detection is crucial. Biometric recognition addresses this problem in a very promising point of view. The human, voice, fingerprint, face, and iris are examples of individual characteristics currently used in biometric recognition systems [1]. Recently, several works have studied the electrocardiography (ECG) signal as an intrinsic subject parameter, exploring its potential as a human identification tool [2–4].

Biometry based in ECG is essentially done by the detection of fiducial points and subsequent feature extraction (Fig. 1) [5]. Nevertheless there are some works that use a classification approach without fiducial points detection [6], referring computational advantages, better identification performance and peak synchronization independence.

Since 2007, Institute of Telecommunications (IT) research group has explored this theme addressing it, essentially, in two ways: (i) analysis of the ECG time persistent information, with possible applicability in biometrics over time; and (ii) development of acquisition methods which enabled the ECG signal acquisition with less obtrusive setups, particularly using hands as signal acquisition point.

Following these goals, a recent work proposed a finger-based ECG biometric system, collecting the signals through a minimally intrusive 1-lead ECG setup at the fingers and recurring to Ag/AgCl electrodes without gel [5]. In the same work, an algorithm was developed for comparison between the R peak amplitude from the

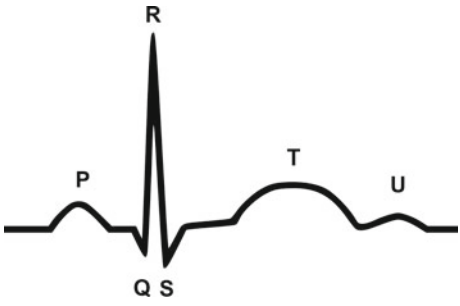

Fig. 1 The ecg fiducial points

heartbeats of test patterns and the R peak from the enrollment template database. The results revealed that this could be a promising technique.

In this work we used the IT ECG database and follow the same methodology as described before, but using a new biometrics classification algorithm based on the heartbeat meanwave's Euclidean distances.

In the following section we will depict the procedure for the ECG data acquisition and pre-processing. We also explain the methodology followed in this study to efficiently classify the heartbeat waves into the respective subject. The results of the classification procedure are exposed and discussed in section three. Conclusions are taken in section four of this paper.

2 Procedure

2.1 Data Collection

ECG data were collected from 63 subjects, 166.55 ± 8.26 cm, 61.82 ± 11.7 Kg and 21 ± 4.46 years old, during two data-recording sessions with six months between them. We divided those acquisitions in two groups, T1 and T2, referring respectively to the first recording instance and the second recording six months after. The subjects were asked to be seated and relaxed in both recordings.

2.2 Signal Acquisition and Conditioning

The signals were acquired by two dried electrodes assembled in a differential configuration [5]. The sensor uses a virtual ground, an input impedance over 1 MΩ, 110 dB of CMRR and gain of 10 in the first stage. The conditioning circuit consists of two filtering levels: (i) bandpass between 0.05 and 1,000 Hz and (ii) notch filter centered in 50 Hz to remove network interference. The final amplification stage has a gain of 100 to improve the resolution of the acquired signal. This system also magnifies the signal after filtering undesired frequencies in each conditioning stage. The signal is then digitalized for further digital processing. This processing consists in: (a) bandpass digital filter (FIR) of 301 order and bandwidth from 5 to 20 Hz, obtained using a hamming window, (b) detection of QRS complexes, (c) segmentation of ECG and determination RR intervals, (d) outliers removal, (e) meanwave computation and feature extraction, and finally (f) the data classification. The signal acquisition and the processing steps (a), (b) and (c) were done by the methodology developed in IT [5].

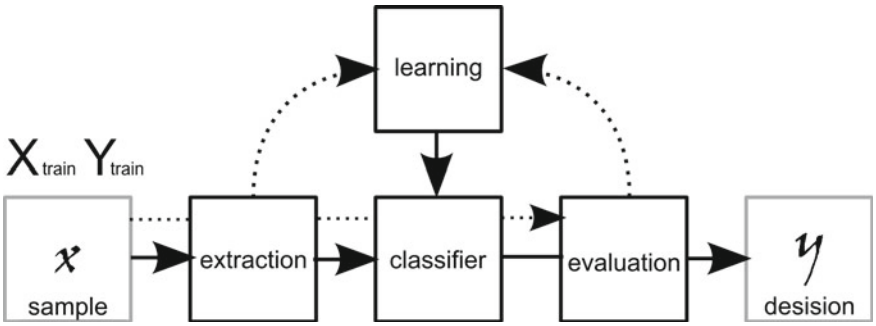

Fig. 2 The process of data classification

In the following section the methodology designed for the implementation of the remaining steps ((d), (e) and (f)) will be described.

2.3 The Process of Data Classification

Data classification is a machine learning technique used to predict group membership for data instances. The main goal of this study was to successfully use the patterns of ECG heartbeats to make subject's identification in different time periods, using a classification method.

Figure 2 depicts the usual process that is followed to classify a set of data.
This process comprises a first stage of feature extraction, making data transformations to generate useful and novel features from a set of candidates. In the data classification there's a supervised learning process.
A first set of data, called training set, is received as input by the classifier, then, with those inputs, it will learn about the features and correspondent classes. The new set of data given, called test set, will match the features with the input training set and associate each sample to the correspondent classes.

2.4 Feature Extraction

The Fig. 3 provides a schematics of the methodology followed in this work.

The data used in this study were divided in two groups: the T1 and T2 acquisitions. In the first test we work with only T1 waves, and in the second test we compare the T2 waves with the T1 template—therefore we can check the differences in classification accuracy when working with acquisitions separated in time from the same subject, simulating a real biometric identification problem.

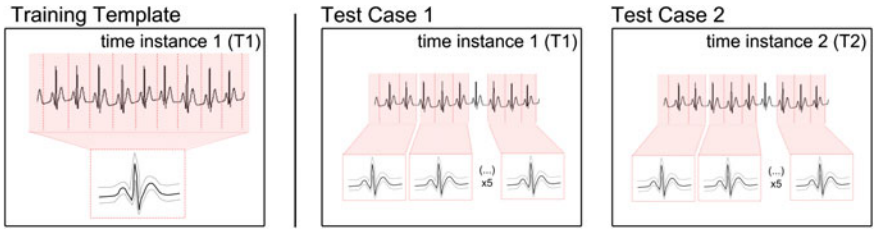

Fig. 3 Template and Tests of the classification process

The dataset defined as template is composed with the T1 subjects' meanwaves, and the features computed for the classification process will be the distance value between the template meanwaves and the meanwaves of future acquisitions (tests).

To compose the template, the first step was to compute a meanwave [7] by the averaging of all T1 waves (which were already segmented into RR-aligned heartbeats). An outliers removal procedure followed, by computing the mean square error distance of each wave to the resulting meanwave. Equation 1 displays the expression for the computation of this distance for only one heartbeat (being l the length, in samples, of the normalized cycle and meanwave). After gathering the distance of each wave to the meanwave, the mean distance value was computed and the waves which presented a distance value higher than two times the mean were removed from the template.

A new meanwave for each subject was then computed without the outliers. Each subject's meanwave was composed with 100 heartbeat waves. This completed the template for the classifier.

$$distance = \sqrt{\frac{\sum_{i=1}^{l}(cycle_i - meanwave_i)^2}{l}} \tag{1}$$

For the first Test dataset, we also used the T1 waves, but divided them randomly into 5 groups, computing one meanwave for each group. Each meanwave was composed with 20 heartbeat waves. Those five test meanwaves were compared, using a distance metric, with the T1 template, for each subject. The distance metric used was the same presented before in Eq. 1, where we used the meanwaves computed from each group instead of each subject's cycle.

For the second Test we followed the same procedure as before but with a calculation of the distance between the T1 template meanwave and the 5 meanwaves from T2 for each subject.

With the distance values computed for both tests we composed two distances' matrices with 63 columns or features, representing the distance of each sample (the Test meanwaves) to each subject's meanwave of the template T1, and 315 (5×63) rows or samples, representing the 5 meanwaves we gathered for each subject and each Test.

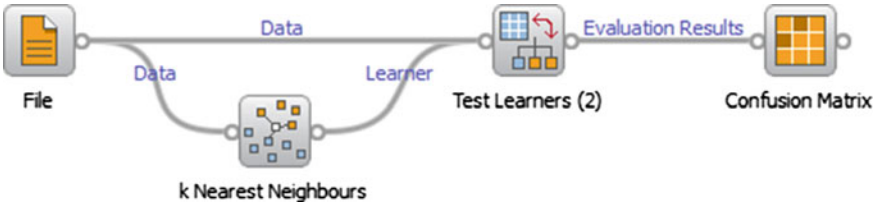

Fig. 4 Schematics used in Orange for classification

2.5 Classifier

To classify the data, a user friendly toolbox [8] was used. As input, it received the distances matrices and used a k-Nearest Neighbor (kNN) classifier with a "leave one out" criterion to learn about the data given. Figure 4 shows the Orange schematics of the data classification and results gathering.

In this image the icons represent the steps of the data classification process: The File icon represents the distance matrices given as input to be classified; The k Nearest Neighbor classifies samples based on the closest class amongst its k nearest neighbors (we used $k = 5$); The test learner represents the stage where the data given is processed by the classification algorithm and the classifier learns about the samples and correspondent classes; The confusion matrix confronts the predictions with the expected results to return the detailed results of the specified classifier.

3 Results and Discussion

3.1 Distance Matrix

Figure 5 shows the distances matrices given as input to the classifier for Test 1 and Test 2 in a color scale image.

The darker colors represent minimum distance values, which are associated to the heartbeat intra-subject distances. For both tests five samples per subject were compared with the meanwave template. Therefore, it is expected to see a diagonal composed with 5 dark cells and all the other cells with lighter colors (in the best case scenario, they would be totally white). As we can see in Fig. 5, the test 1 is closer to the ideal result, as this test comprises waves from the same acquisition both in template and test sets. In the second test the subjects are not so easily visually identified by the distance metric, and therefore it is expected to see a decrease in accuracy for the second test (Table 1).

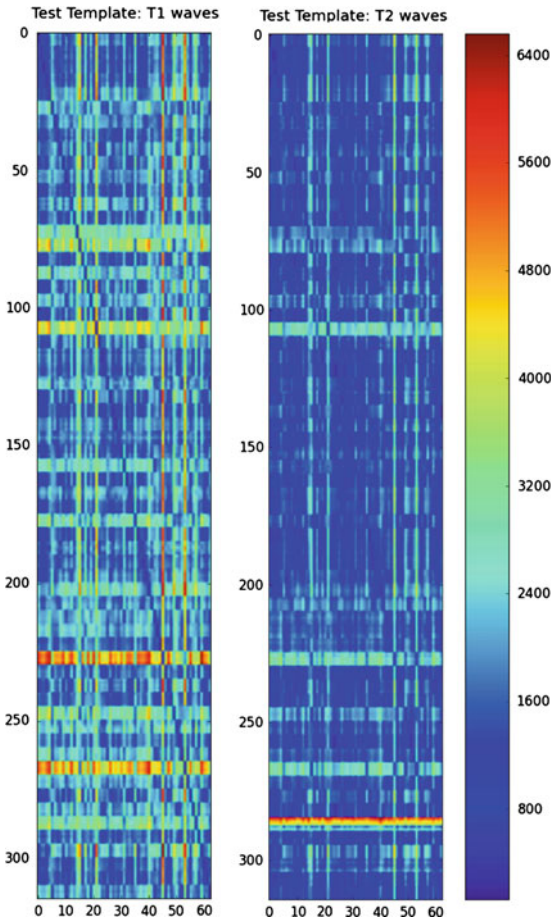

Fig. 5 Distance matrices for Test 1 and Test 2 given as input to the classifier

3.2 Classification Accuracy

After the learning process in Orange, a confusion matrix returned the depicted results of the classifier. An example of that matrix is shown in Table 2.

This matrix gathers the results of the classification for each class (each subject). The ideal case was to have a diagonal always with 5 samples—it represents that all samples were efficiently classified, as we had 5 samples per subject. A cell presenting an inferior value represents that at least one misclassification was made, associating a sample to other class (at least one heatbeat's meanwave was classified as belonging to a different subject).

The final classification results for test 1 and 2, concerning all subjects are included in Table 1.

Table 1 Results for the classification accuracy

	Test 1	Test 2
Accuracy	95.2 %	90.5 %

Table 2 Part of the confusion matrix returned from the classifier

	1	2	3	4	5	6	7	8	9	10	(...)	60	61	62	63
1	**5**	0	0	0	0	0	0	0	0	0	...	0	0	0	0
2	0	**5**	0	0	0	0	0	0	0	0	...	0	0	0	0
3	0	0	**5**	0	0	0	0	0	0	0	...	0	0	0	0
4	0	0	0	**5**	0	0	0	0	0	0	...	0	0	0	0
5	0	0	0	0	**4**	0	0	0	0	**1**	...	0	0	0	0
6	0	0	0	0	0	**5**	0	0	0	0	...	0	0	0	0
7	0	**2**	0	0	0	0	**3**	0	0	0	...	0	0	0	0
8	0	0	0	0	0	0	0	**5**	0	0	...	0	0	0	0
9	0	0	0	0	0	0	0	0	**5**	0	...	0	0	0	0
10	0	0	0	0	0	0	0	0	0	**5**	...	0	0	0	0
(...)	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
60	0	0	0	0	0	0	0	0	0	0	...	**5**	0	0	0
61	0	0	0	0	0	0	0	0	0	0	...	0	**5**	0	0
62	0	0	0	0	0	0	0	0	0	0	...	0	0	**5**	0
63	0	0	0	0	0	0	0	0	0	0	...	0	0	0	**5**

Table 3 Classification accuracy results for Test 1 and Test 2 with and without removal of outliers

	Test 1 (%)	Test 2 (%)
w/ outliers removal	95.2	90.5
w/o outliers removal	88.2	85.4

3.3 Algorithm Parameterization Versus Classification Accuracy

The methodology followed to achieve the depicted results was designed to optimize the classification rate. Before gathering the meanwaves for each subject, an outlier removal algorithm was applied to remove waves which were distant from the template wave. The outliers removal algorithm is relevant to the classification process, as seen in the accuracy rates shown in Table 3. The classification accuracy increases by 2 % and 5 % after removal of the outlier heartbeat waves.

Also stated in the methodology of this work, each of the test sample meanwaves were composed with 20 heartbeat waves from each subject. This was the optimal number of waves to achieve the higher classification rate, as shown in Fig. 6.

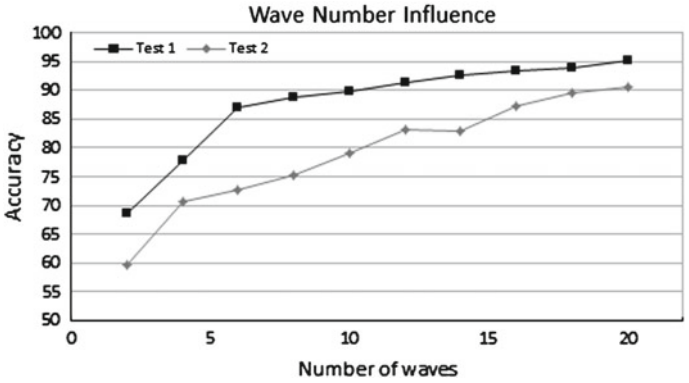

Fig. 6 Influence of the number of waves composing the test sample meanwaves

4 Conclusions

A new biometric classification procedure based on electrocardiogram (ECG) heartbeats meanwave's distances was implemented and depicted in this study. Our goal was to successfully use the patterns of ECG heartbeats to make subject's identification. In order to validate the developed solutions, the methods were tested in a real ECG database. The database was composed by two finger-based ECG acquisitions from 63 subjects. The acquisitions from each subject were separated by six months between them. This fact enabled the evaluation of the algorithm accuracy in a test case scenario, where the test and enrollment template belonged to the first acquisitions, and a real case scenario where we used the first acquisitions as the enrollment template and the second one as test. Using our approach it was possible to obtain accuracy rates of 95.2 % for the test scenario (Test 1) and 90.5 % for the real case scenario (Test 2). Compared with a previous state-of-the-art approach, the results outperform the recent studies on finger-ECG based identifications. Previous works present 89 % [9] and 94.4 % [5] of accuracy.

Future work will be focused on improving the feature extraction process and add features to the classifier, such as the correlation between waves or the intra-subject variability—as we noticed that some subjects had an higher variability in their meanwaves, and therefore the distance computed isn't the best feature per se.

Acknowledgments The authors would like to thank the Escola Superior de Saúde-Cruz Vermelha Portuguesa (ESSCVP) for the data collections infrastructures and subjects providence.

References

1. Jain, A., Hong, L., Pankanti, S.: Biometric identification. Commun. ACM. **42**(2), 90–98 (2000)
2. Silva, H., Gamboa, H., Fred, A.: Applicability of lead v2 ecg measurements in biometrics. In: Proceedings of Med-e-Tel (2007)
3. Coutinho, D. P., Fred, A. L. N., Figueiredo, M. A. T.: Personal identification and authentication based on one-lead ecg using ziv-merhav cross parsing. In: 10th International Workshop on Pattern Recognition in Information Systems (2010)
4. Li, M., Narayanan, S.: Robust ecg biometrics by fusing temporal and cepstral information. In: 20th International Conference on Pattern Recognition (2010)
5. Lourenco, A., Silva, H., Fred, A.: Unveiling the biometric potential of finger-based ecg signals. In: Computational Intelligence and Neuroscience (2011)
6. Plataniotis, K.N., Hatzinakos, D., Lee, J.K.M.: Ecg biometric recognition without fiducial detection. In: Biometric Consortium Conference, Biometrics Symposium (2006)
7. Nunes, N., Araujo, T., Gamboa, H.: Time series clustering algorithm for two-modes cyclic biosignals. In: Fred, A., Filipe, J., Gamboa, H. (eds.) BIOSTEC 2011, CCIS 273, pp. 233–245. Springer, Heidelberg (2012)
8. Orange. http://orange.biolab.si/ (2012)
9. Chan, A.D.C., Hamdy, M.M., Badre, A., Badee, V.: Wavelet distance measure for person identification using electrocardiograms in IEEE Transactions on Instrumentation and Measurement (2008)