Impact of the Acquisition Time on ECG Compression-based Biometric Identification Systems

João M. Carvalho[1], Susana Brás[1], Jacqueline Ferreira[23], Sandra C. Soares[245], and Armando J. Pinho[1]

[1] IEETA - Dept. of Electronics Telecommunications and Informatics, University of Aveiro, Portugal
[2] Dept. of Education and Psychology, University of Aveiro, Portugal
[3] IBILI, Faculty of Medicine, University of Coimbra, Portugal
[4] CINTESIS-UA, University of Aveiro, Portugal
[5] Dept. of Clinical Neurosciences, Karolinska Institute, Stockholm, Sweden
{joao.carvalho|susana.bras|sandra.soares|jacquelineferreira|ap}@ua.pt

Abstract. The ECG signal conveys desirable characteristics for biometric identification (universality, uniqueness, measurability, acceptability and circumvention avoidance). However, based on the current literature review, there are no results that evaluate the number of heartbeats needed for personal identification. This information is undoubtedly useful when building a biometric identification system – any system should ask participants to provide data for identification, using the smallest time interval that is possible, for practical reasons. In this paper, we aim at exploring this topic using a measure of similarity based on the Kolmogorov Complexity, called the Normalized Relative Compression (NRC). To attain the goal, we built finite-context models to represent each individual – a compression-based approach that has been shown successful for several other pattern recognition applications like image similarity, DNA sequences or authorship attribution.

Keywords: Kolmogorov complexity, biometric identification, finite-context models, similarity metrics, compression algorithms, ECG

1 Introduction

1.1 Motivation

Data compression models have been used to address several data mining and machine learning problems, usually by means of a formalization in terms of the information content of a string or of the information distance between strings [5][3] [6] [4]. This approach relies on solid foundations of the concept of algorithmic entropy and, because of its non-computability, approximations provided by data compression algorithms [7].

The ECG is a well-known and studied biomedical signal. To understand pathological characteristics, in clinical practice, it is usual to try to reduce the inter-variability that characterizes the signal. This inter-variability is precisely the source of richness that renders the ECG an interesting signal for biometric applications. Recent work, using Ziv-Merhav cross parsing algorithm [8][9], as well as finite-context models (FCM) [10], has shown that compression-based approaches are suitable to ECG biometric identification. Nonetheless, a good acceptability when identifying a person for biometry should be fulfilled in as short time as possible. For that, we need to evaluate the minimal number of heartbeats that is needed to be collected, which has not yet been explored in the literature.

1.2 Database used

The database used in our experiments, and in previous works, was collected *in house* [10], where 25 participants were exposed to different external stimuli – *disgust*, *fear* and *neutral*. Data were collected on three different days (once per week), at the University of Aveiro, using a different stimulus per day.

The data signals were collected during 25 minutes on each day, giving a total of around 75 minutes of ECG signal per participant. Before being exposed to the stimuli, during the first 4 minutes of each data acquisition, the participants watched a movie with a beach sunset and an acoustic guitar soundtrack, and were instructed to try to relax as much as possible.

The ECG was sampled at 1000 Hz, using the MP100 system and the software AcqKnowledge (Biopac Systems, Inc.). During the preparation phase, the adhesive disposable Ag/AgCL-electrodes were fixed in the right hand, as well as in the right and left foot. We are aware that such an intrusive set-up is not desirable for a real biometric identification system. However, for testing purposes, it seems appropriate, as this approach is more reliable – produces less noise.

1.3 Compression-based measures

Compression-based distances are tightly related to the Kolmogorov notion of complexity, also known as algorithmic entropy. Let x denote a binary string of finite length. Its **Kolmogorov complexity**, $K(x)$, is the length of the shortest binary program x^* that computes x in a universal Turing machine and halts. Therefore, $K(x) = |x^*|$, the length of x^*, represents the minimum number of bits from which x can be computationally retrieved [11].

The **Information Distance** (ID) and its normalized version, the **Normalized Information Distance** (NID), were proposed by Bennett *et al.* almost two decades ago [12] and are defined in terms of the Kolmogorov complexity of the strings involved, as well as the complexity of one when the other is provided.

However, since the Kolmogorov Complexity of a string is not computable, an approximation (upper bound) for it can be used by means of a compressor. Let $C(x)$ be the number of bits used by a compressor to represent the string x. We will use a measure based on the notion of *relative compression* [4], denoted by $C(x||y)$, which represents the compression of x relatively to y. This measure obeys the following rules:

- $C(x||y) \approx 0$ iff string x can be built efficiently from y;
- $C(x||y) \approx |x|$ iff $K(x|y) \approx K(x)$.

Based on this rules, the **Normalized Relative Compression** (NRC) of string x given string y is defined as

$$\mathrm{NRC}(x, y) = \frac{C(x||y)}{|x|}, \tag{1}$$

where $|x|$ denotes the length of x.

2 Method

In ECG biometric identification, the signal should be processed, in order to extract the useful information for similarity evaluation. The workflow for biometric identification is represented in Fig. 1. Each block will be explained through this section.

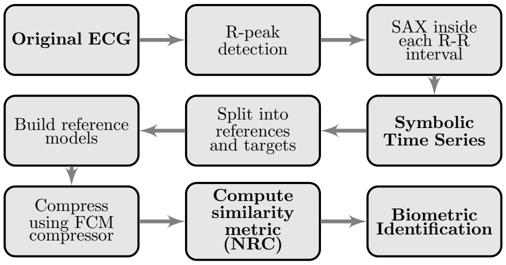
Original ECG
R-peak detection
SAX inside each R-R interval
Symbolic Time Series
Split into references and targets
Build reference models
Compress using FCM compressor
Compute similarity metric (NRC)
Biometric Identification

Fig. 1. Overview of the Biometric Identification method used in this work.

2.1 R-peak detection

The development of a robust automatic *R-peak* detector is essential, but it is still a challenging task, due to irregular heart rates, various amplitude levels and *QRS* morphologies, as well as all kinds of noise and artifacts [13].

We have decided to use a *partially fiducial* method for segmenting the ECG signal and, since this was not the major focus of the work, we used a preexisting implementation to detect *R-peaks*, based on [13]. This method detects the *R-peak* by calculating the average point between the *Q* and *S* peaks (from the *QRS complex*) – this may not give the real local maximum of the *R-peak*, but it produces a very close point. Some evaluations were done using *R-peak* detection performed by humans, in order to validate this step.

The process used for detecting the QRS complexes is somewhat similar to the one described in [13]. It uses bandpass filtering and differentiation operations, aiming to enhance the QRS complexes and to reduce out-of-band noise. A nonlinear transformation is used to obtain a positive-valued feature signal, which includes large candidate peaks corresponding to the QRS complex regions.

2.2 Quantization

We consider that the signal is already discrete in the time domain, i.e., that it is already sampled. However, we perform re-sampling using the previously detected R-peaks (Fig. 1).

The design of the *quantizer* has a significant impact on the amount of compression obtained and loss incurred in a lossy compression scheme. We have used the widely known Symbolic Aggregate ApproXimation [14], SAX, in order to quantize the ECG values into a discrete alphabet.

There is a fundamental trade-off to take into account while performing the choice of the *alphabet size*: the quality produced versus the amount of data necessary to represent the sequence [15]. From experiments, we found that using an alphabet size of 6 and 200 symbols per each R-R segment (per heartbeat) produced good results for biometric identification. However, this result does not guarantee that the same will hold true for a different dataset or application.

2.3 Compressing using finite-context models

Finite-context modeling (FCM) has been used in several areas, such as in text and image. Recent work has shown that these models have the ability to measure similarity (or dissimilarity), relying on the data algorithmic entropy [16][17][2].

A finite-context model complies to the Markov property, i.e., it estimates the probability of the next symbol of the information source using the $k > 0$ immediate past symbols (order-k context) to select the probability distribution [18]. Therefore, assuming that the k past outcomes are given by $x_{n-k+1}^{n} = x_{n-k+1} \cdots x_n$, the probability estimates, $P(x_{n+1}|x_{n-k+1}^{n})$ are calculated using symbol counts that are accumulated while the information source is processed, with

$$P(s|x_{n-k+1}^{n}) = \frac{v(s|x_{n-k+1}^{n}) + \alpha}{v(x_{n-k+1}^{n}) + \alpha|\mathcal{A}|}, \tag{2}$$

where $\mathcal{A} = \{s_1, s_2, \dots s_{|\mathcal{A}|}\}$ is the alphabet that describes the objects of interest, $v(s|x_{n-k+1}^{n})$ represents the number of times that, in the past, symbol $s \in \mathcal{A}$ was found having x_{n-k+1}^{n} as the conditioning context and where

$$v(x_{n-k+1}^{n}) = \sum_{a \in \mathcal{A}} v(a|x_{n-k+1}^{n}) \tag{3}$$

denotes the total number of events that has occurred within context x_{n-k+1}^{n}. The parameter α allows balancing between a maximum likelihood estimator and a uniform distribution. Notice that when the total number of events, n, is large (2) behaves as a maximum likelihood estimator [18].

After processing the first n symbols of x, the total number of bits generated by an order-k FCM is given by

$$-\sum_{i=1}^{n} \log_2 P(x_i|x_{i-k}^{i-1}), \tag{4}$$

where $P(x_i|x_{i-k}^{i-1}) = \frac{1}{|\mathcal{A}|}, i = \{1, \ldots, k\}$, or, in other words, we assume that the first k symbols follow a uniform distribution.

3 Experimental Results

Parameter free data mining methods are reported in the literature as efficient in classification and extraction of information. Since there is no pre-assumption about the premises, it allows true exploratory data mining. The use of FCM based compressors are examples of such methods. These methods correctly deal with some of the problems reported in the literature for ECG biometric processing, such as: variability, noise, and others.

In the models design, the algorithm takes some time (around one second). However, in testing, they are characterized by being fast. After the model is built, we found that a regular computer can run hundreds of similarity evaluations per second, using just one processor[6]. In other words, when a small ECG signal is collected, we can obtain the similarity measures to hundreds of models (one model per participant, in the case of biometric identification) that are previously built in our database and loaded into RAM. Another important factor is that this process is easily parallelizable, which means that this computation can scale as much as we want.

Since memory usage was not a concern for these preliminary tests, we computed all possible context depths k from 1 up to 40. Theoretically speaking, the number of possible contexts found by an FCM of $k = 40$, with an alphabet size of 6, would be 6^{40} (higher than 10^{31}). However, using hash tables, we do not need to compute all those combinations. In fact, from all the contexts that we computed, no model used more than 10^5 different contexts – different participants have a tendency to produce, statistically speaking, very different contexts, which is in fact what we exploit in order to distinguish amongst them.

As mentioned in [19], there is an intra-variability for each participant from one day to another, which makes the biometric identification more challenging. The tests (**a**) Day 1, (**b**) Day 2 and (**c**) Day 3 contained no information of the ECG being tested when building the models for each participant. Test (**d**) was performed using the baselines from all days (the first 350 heartbeats) and then by running the biometric identification tests using different segments from all days.

[6] All the experiments were done using Python 3.5 on an Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz, with 32GB of RAM.

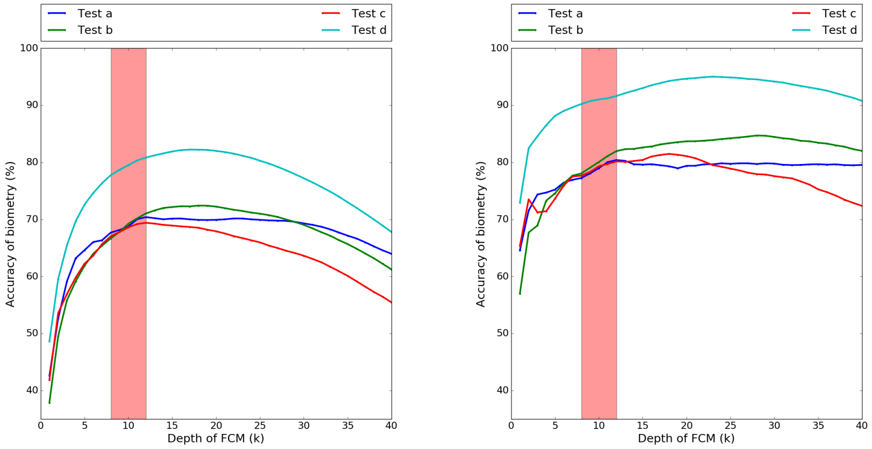

Fig. 2. Comparison of accuracies obtained using test segments with (**left**) 1 heartbeat and (**right**) 10 heartbeats, for all possible contexts $k = \{1, \ldots, 40\}$. The area in red represents what we consider the best choices for k, taking into account the complexity of the models produced.

Given that the goal of this work was to measure the minimal number of heartbeats in which it was possible to identify subjects, even in situations where they are under the effect of fear or disgust, we tried different setups: using from just one heartbeat for biometric identification, up to twenty heartbeats, in order to see the differences obtained in accuracy. In Fig. 2, it is possible to see very significant gains in performance from using one heartbeat to using ten heartbeats, for all possible contexts $k = \{1, \ldots, 40\}$.

However, in order to have a more accurate way of choosing the "ideal" number of heartbeats to collect for testing, we show a plot with the maximum performance obtained for each of the tests ran in Fig. 3. It is possible to see that the performance does not improve significantly when using more than a certain number of heartbeats for testing (area marked in red). In fact, the only test which accuracy does continue to increase is test (**d**), which is the only one that includes the baseline for all the days – even the one we are testing, which does not simulate a real world biometric identification.

4 Conclusions and Future Work

Our results showed that it is possible to identify participants accurately around 75-80% of the time with only 5 to 12 heartbeats, at least for the database

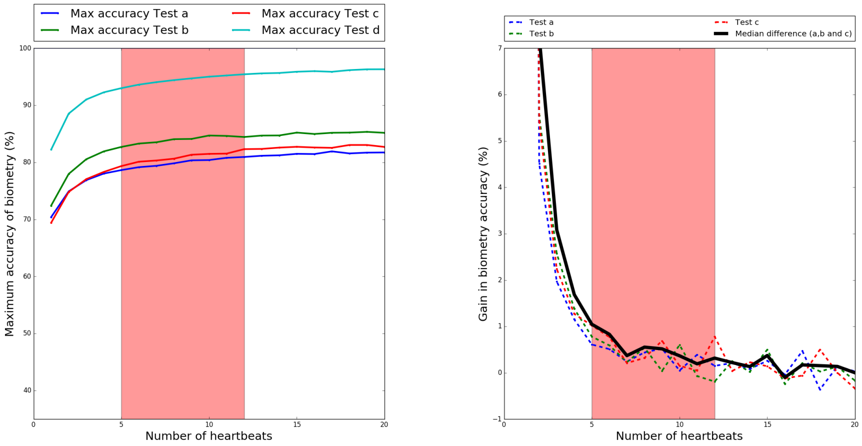

Fig. 3. (**left**) Best accuracy obtained using all individual contexts tested ($k = \{1, \ldots, 40\}$) for each number of heartbeats as target (from 1 to 20). (**right**) Difference in gain of accuracy when adding more heartbeats for the testing segments.

used. Even though the result is lower than what would be desirable for a real system, each of the experiments was performed using only one context depth ($k = \{1, \ldots, 40\}$). Previous works regarding similarities have shown empirically that collaborative models of FCMs [2][3][4] produce better approximations to the Kolmogorov Complexity (better compression ratios) and, therefore, better similarity metrics. We plan on improving the current results by using a similar approach in the near future.

Besides that, future work includes developing a compressor based on FCMs that is specific for quantized time-series. The idea is to use the fact that there is a periodical repetition, which is not currently being taken into account. This compressor will then be tested for biometric identification using the ECG.

5 Acknowledgments

This work was partially supported by national funds through the FCT - Foundation for Science and Technology, and by european funds through FEDER, under the COMPETE 2020 and Portugal 2020 programs, in the context of the projects UID/CEC/00127/2013 and PTDC/EEI-SII/6608/2014. S. Brás acknowledges the Postdoc Grant from FCT, ref. SFRH/BPD/92342/2013.

References

1. Karimian, N., Wortman, P.A., Tehranipoor, F.: Evolving authentication design considerations for the internet of biometric things (IoBT). In: Proc. Elev. IEEE/ACM/IFIP Int. Conf. Hardware/Software Codesign Syst. Synth. - CODES '16, New York, New York, USA, ACM Press (2016) 1–10
2. Pinho, A., Ferreira, P.: Image similarity using the normalized compression distance based on finite context models. 18th IEEE Int. Conf. Image Process. (2011)
3. Pratas, D., Pinho, A.J.: A Conditional Compression Distance that Unveils Insights of the Genomic Evolution. In: 2014 Data Compression Conf., IEEE (mar 2014) 421–421
4. Pinho, A.J., Pratas, D., Ferreira, P.J.S.G.: Authorship Attribution using Compression Distances. In: Data Compression Conf. (2016)
5. Pinho, A., Ferreira, P.: Finding unknown repeated patterns in images. 19th Eur. Signal Process. Conf. (EUSIPCO 2011) (2011)
6. Coutinho, D.P., Figueiredo, M.A.T.: Text Classification Using Compression-Based Dissimilarity Measures. Int. J. Pattern Recognit. Artif. Intell. **29**(05) (aug 2015)
7. Li, M., Chen, X., Li, X.: The similarity metric. IEEE Trans. Inf. Theory **50**(12) (2004) 3250–3264
8. Coutinho, D.P., Silva, H., Gamboa, H., Fred, A., Figueiredo, M.: Novel fiducial and non-fiducial approaches to electrocardiogram-based biometric systems. IET Biometrics **2**(2) (2013)
9. Coutinho, D., Fred, A., Figueiredo, M.: One-lead ECG-based personal identification using Ziv-Merhav cross parsing. In: Pattern Recognit. (ICPR), 20th Int. Conf. (2010) 3858–3861
10. Brás, S., Pinho, A.J.: ECG biometric identification: A compression based approach. In: 2015 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (aug 2015) 5838–5841
11. Li, M., Vitányi, P.: An introduction to Kolmogorov complexity and its applications. 3rd edn. Springer (1997)
12. Bennett, C., Gács, P., Li, M.: Information distance. IEEE Trans. Inf. Theory **44**(4) (1998) 1407 – 1423
13. Kathirvel, P., Sabarimalai Manikandan, M., Prasanna, S.R.M., Soman, K.P.: An Efficient R-peak Detection Based on New Nonlinear Transformation and First-Order Gaussian Differentiator. Cardiovasc. Eng. Technol. **2**(4) (oct 2011) 408–425
14. Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A symbolic representation of time series, with implications for streaming algorithms. In: DMKD '03 Proc. 8th ACM SIGMOD Work. Res. issues data Min. Knowl. Discov. (2003) 2–11
15. Gonzalez C., R., Woods, R.E.: Sampling and Quantization. In: Digit. Image Process. Prentice Hall PTR (2007)
16. Garcia, S., Rodrigues, J., Santos, S., Pratas, D., Afreixo, V., Bastos, C., Ferreira, P., Pinho, A.: A Genomic Distance for Assembly Comparison Based on Compressed Maximal Exact Matches. IEEE/ACM Trans. Comput. Biol. Bioinforma. **10**(3) (may 2013) 793–798
17. Pratas, D., Pinho, A.J., Sara P. Garcia: Computation of the Normalized Compression Distance of DNA Sequences using a Mixture of Finite-context Models. In: BIOINFORMATICS. (2012) 308–311
18. Brás, S., Ferreira, J., Soares, S.C., Pinho, A.J.: Biometric and Emotion Identification: an ECG compression based method. (Submitted) (2016)
19. Agrafioti, F., Hatzinakos, D., Anderson, A.K.: ECG Pattern Analysis for Emotion Detection. IEEE Trans. Affect. Comput. **3**(1) (jan 2012) 102–115