Abstract—The advance of falsification technology increases security concerns and gives biometrics an important role in security solutions. The electrocardiogram (ECG) is an emerging biometric that does not need liveliness verification. There is strong evidence that ECG signals contain sufficient discriminative information to allow the identification of individuals from a large population. Most approaches rely on ECG data and the fiducia of different parts of the heartbeat waveform. However non-fiducial approaches have proved recently to be also effective, and have the advantage of not relying critically on the accurate extraction of fiducia data. In this paper, we propose a new non-fiducial ECG biometric identification method based on data compression techniques, namely the Ziv-Merhav cross parsing algorithm for symbol sequences (strings). Our method relies on a string similarity measure which can be seen as a compression-based approximation of the algorithmic cross complexity. We present results on real data, one-lead ECG, acquired during a concentration task, from 19 healthy individuals. Our approach achieves 100% subject recognition rate despite the existence of differentiated stress states.

I. INTRODUCTION

Biometrics deals with identification of individuals based on their physiological or behavioral characteristics [1] and plays an important role in security systems. Traditional methods of biometric identification, such as those using fingerprints or iris, provide accurate identification but lack robustness against falsification.

The electrocardiogram (ECG) is an emerging biometric tool exploiting a physiological feature that exists in all humans; there is a strong evidence that the ECG is sufficiently discriminative to identify individuals in a large population. The ECG has intrinsic liveliness verification, and allows personal identification and authentication, and detection of different stress or emotional states [2]. The ECG can also be used together with other biometric measures [3], as a complementary feature, for fusion in a multimodal system [4, Ch. 18] and for continuous verification where biological signatures are continuously monitored (easily done by using new signal acquisition technologies, such as the Vital Jacket [5]) in order to guarantee the identity of the operator throughout the whole process [6].

A typical ECG signal of a normal heartbeat can be divided into 3 parts, as depicted in Figure 1: the P wave (or P complex), which indicates the start and end of the atrial depolarization of the heart; the QRS complex, which corresponds to the ventricular depolarization; and, finally, the T wave (or T complex), which indicates the ventricular repolarization. It is known that the shape of these complexes differs from person to person, a fact which has stimulated the use of the ECG as a biometric [7].

In a broad sense, one can say there are two different approaches in the literature concerning feature extraction from ECG: fiducial [7] [8] [9] [10] and non-fiducial [11] [12]. Fiducial methods use points of interest within a single heartbeat waveform, such as local maxima or minima; these points are used as reference to allow the definition of latency times (features), as shown in Figure 1. Several methods exist that extract different time and amplitude features, using these reference points. Non-fiducial techniques aim at extracting discriminative information from the ECG waveform without having to localize fiducial points. In this case, a global
pattern from several heartbeat waveforms may be used as a feature. Some methods combine these two different approaches or are partially fiducial [13] (e.g., they use only the R peak as a reference for segmentation of the heartbeat waveforms). Table I summarizes several approaches found in the literature; for more details on each method, see the corresponding publication.

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Feature</th>
<th>Method</th>
<th>Subjs</th>
<th>Accur.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[7]</td>
<td>Fiducial</td>
<td>PCA</td>
<td>20</td>
<td>100%</td>
</tr>
<tr>
<td>[8]</td>
<td>Fiducial</td>
<td>Templ. matching+DBNN</td>
<td>20</td>
<td>100%</td>
</tr>
<tr>
<td>[9]</td>
<td>Fiducial</td>
<td>LDA</td>
<td>29</td>
<td>98 %</td>
</tr>
<tr>
<td>[10]</td>
<td>Fiducial</td>
<td>FSE</td>
<td>26</td>
<td>99.97%</td>
</tr>
<tr>
<td>[11]</td>
<td>Non-fiducial</td>
<td>Wavelet Distance</td>
<td>50</td>
<td>95%</td>
</tr>
<tr>
<td>[12]</td>
<td>Non-fiducial</td>
<td>Wavelet Distance</td>
<td>35</td>
<td>100%</td>
</tr>
<tr>
<td>[13]</td>
<td>Non-fiducial</td>
<td>AC/DCT+KNN</td>
<td>13</td>
<td>97.8%</td>
</tr>
<tr>
<td>Ours</td>
<td>Non-fiducial</td>
<td>Cross Parsing+MDL</td>
<td>19</td>
<td>100%</td>
</tr>
</tbody>
</table>

This paper introduces a new non-fiducial ECG biometric identification method that uses averaged single heartbeat waveforms and is based on data compression techniques, namely the Ziv-Merhav cross parsing algorithm for sequences of symbols, which derives from algorithmic cross complexity concept and its compression-based approximation. We present results on real data, using one-lead ECG acquisition during a concentration task. On a set of 19 healthy individuals, our method achieves 100% subject recognition rate despite the existence of differentiated stress states in the ECG signals [2].

The outline of the paper is as follows. In Section 2, we review the fundamental tools underlying our approach: Lempel-Ziv string parsing and compression; the Ziv-Merhav cross parsing algorithm. Section 3 presents the proposed classification method. Experimental results are presented in Section 4, while Section 5 concludes the paper.

II. THE LEMPEL-ZIV AND ZIV-MERHAV ALGORITHMS

The Lempel-Ziv (LZ) algorithm is a well-known tool for text compression [14] [15] [16] [17], which in recent years has also been used for classification purposes (see [18] and references therein). In particular, in [18], we have shown how the Ziv-Merhav (ZM) method for measuring relative entropy [19] (which is based on Lempel-Ziv-type string parsing) achieves state-of-the-art performance in a specific text classification task. We will now briefly review these algorithms.

• The incremental LZ parsing algorithm [15], is a self parsing procedure of a sequence into \( c(z) \) distinct phrases such that each phrase is the shortest sequence that is not a previously parsed phrase. For example, let \( n = 11 \) and \( z = (01111000110) \), then the self incremental parsing yields \((0, 1, 11, 10, 00, 110)\), namely, \( c(z) = 6 \).

• The ZM algorithm, a variant of the LZ parsing algorithm, is a sequential parsing of a sequence \( z \) with respect to another sequence \( x \) (cross parsing). Let \( c(z|x) \) denote the number of phrases in \( z \) with respect to \( x \). For example, let \( z \) be as above and \( x = (10010100110) \); then, parsing \( z \) with respect to \( x \) yields \((011, 110, 00110)\), that is \( c(z|x) = 3 \).

Roughly speaking, we can see \( c(z) \) as a measure of the complexity of the sequence \( z \), while \( c(z|x) \), the description length obtained when coding \( z \) using a model for \( x \) (cross parsing), can be seen as an estimate of the cross complexity [20]. It is expectable that the cross complexity is low when the two sequences are very similar; this is the key idea behind the use of ZM cross parsing in classification [18], which in this paper will be adopted for ECG-based personal identification.

An implementation of the ZM cross parsing algorithm was proposed in [18], based on a modified LZ77 [14] algorithm, where the dictionary is static and only the lookahead buffer (LAB) slides over the input sequence.

An implementation of the ZM cross parsing algorithm was proposed in [18], based on a modified LZ77 [14] algorithm, where the dictionary is static and only the lookahead buffer slides over the input sequence, as shown in Figure 2 (for more details, see [18]). This very same implementation, using a 64 Kbyte dictionary and a 256 byte lookahead buffer, was used in the experiments reported below.

III. PROPOSED CLASSIFICATION METHOD

To use ZM-based tools for classification, a necessary first step is the conversion of the ECG (discrete-time analog) signal into a sequence of symbols. In this paper, we propose a very simple approach based on quantization. Assuming we are given a set of single heartbeat waveforms (resulting from a segmentation preprocessing stage), we simply apply 8-bit (256 levels) uniform quantization, thus obtaining a sequence...
of symbols (from a 256 symbols alphabet) from each single heartbeat.

Consider a collection of training samples partitioned into \( K \) classes (the set of subjects to be identified): \( \mathcal{X} = \{ \mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_K \} \). For each subject/class \( k \), \( \mathcal{X}_k \) contains \( n \) strings obtained from the same number of heartbeats using the quantization procedure described in the previous paragraph. A string \( x_k \) is formed by concatenating the \( n \) training strings of subject \( k \); string \( x_k \) is, in some sense, a “model” representing the shape of the heartbeats of subject \( k \).

Given a test sample \( z \) (containing the string representing \( m \) heartbeats) obtained from an unknown subject (assumed to be one from which the training set was obtained), its identity is estimated as follows:

\[
\hat{k}(z) = \arg \min_{k \in \{1, \ldots, K\}} c(z|x_k),
\]

where \( c(z|x_k) \) is computed by the ZM cross parsing algorithm, as described in Section II. In other words, the test sample is classified as belonging to the subject that leads to its shortest description. Although using different tools, this approach is related in spirit with the minimum incremental coding length (MICL) approach [21].

IV. EXPERIMENTS

A. Data collection

The ECG waveform dataset used was acquired using one lead, in the context of the Himotion project. The dataset contains ECG recordings from 19 subjects acquired during a concentration task on a computer, designed for an average completion time of 10 minutes. All the acquired ECG signals were normalized and band-pass filtered (2–30Hz) in order to remove noise. Each heartbeat waveform was sequentially segmented from the full recording and then all the obtained waveforms were aligned by their R peaks. From the resulting collection of ECG heartbeat waveforms, the mean wave for groups of 10 consecutive waveforms (without overlap) was computed. Each of these mean waveforms is what we call a single heartbeat in Section III. Notice that an intra-class study [2] with the dataset, in the context of the exploration of electrophysiological signals for emotional states detection, showed the existence of differentiated states in the data that represent the ECG signal of a subject.

B. Experimental Results

The reported results are averages over 30 runs. In each run, we partition the set of heartbeats of each subject into two mutually exclusive subsets: one of these subsets is used to form the training data set \( \mathcal{X} = \{ \mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_K \} \), while the other is used to build the test waveforms. We consider several values for \( n \) (the number of “model” strings) as well as for \( m \) (the number of test waveforms).

The results of this experiment, which are depicted in Figure 3, show that the proposed method achieves 100% accuracy for \( m = 12 \) and \( n = 13 \) or \( n = 20 \). This is better than the results reported in [10] over the same dataset. The approaches in [7], [8], [12], were not tested on the same dataset, so the results are not directly comparable. Notice that using only \( m = 5 \) waveforms for the test patterns, we already reach an accuracy around 99.5%. As expected, the accuracy increases both with \( n \) and \( m \).

IV. EXPERIMENTS

A. Data collection

The ECG waveform dataset used was acquired using one lead, in the context of the Himotion project. The dataset contains ECG recordings from 19 subjects acquired during a concentration task on a computer, designed for an average completion time of 10 minutes. All the acquired ECG signals were normalized and band-pass filtered (2–30Hz) in order to remove noise. Each heartbeat waveform was sequentially segmented from the full recording and then all the obtained waveforms were aligned by their R peaks. From the resulting collection of ECG heartbeat waveforms, the mean wave for groups of 10 consecutive waveforms (without overlap) was computed. Each of these mean waveforms is what we call a single heartbeat in Section III. Notice that an intra-class study [2] with the dataset, in the context of the exploration of electrophysiological signals for emotional states detection, showed the existence of differentiated states in the data that represent the ECG signal of a subject.

B. Experimental Results

The reported results are averages over 30 runs. In each run, we partition the set of heartbeats of each subject into two mutually exclusive subsets: one of these subsets is used to form the training data set \( \mathcal{X} = \{ \mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_K \} \), while the other is used to build the test waveforms. We consider several values for \( n \) (the number of “model” strings) as well as for \( m \) (the number of test waveforms).

The results of this experiment, which are depicted in Figure 3, show that the proposed method achieves 100% accuracy for \( m = 12 \) and \( n = 13 \) or \( n = 20 \). This is better than the results reported in [10] over the same dataset. The approaches in [7], [8], [12], were not tested on the same dataset, so the results are not directly comparable. Notice that using only \( m = 5 \) waveforms for the test patterns, we already reach an accuracy around 99.5%. As expected, the accuracy increases both with \( n \) and \( m \).

IV. EXPERIMENTS

A. Data collection

The ECG waveform dataset used was acquired using one lead, in the context of the Himotion project. The dataset contains ECG recordings from 19 subjects acquired during a concentration task on a computer, designed for an average completion time of 10 minutes. All the acquired ECG signals were normalized and band-pass filtered (2–30Hz) in order to remove noise. Each heartbeat waveform was sequentially segmented from the full recording and then all the obtained waveforms were aligned by their R peaks. From the resulting collection of ECG heartbeat waveforms, the mean wave for groups of 10 consecutive waveforms (without overlap) was computed. Each of these mean waveforms is what we call a single heartbeat in Section III. Notice that an intra-class study [2] with the dataset, in the context of the exploration of electrophysiological signals for emotional states detection, showed the existence of differentiated states in the data that represent the ECG signal of a subject.

B. Experimental Results

The reported results are averages over 30 runs. In each run, we partition the set of heartbeats of each subject into two mutually exclusive subsets: one of these subsets is used to form the training data set \( \mathcal{X} = \{ \mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_K \} \), while the other is used to build the test waveforms. We consider several values for \( n \) (the number of “model” strings) as well as for \( m \) (the number of test waveforms).
Vital Jacket [5], and included in a multimodal system. Current work include further evaluation of our method when used for authentication purposes, with ROC curve design for false acceptance rate (FAR) and false rejection rate (FRR) analysis [22].

ACKNOWLEDGMENTS

We acknowledge the following financial support: the FET programme, within the EU FP7, under the SIMBAD project (contract 213250); Fundação para a Ciência e Tecnologia (FCT), under grants PTDC/EEA-TEL/72572/2006 and QREN 3475.

REFERENCES


