Available online at www.sciencedirect.com

ScienceDirect

Pattern Recognition Letters 29 (2008) 501–514

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Matching of quasi-periodic time series patterns by exchange of block-sorting signatures

Bachir Boucheham *

University of Skikda, 20 Août 1955, Department of Informatics, BP 26, Route El-Hadaek, Skikda DZ21000, Algeria

Received 15 June 2007; received in revised form 4 November 2007
Available online 21 November 2007

Communicated by R.C. Guido

Abstract

We propose a novel method for quasi-periodic time series patterns matching, through signature exchange between the two patterns. The signature is obtained through sorting of the time series on magnitude. The advantage is that the difficult task of comparing the two patterns can be easily performed as a result of this exchange: The original time series is compared (point to point matching) to the reconstructed time series obtained through the reverse process, using the other time series signature. The matching is such that periods in one time series are put into correspondence with periods in the other time series, even if the time series is of different basic patterns and/or different lengths. The method is simple to implement and requires no parameters. It was compared to the very appreciated DTW algorithm on execution time, space and accuracy. Due to the quasi-periodic nature of the electrocardiogram, the tests were performed on ECG traces, selected from the Massachusetts Institute of Technology – Beth Israel Hospital (MITBIH) public database. Results show that the proposed method outperforms DTW on all aspects. This suggests that our method could be a good alternative to the classical DTW technique for quasi-periodic signals comparison. Specific applications are foreseen for our method: Novelty detection and person identification using ECG.

Keywords: Pattern matching; Time series; Block-sorting signature; Shape exchange; DTW

1. Introduction

Time series are measurements of different sorts of phenomena taken at regular intervals. For instance, we can cite natural phenomena (e.g. river levels), physiological phenomena (e.g. electrocardiogram), economical phenomena (e.g. stock exchange index) or artificial phenomena (e.g. traffic density in a computer network) as examples of useful time series. Analysis and interpretation of such time series are of great importance for experts in their respective fields. Common applications in that respect include Creation and querying of specialized databases, compression and storage or transmission of time series, and interpretation of time series patterns, including event and abnormal behavior detection. This study is exclusively dedicated to quasi-periodic time series. These are time series that are concatenations of quasi-similar patterns called pseudo periods (periods for short). Our interest in quasi-periodic time series is due to the fact that many phenomena are of quasi-periodic nature. We particularly focus the study on the electrocardiogram (ECG): A quasi-periodic physiological time series, intensively used in medical care. A typical electrocardiogram is a sequence of quasi-regular periods reflecting heart cycles. Each cycle is itself a concatenation of three consecutive basic patterns: P wave, QRS complex and T wave. Fig. 1 presents a two period ECG trace with its main patterns.

In previous works, we developed algorithms and methods for ECG baseline wandering detection and correction (Boucheham et al., 2005), QRS detection (Boucheham et al., 2003), short-term compression (Boucheham et al.,

* Tel./fax: +213 038 701 700.
E-mail address: boucbac@yahoo.com

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.11.004

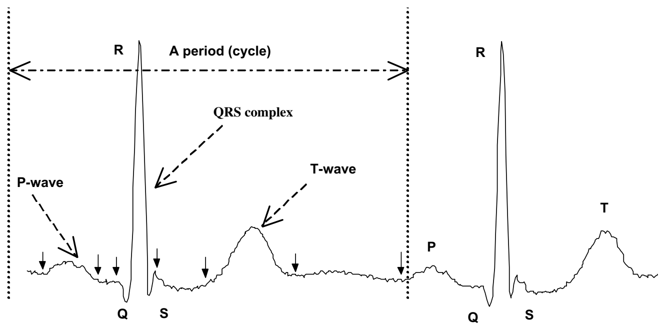

Fig. 1. A typical ECG segment, where up–down arrows delimit the sub-patterns. The figure shows a two period ECG. Each period is composed of three clinically significant basic patterns: *P* wave, QRS complex and *T* wave. Each period corresponds to one heartbeat cycle.

2006) and long-term compression (Boucheham, 2007). In the current work, we focus on resolving the pattern matching problem for this type of time series. This problem can be stated this way: Given two ECG time series, say X and Y, find a method to decide whether they are similar. This is an important problem, the solution of which could have impact on many interesting ECG applications: Clustering, data-mining, novelty detection, automatic diagnosis, person identification, and perhaps other applications. We stress the point that, although this work is centered on ECG time series, the study is also suitable to other quasi-periodic time series, since we do use no priori knowledge on the nature of the data. For this reason, we indifferently refer to the processed data by ECG or (quasi-periodic) time series.

In (Boucheham, 2007), we noticed that sorting of ECG signals on magnitude yields a kind of signature of the time series, in the sense that the obtained trace was a global and a stable characteristic of the used time series, in each test. This observation was behind the idea to use this signature to compare quasi-periodic time series patterns. This last operation would allow quantifying the similarity between two given time series patterns. However, it is desirable to derive a detailed matching using this signature. This desire is motivated by the observation that human experts can visually tell in all cases if two time series are similar or not. In case of dissimilarity, they can even indicate the regions where there is disagreement between the two patterns. In a sense, when comparing time series visually, we somehow try to compare periods in each time series. Therefore, ideally, periods in one time series should be put into correspondence with periods in the other time series, even if the number of periods, lengths or the sources of the two times series were different. We show that this operation is possible, no matter the number of periods in each pattern and no matter if the two patterns come from different sources or not. The comparison operation is made possible through exchange of the two time series respective signatures, while using the respective time indexes for each series. For this reason, we use the acronym SEA (Shape Exchange Algorithm) for our method. We also show that the very appreciated DTW method, as far as time series comparison is concerned, is unable to deal with all situations, whereas our method never fails. In addition, we show that the proposed SEA method is more accurate, faster and less memory consuming than DTW. We also illustrate some specific applications we foresee for our method.

The rest of this work is organized as follows. In Section 2, the problem of similarity measure for time series comparison is stated in a more detailed manner and related works are cited and commented. In particular, the DTW method is presented in brief in this section. In Section 3, the main materials and methods for the proposed solution are presented. Especially, the proposed SEA method is presented in this section. In Section 4, a thorough comparison SEA vs. DTW on ECG time series is presented and applications using the SEA method are illustrated. In Section 5, results of the conducted tests are discussed and concluding thoughts and future works are announced.

2. Similarity measure for time series matching

2.1. Problem statement

Given two time series $X = (x_i)$, $i = 1{:}n$ and $Y = (y_j)$, $j = 1{:}m$, the problem of comparing X and Y consists in proposing a similarity measure $d(X, Y)$ capable of quantifying the degree of resemblance between X and Y. In other words, quantity $d(X, Y)$ should allow concluding if the two patterns 'fit' and how good the fit is. One of the earliest pattern matching techniques is the Euclidian distance. This measure is defined only for equal lengths time series ($n = m$). In this case, the Euclidian distance between X and Y is defined by Eq. (1)

$$\text{Euclidian}(X, Y) = \sqrt{\sum\nolimits_{i=1}^{N}[x_i - y_i]^2} \qquad (1)$$

The advantage of the Euclidian distance is its simplicity. It also has an advantageous $O(n)$ computational complexity. However, this measure is not efficient in direct data comparison due to its sensitivity to small distortions in the time axis (Keogh and Pazzani, 2000; Keogh and Ratanamahatana, 2005). For instance, Fig. 2a illustrates the inefficiency of this measure on a small ECG time series. It is easy to derive from the figure that the time series $S1$ and $S2$ are similar. However, due to a shift in the time axis of one time series with respect to the other, the Euclidian distance would rather induce a very large dissimilarity measure because it assumes that the ith point in one series is aligned with the ith point in the other series. More generally, the problem with the Euclidian distance is its inability to deal with four basic difficulties encountered in time series matching: Noise (Fig. 3a), offset translation (Fig. 3b), amplitude scaling (Fig. 3c) and time axis scaling (Fig. 3d). Many methods have been developed to deal with these problems. Obviously, time scaling is the most difficult aspect in time series pattern matching.