PCA-based Time Series Similarity Search

Leonidas Karamitopoulos, Georgios Evangelidis, and Dimitris Dervos

Abstract We propose a novel approach in multivariate time series similarity search for the purpose of improving the efficiency of data mining techniques without substantially affecting the quality of the obtained results. Our approach includes a representation based on Principal Component Analysis (PCA) in order to reduce the intrinsically high dimensionality of time series, and utilizes as a distance measure a variation of the Squared Prediction Error (SPE), a well-known statistic in the Statistical Process Control community. Contrary to other PCA-based measures proposed in the literature, the proposed measure does not require applying the computationally expensive PCA technique on the query. In this paper, we investigate the usefulness of our approach in the context of query by content and 1-NN classification. More specifically, we consider the case where there are frequently arriving objects that need to be matched with the most similar object in a database or that need to be classified into one of several pre-determined classes. We conduct experiments on four datasets used extensively in the literature, and we provide the results of the performance of our measure and other PCA-based measures with respect to classification accuracy and precision/recall. Experiments indicate that our approach is at least comparable to other PCA-based measures and a promising option for similarity search within the Data Mining context.

Leonidas Karamitopoulos, Georgios Evangelidis
University of Macedonia, Department of Applied Informatics, 156 Egnatia Str., GR-54006, Thessaloniki, Greece, e-mail: {lkaramit, gevan}@uom.gr

Dimitris Dervos
Alexander Technology Educational Institute of Thessaloniki, Department of Informatics, P.O. Box 141, GR-57400 Sindos, Greece, e-mail: dad@it.teithe.gr

1 Introduction

Rapid advances in automated monitoring systems and storing devices have led to the generation of huge amounts of data in the form of time series, that is, series of measurements recorded through time. Inevitably, (most of) this volume of data remains unexploited, since the traditional methods of analyzing data do not adequately scale to the massive datasets frequently encountered. In the last decade, there has been an increasing interest in the Data Mining field, which involves techniques and algorithms capable of efficiently extracting patterns that can potentially constitute knowledge from very large databases.

The field of time series data mining mainly considers methods for the following tasks: clustering, classification, novelty detection, motif discovery, rule discovery, segmentation and indexing [30]. At the core of these tasks lies the concept of similarity, since most of them require searching for similar patterns [21]. Two time series can be considered similar when they exhibit similar shape or pattern. However, the presence of high levels of noise demands the definition of a similarity/distance measure that allows imprecise matches among series [8]. In addition to that, the intrinsically high dimensionality of time series affects the efficiency of data mining techniques. Note that the dimensionality is defined by the length of the time series. In other words, each time point can be considered as a feature whose value is recorded. Thus, an appropriate representation of the time series is necessary in order to manipulate and efficiently analyze huge amounts of data. The main objective is to reduce the dimensionality of a time series by representing it in a lower dimension and analyze it in this dimension. There have been several time series representations proposed in the literature for the purpose of dealing with the problem of the “dimensionality curse” that appears frequently within real world data mining applications [7, 9].

In this paper, we consider the case of multivariate time series, that is, a set of time series recorded at the same time interval. Contrary to the univariate case, the values of more than one attribute are recorded through time. The objects under consideration can be expressed in the form of matrices, where columns correspond to attributes and rows correspond to time instances. Notice that a univariate time series can be expressed as a column (or row) vector that corresponds to the values of one attribute at consecutive time instances. Multivariate time series frequently appear in several diverse applications. Examples include human motion capture [27], geographical information systems [7], statistical process monitoring [20], or intelligent surveillance systems [34]. For instance, it is of interest to form clusters of objects that move similarly by analyzing data from surveillance systems or classify current operating conditions in a manufacturing process into one of several operational states.

As a motivating example, consider the task of automatically identifying people based on their gait. Suppose that data is generated using a motion capture system, which transmits the coordinates of 22 body joints every second (i.e. 66 values) for two minutes (i.e. 120 seconds). The resulting dataset consists of 66 time series and 120 time instances, and corresponds to a specific person. This dataset can be expressed as a matrix $X_{120\times66}$. Also, suppose that we have obtained gait data for every known person under different conditions, for example, under varying gait speeds, and stored it in a database. Each record corresponds to one person and holds the gait data, which can be considered as a matrix, along with a label that indicates the identity of this person. Note that there is more than one record that corresponds to the same person, since we have obtained gait data under different conditions for every known person. Given this database, the objective is to identify a person under surveillance. In this case, we search the database for the most similar matrix to the one that is generated by this person. This task can be considered as a classification task. Each known person represents a class that consists of the gait data of this person generated under different conditions. The task is to classify (identify) a person under surveillance into a class.

This classification problem can be virtually handled by (other) classic classification techniques [35, 29], if each matrix is represented as a vector by concatenating its columns (i.e., the values of the corresponding attributes). However, we have to consider two issues with respect to this approach. The first issue is that the problem of high dimensionality deteriorates in the case of multivariate time series, since it is not only the length of the time series, but also the number of attributes that determine the dimensionality. In the previous example, the matrix $X_{120\times66}$ that corresponds to the gait data of one person constitutes an object of 7920 (120×66) dimensions. The second issue is that the correlations among attributes of the same multivariate time series are ignored. This loss of information may be of serious importance within a classification application.

We introduce a novel approach in identifying similar multivariate time series, which includes a PCA-based representation for the purpose of dimensionality reduction and a distance measure that is based on this representation. Principal Component Analysis (PCA) is a well-known statistical technique that can be used to reduce the dimensionality of a multivariate dataset by condensing a large number of interrelated variables into a smaller set of variates, while retaining as much as possible of the variation present in the original dataset [16]. In our case, the interrelated variables are in the form of time series. We provide a novel PCA-based measure that is a variation of the Squared Prediction Error (SPE) or Q-statistic, which is broadly utilized in Multivariate Statistical Process Control [22]. Contrary to other PCA-based measures proposed in the literature, this measure does not require applying the computationally expensive PCA technique on the query. Moreover, we provide a method that further speeds up the calculations of the proposed measure by reducing the dimensionality of each one of the time series that form the query object during the pre-processing phase. Although our approach can be applied on other types of data, we concentrate on time series for two reasons. First, this type of data differs from other domains in that it exhibits high dimensionality, high feature correlation, and high levels of noise. Second, a large portion of data is generated in the form of time series in almost all real- world applications.

The objective of our approach is to provide a means for improving the efficiency of data mining techniques without substantially affecting the quality of the corresponding results. In particular, the dimensionality reduction of the original data improves the scalability of any data mining technique that will be applied subsequently, and the proposed measure aims at maintaining the quality of the results. In this paper, we investigate the potential usefulness of our approach, mainly in the context of query by content and 1-NN classification. More specifically, we consider the case where there are frequently arriving objects that need to be matched with the most similar object in a database or that need to be classified into one of several pre-determined classes.

In Section 2, we discuss PCA with respect to similarity search and we provide related work. Section 3 introduces our approach and provides a distance measure that is based on Multivariate Statistical Process Control. In Section 4, we describe the experimental settings with respect to the datasets, the methods, and, the rival measures. The results of our experiments are presented and discussed in Section 5. Finally, conclusions and future work are provided in Section 6.

2 Background

We briefly review Principal Component Analysis on multivariate data in Section 2.1. Similarity search is based on shapes, meaning that two time series are considered similar when their shapes are considered to be "close enough". Apparently, the notion of "close enough" depends heavily on the application itself, a fact that affects the decision of the pre-processing phase steps to be followed, the similarity measure to be utilized and the representation to be applied on the raw data (Section 2.2). Finally, in Section 2.3, we review several PCA-based measures.

2.1 Review of PCA

PCA is applied on a multivariate dataset, which can be represented as a matrix $X_{n\times p}$. In the case of time series, n represents their length (number of time instances), whereas p is the number of variables being measured (number of time series). Each row of X can be considered as a point in p-dimensional space. The objective of PCA is to determine a new set of orthogonal and uncorrelated composite variates $Y_{(j)}$, which are called principal components:

$$Y_{(j)} = a_{1j}X_1 + a_{2j}X_2 + \ldots + a_{pj}X_p, \quad j = 1,2,\ldots p \tag{1}$$

The coefficients a_{ij} are called component weights and X_i denotes the i^{th} variable. Each principal component is a linear combination of the original variables and is derived in such a manner that its successive component accounts for a smaller portion of variation in X. Therefore, the first principal component accounts for the largest portion of variance, the second one for the largest portion of the remaining variance, subject to being orthogonal to the first one, and so on. Hopefully, the first m components will retain most of the variation present in all of the original variables (p). Thus, an essential dimensionality reduction may be achieved by projecting the original data on the new m-dimensional space, as long as, $m \ll p$.

The derivation of the new axes (components) is based on Σ, where Σ denotes the covariance matrix of X. Each eigenvector of Σ provides the component weights a_{ij} of the $Y_{(j)}$ component, while the corresponding eigenvalue, denoted λ_j, provides the variance of this component. Alternatively, the derivation of the new axes can be based on the correlation matrix, producing slightly different results. These two options are equivalent when the variables are standardized (i.e. they have mean equal to zero and standard deviation equal to one).

Intuitively, PCA transforms a dataset X by rotating the original axes of a p-dimensional space and deriving a new set of axes (components), as in Fig. 1. The component weights represent the angles between the original and the new axes. In particular, the component weight a_{ij} is the cosine of the angle between the i^{th} original axis and the j^{th} component [12]. The values of $Y_{(j)}$ calculated from Eq. 1 provide the coordinates of the original data in the new space.

Conclusively, the application of PCA on a multivariate dataset $X_{n\times p}$ results in two matrices, in particular the matrix of component weights $A_{p\times p}$ and the matrix of variances $\Lambda_{p\times 1}$. In addition to that, the matrix of the new coordinates $Y_{n\times p}$ of the original data can be calculated from A, since $Y = X \cdot A$.

2.2 Implications of PCA in Similarity Search

Regarding the pre-processing phase, there are four main distortions that may exist in raw data, namely, offset translation, amplitude scaling, time warping and noise. Distance measures may be seriously affected by the presence of any of these distortions, resulting most of the times in missing similar shapes. Offset translation refers to the case where there are differences in the magnitude of the values of two time series, while the general shape remains similar (Fig. 2). This distortion is inherently handled by PCA, since it is based on covariances, which are not affected by the magnitude of the values. This is a potential disadvantage of PCA, if simi-

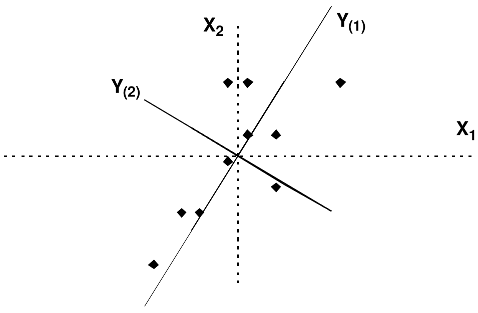

Fig. 1 A multivariate time series consisting of two variables (X_1 and X_2) and ten time instances. Dots represent the time instances, while solid lines represent the principal components that have been derived by PCA. A dimensionality reduction can be achieved, if only the first component $Y_{(1)}$ is retained and data is projected on it.

larity search is to be based also on the magnitude of the values. Amplitude scaling refers to the case where there are differences in the magnitude of the fluctuations of two time series, while the general shape remains similar (Fig. 2). In this case, PCA representation can be based on the correlations among variables, instead of the covariances. This is an alternative way of deriving the principal components that produces slightly different results, but not essentially different in the context of dimensionality reduction. Time warping, which may be global or local, refers to the acceleration or deceleration of the evolvement of a time series through time. In the case of global time warping (i.e. two multivariate time series evolve in different rates), PCA representation is expected to be similar, since the shorter time series can be considered as a systematic random sample of the longer one, resulting to a similar covariance matrix. Intuitively, the existence of local time warping distortions may be captured by the covariances of the corresponding variables. Finally, noise is intrinsically handled by PCA, since the discarded principal components account mainly for variations due to noise.

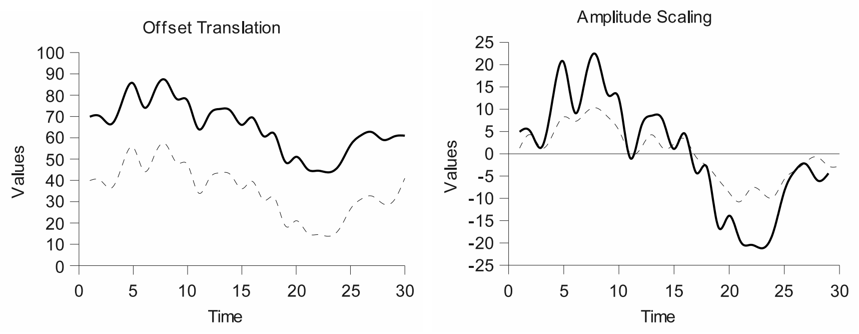

Fig. 2 Two of the distortions that may exist in raw time series data

Another issue in the pre-processing phase is the handling of time series of different lengths. PCA requires variables (time series) of equal length for the same object. For example, an object $X_{n\times p}$ consists of p time series that all have the same length n. Therefore, this is a limitation of this technique. However, similarity search is performed among objects, and thus, it is based on the produced matrices $A_{p\times p}$ and $\Lambda_{p\times 1}$, which are independent of the lengths of the series. For example, the comparison of two objects $X_{n\times p}$ and $Q_{m\times p}$ is feasible, since their PCA representations are independent of n and m, respectively.

Similarity search also requires a measure that quantifies the similarity or dissimilarity between two objects. Under PCA transformation, this measure should be based on at least one of the produced matrices, mentioned in the previous paragraph, $A_{p\times p}$, $\Lambda_{p\times 1}$, and $Y_{n\times p}$. The central concept is that, if two multivariate time series are similar, their PCA representations will be similar, that is, the produced matrices will be close enough. Searching similarity based on $A_{p\times p}$, means to compare the angles of principal components derived from two multivariate time series, whereas searching based solely on $Y_{n\times p}$ is useless, since these values are coordinates in different spaces. $\Lambda_{p\times 1}$ contains information about the shape of the time series and it may be used in conjunction with $A_{p\times p}$ for further distinguishing power.

The PCA representation of a dataset $X_{n\times p}$ consists of the component weight matrix $A_{p\times p}$ and the variances matrix $\Lambda_{p\times 1}$. The data reduction may be substantial as long as the number of time instances n is much greater than the number of variables p. Moreover, a further data reduction can be achieved, if only m components are retained, where $m < p$. There are several criteria for determining the number of components to retain, such as the scree graph or the cumulative percentage of total variation [16]. According to the latter criterion, one could select that value for m, for which the first m components retain more than 90% of the total variation present in the original data.

Although PCA-based similarity search is complicated and usually requires expensive computations, it may improve the quality of similarity search providing at the same time useful information for post hoc analysis.

2.3 Related Work

Although there is a vast literature in univariate time series similarity search, the case of multivariate time series has not been extensively explored. Most of the papers concentrate on indexing multidimensional time series and provide an appropriate representation scheme and/or a similarity measure. In addition to that, most of the research interest lays on trajectories, which usually consist of 2 or 3 dimensional time series.

The authors of [37] and [5] suggest similarity measures based on the Longest Common Subsequence (LCSS) model, whereas a modified version of the Edit Distance for real-valued series is provided in [8]. Bakalov et al. [2] extend the Symbolic Aggregate Approximation (SAX) [26] and the corresponding distance measure for multivariate time series. Vlachos et al. [36] propose an indexing framework that supports multiple similarity/distance functions, without the need to rebuild the index. Several researchers approach similarity search by applying a measure and/or an indexing method on transformed data. Kahveci et al. [18] propose to convert a p-dimensional time series of length n to a univariate time series of length np by concatenation, and then apply a representation scheme for the purpose of dimensionality reduction. Lee et al. [24] propose a scheme for searching a database, which, in the pre-processing phase includes the representation (e.g. DFT) of each one of the p time series separately. Cai & Ng [6] approximate and index multidimensional time series with Chebyshev polynomials. In the latter three papers, the Euclidean distance is applied as a distance measure.

On the other hand, there are several PCA-based measures that have been proposed in order to compare two objects, which are in the form of multivariate time series. The main idea is to derive the principal components for each one and then to compare the produced matrices.

Suppose that we have two multivariate time series denoted $X_{n\times p}$ and $Q_{n\times p}$. Applying PCA on each one results in the matrices of component weights A_X and A_Q and variances Λ_X and Λ_Q respectively. All the following measures assume that the number of variables p is the same for all series. This is a rational assumption, since these series are usually generated by the same process within a specific application.

One of the earliest measures has been proposed by Krzanowski [23]. This measure (Eq.2) is applicable to time series, although originally it was not applied on such type of data. The proposed approach is to retain m principal components and compare the angles between all the combinations of the first m components of the two objects.

$$Sim_{PCA}(X,Q) = trace(A_X^T A_Q A_Q^T A_X) = \sum_{i=1}^{m}\sum_{j=1}^{m} cos^2\theta_{ij}, \quad 0 \leq Sim_{PCA} \leq m \tag{2}$$

where θ_{ij} is the angle between the i^{th} principal component of X and the j^{th} principal component of Q.

Johannesmeyer [15] modified the previous measure by weighting the angles with the corresponding variances as in Eq. 3.

$$S_{PCA}^{\lambda}(X,Q) = \sum_{i=1}^{m}\sum_{j=1}^{m}(\lambda_{X_i}\cdot\lambda_{Q_j}\cdot cos^2\theta_{ij})/\sum_{i=1}^{m}\lambda_{X_i}\cdot\lambda_{Q_j}, \quad 0 \leq S_{PCA}^{\lambda} \leq 1 \tag{3}$$

Yang & Shahabi [38] propose a similarity measure, Eros, which is based on the acute angles between the corresponding components from two objects X and Q (Eq. 4). Contrary to the previous measures, all components are retained from each object and their variances form a weight vector w. More specifically, the variances obtained from all the objects in a database are aggregated into one weight vector, which is updated when objects are inserted or removed from the database. Finally, the authors provide lower and upper bounds for this measure.

$$Eros(X,Q,w) = \sum_{i=1}^{p} w(i)\cdot|cos\theta_i|, \quad 0 \leq Eros \leq 1 \tag{4}$$

Li & Prabhakaran [25] propose a similarity measure for recognizing distinct motion patterns in motion streams in real time. This measure, which is called k Weighted Angular Similarity (kWAS), can be obtained by applying singular value decomposition on the transformed datasets, $X^T X$ and $Q^T Q$, and retaining the first m components. kWAS is based on the acute angles between the corresponding components weighted by the corresponding eigenvalues (Eq. 5).

$$\Psi(X,Q) = \frac{1}{2}\sum_{i=1}^{m}((\sigma_i/\sum_{i=1}^{n}\sigma_i + \lambda_i/\sum_{i=1}^{n}\lambda_i)|u_i\cdot v_i|), \quad 0 \leq \Psi(X,Q) \leq 1 \tag{5}$$

where σ_i and λ_i are the eigenvalues corresponding to the i^{th} eigenvectors u_i and v_i of matrices $X^T X$ and $Q^T Q$. When the original datasets are mean centered, the above procedure is equivalent to applying PCA on the original data. The eigenvectors u_i and v_i are the corresponding principal components, while the eigenvalue-based weight in Eq. 5 is equal to the one obtained, if σ_i and λ_i are replaced by the variances of the corresponding components. The absolute value implies that the cosine of the acute angles is computed.

Singhal & Seborg [32] extend Johannesmeyer's [15] measure by incorporating an extra term, which expresses the distance between the original values of the two objects. This term is based on Mahalanobis distance and on the properties of the Gaussian distribution.

Another measure that incorporates the distance between the original values of two objects has been proposed by Otey & Parthasarathy [28]. The authors define a distance measure in terms of three dissimilarity functions that take into account the differences among the original values, the angles between the corresponding components and the difference in variances. For the first term, the authors propose to use either the Euclidean or the Mahalanobis distance, whereas the second term is defined as the summation of the acute angles between the corresponding components, given that all components are retained. The third term accounts for the differences in the distributions of the variance over the derived components and is based on the symmetric relative entropy [9].

In the context of Statistical Process Control, Kano et al. [19] propose a distance measure for the purpose of monitoring processes and identifying deviations from normal operating conditions. This measure is based on the Karhunen-Loeve expansion, which is mathematically equivalent to PCA. However, it involves applying eigenvalue decomposition twice during its calculation, which is the most computationally expensive part.

3 Proposed Approach

In this paper, we propose a novel approach in multivariate time series similarity search that is based on Principal Component Analysis. The main difference to other proposed methods is that it does not require applying PCA on the query object. Remember that an object is a multivariate time series that is expressed in the form of a matrix.

More specifically, PCA is applied on each object $X_{n \times p}$ of a database and the derived matrix of component weights $A_{p \times m}$ is stored (where m is the number of the retained components). Although this task is computationally expensive, it is performed only once during the preprocessing phase.

When a query object arrives, the objective is to identify the most similar object in the database. We propose a distance measure that relates to the Squared Prediction Error (SPE), a well-known statistic in Multivariate Statistical Process Control [22].

In particular, each time instance q_i of a query object $Q_{v \times p}$ is projected on the plane derived by PCA and its new coordinates (q'_i) are obtained (Eq. 6).

$$q'_i = q_i \cdot A, \quad i = 1, 2, \ldots v \tag{6}$$

In order to determine the error that this projection introduces to the new values, we need to calculate the predicted values ($\hat{q}_i$) of q_i (Eq. 7).

$$\hat{q}_i = q'_i \cdot A^T, \quad i = 1, 2, \ldots v \tag{7}$$

SPE is the sum of the squared differences between the original and the predicted values, and represents the squared perpendicular distance of a time instance from the plane (Eq. 8).

$$SPE_i = \sum_{j=1}^{p} (q_{ij} - \hat{q}_{ij})^2, \quad i = 1, 2, \ldots v \tag{8}$$

This measure can be extended in order to incorporate all time instances of the query object $Q_{v \times p}$ (Eq. 9). We call this new distance measure SPEdist (Squared Prediction Error Distance).

$$SPE_{dist}(X, Q) = \sum_{i=1}^{v} \sum_{j=1}^{p} (q_{ij} - \hat{q}_{ij})^2 \tag{9}$$

SPE is particularly useful within statistical process control because it is very sensitive to outliers, and thus, it can efficiently identify possible deviations from the normal operating conditions of a process. However, this sensitivity may be problematic in other applications that require more robust measures. Therefore, we propose a variation of SPEdist, that utilizes the absolute differences between the original and the predicted values (Eq. 10). We call this measure APEdist (Absolute Prediction Error Distance).

$$APE_{dist}(X, Q) = \sum_{i=1}^{v} \sum_{j=1}^{p} |q_{ij} - \hat{q}_{ij}| \tag{10}$$

The main concept is that, the most similar object in a database is defined to be the one, whose principal components describe more adequately the query object with respect to the reconstruction error. A similar approach can be found in the work of Barbic et al. [3], who propose a technique for the purpose of segmenting motion capture data into distinct motions. However, the authors utilize the squared error of the projected values and not the predicted values, as we propose in our work. Moreover, they focus on an application that involves one multivariate time series, which should be segmented.

As it was mentioned earlier, the proposed approach does not apply the computationally expensive PCA technique on the query object. Moreover, we provide a method that further speeds up the calculation of APEdist, hopefully, without substantially affecting the quality of similarity search. This method involves applying a dimensionality reduction technique on each one of the time series that form the query object, as a pre-processing step. The proposed technique is the Piecewise Aggregate Approximation (PAA) that was introduced independently by Keogh et al. [21] and, Yi & Faloutsos [39]. PAA is a well-known representation in the data mining community that can be extremely fast to compute. This technique segments a time series of length n into N consecutive sections of equal-width and calculates the corresponding mean for each one. The series of these means is the new representation of the original series. According to this approach, a query object that consists of p time series of length n is transformed to an object of p time series of length N. Under this transformation, we only need a fraction (N/n) of the required calculations in order to compute APEdist. Equivalently, the required calculations will be executed n/N times faster than the original ones. The consequence of this method in the quality of similarity search depends mainly on the quality of PAA representation within a specific dataset. Intuitively, APEdist is computed on a set of time instances, which may be considered as representatives of the original ones.

In general, our approach can be applied on data types other than time series. For example, suppose that we have customer data, such as age, income, gender, from several stores. Each store is represented by a matrix whose rows correspond to customers and columns correspond to their attributes. The objective is to identify similar stores with respect to their customer profiles. The PCA representation is based on the covariance matrix, which is independent of the order of the corresponding rows (time instances), and thus, the time dimension is ignored under the proposed representation.

In this paper, we focus on time series because this type of data is generated at high rates and is of high dimensionality. Our approach has two advantages. First, PCA-based representation dramatically reduces the size of the database while retaining most of the important information present in the original data. Second, the proposed distance measure does not require applying the computationally expensive PCA technique on the query.

4 Experimental Methodology

The experiments are conducted on three real-world datasets and one synthetically created dataset used extensively in the literature and described in Section 4.1. Section 4.2 presents the evaluation methods and Section 4.3 discusses the rival measures along with their corresponding settings.

4.1 Datasets

The first dataset relates to Australian Sign Language (AUSLAN), which contains sensor data gathered from 22 sensors placed on the hands (gloves) of a native AUS-

LAN speaker. The objective is the identification of a distinct sign. There are 95 distinct signs, each one performed 27 times. In total, there are 2,565 signs in the dataset. More technical information can be found in [17].

The second dataset, HUMAN GAIT, involves the task of identifying a person at a distance. Data are captured using a Vicon 3D motion capture system, which generates 66 values at each time instance. 15 persons participated in this experiment and were required to walk in 3 sessions, at 4 different speeds, 3 times for each speed. In total, there are 540 walk sequences. More technical information can be found in [33].

The third dataset relates to EEG (electroencephalography) data that arises from a large study to examine EEG correlates of genetic predisposition to alcoholism [4]. It contains measurements from 64 electrodes placed on the scalp and sampled at 256 Hz (3.9-msec epoch) for 1 second. The experiments were conducted on 10 alcoholic and 10 control subjects. Each subject was exposed to 3 different stimuli, 10 times for each one. This dataset is provided in the form of a train and a test set, both consisting of 600 EEG's. The test data was gathered from the same subjects as with the training data, but with 10 out-of-sample runs per subject per paradigm.

Finally, the transient classification benchmark (TRACE) is a synthetic dataset designed to simulate instrumentation failures in a nuclear power plant [31]. There are 4 process variables, which generate 16 different operating states, according to their co-evolvement through time. There is an additional variable, which initially takes on the value of 0, until the start of the transient occurs and its value changes to 1. We retain only that part of data, where the transient is present. For each state, there are 100 examples. The dataset is separated into train and test sets each consisting of 50 examples per state.

Table 1 summarizes the profile of the datasets.

Table 1 Description of Datasets

DATASET	# of variables	mean length	# of classes	size of class	size of dataset
AUSLAN	22	57	95	27	2565
HUMAN GAIT	66	133	15	36	540
EEG	64	256	2	600	1200
TRACE	4	250	16	100	1600

4.2 Evaluation Methods

In order to evaluate the performance of the proposed approach, we conducted several experiments in three phases.

First, we perform one-nearest neighbor classification (1-NN) and evaluate it by means of classification error rate. We use 9-fold cross validation for AUSLAN and

HUMAN GAIT datasets taking into account all the characteristics of the experiments, while creating the subsets. The observed differences in the error rates among the various methods were statistically tested. Due to the small number of subsets and to the violation of normality assumption in some cases, Wilcoxon Signed-Rank tests were performed at 5% significance level. For the EEG and TRACE datasets, we use the existing train and test sets.

Second, we perform leave-one-out k-NN similarity search and evaluate it by plotting the recall-precision graph [14]. In particular, every object in the dataset is considered as a query. Then the r most similar objects are retrieved, where r is the smallest number of objects that should be retrieved in order to obtain k objects of the same class with the query ($1 \leq k \leq$ size_of_class-1). The precision and recall pairs corresponding to the values of k are calculated. Finally, the average values of precision and recall are computed for the whole dataset. Precision is defined as the proportion of retrieved objects that are relevant to the query, whereas Recall is defined as the proportion of relevant objects that are retrieved relative to the total number of relevant objects in the dataset. In these experiments, the training and testing datasets of EEG and TRACE are merged.

Third, we evaluate the trade-off between classification accuracy and speed of calculating the proposed measure APEdist by applying 1-NN classification on objects that have been pre-processed as described in Section 3.

All the necessary codes and experiments were developed in MATLAB, whereas the statistical analysis was performed in SPSS.

4.3 Rival Measures

The similarity measures that were tested on our experiments are SimPCA, S^{λ}_{PCA}, Eros, kWAS, SPEdist, and APEdist. We choose to omit the results for SPEdist, because it performed similarly or slightly worse than APEdist in most cases. For comparison reasons, we also included in the experiments the Euclidean distance. Since this measure requires datasets of equal number of time instances, we decided to apply linear interpolation on the original datasets and set the length of the time series equal to the corresponding mean length (Table 1). The transformed datasets were utilized only when Euclidean distance was applied. The rest of the measures we reviewed in Section 2.3 are not included in these experiments because they take into consideration the differences among the original values, whereas in our experiments, the measures are calculated on the mean centered values.

Regarding Eros, the weight vector w was computed by averaging the variances of each component across the objects of the training dataset and normalizing them so that $\sum w_i = 1$, for $i = 1, 2, \ldots p$. In [38] one can find alternative ways for computing the weight vector.

All other measures require determining the number of components m to be retained. For AUSLAN, HUMAN GAIT and EEG, we have conducted classification for consecutive values of m between 1 and 20. For $m = 20$, at least 99% of the total variation is retained for all objects in AUSLAN and HUMAN GAIT, whereas at least 90% of the total variation is retained for all objects in EEG. For TRACE, we have conducted classification for all possible values of m ($m = 1,2,3,4$). Precision-Recall graphs are plotted for the "best" value that it was observed in the classification experiments. In general, this value is different for each measure.

Principal Component Analysis is performed on the covariance matrices. For comparison reasons, the similarity measures kWAS, and APEdist were computed on the mean centered values.

5 Results

We provide and discuss the results of 1-NN classification for each dataset separately in Section 5.1. In particular, we present the classification error rates that the tested measures achieved across various values of m (the number of components retained), and we also report the m that corresponds to the lowest error rate for each measure. In Section 5.2, the results of performing leave-one-out k-NN similarity search are presented in precision-recall graphs for each dataset. Finally, in Section 5.3, we provide and discuss the effect APEdist with PAA has on the classification accuracy for various degrees of speed up.

5.1 1-NN Classification

In the following figures (Fig. 3 to Fig. 6), the classification error rates are presented graphically for various values of m (the number of components retained) for each dataset. For the first three datasets, we show the error rates up to that value of m beyond which the behavior of similarity measures does not change significantly. For the TRACE dataset, which has only four variables, we show error rates for values of m up to three. For Euclidean Distance (ED) and Eros the rates are constant across m.

Regarding the first three datasets (Fig. 3 to Fig. 5), we observe that all measures seem to achieve the lowest error rate, when only a few components are retained. Moreover, as the number of components is further increased, the improvement in error rates seems to be negligible. In AUSLAN (Fig. 3), the performance of APEdist and SimPCA deteriorates with the increase of m. Note that these two measures do not take into account the variance that each component explains, contrary to the other three PCA-based measures. A second observation is that the performance of APEdist is comparable, if not better, to the "best" measure in each one of the three datasets. Regarding TRACE, which consist of only 4 variables, ED achieves considerably lower error rates than any other measure (Fig. 6).

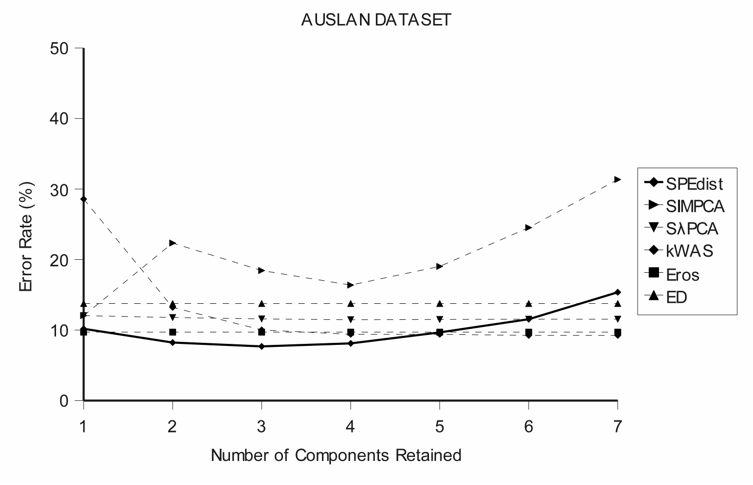

Fig. 3 1-NN Classification Error Rates (AUSLAN dataset)

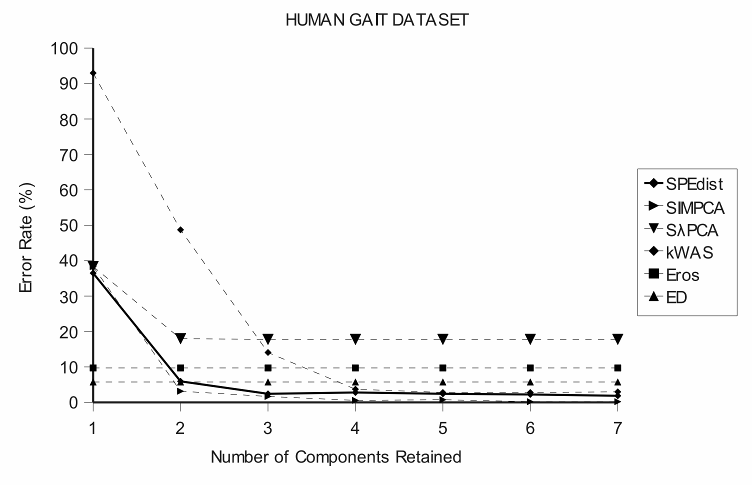

Fig. 4 1-NN Classification Error Rates (HUMAN GAIT dataset)

In Table 2, the lowest classification error rates are presented along with the corresponding number of the retained components. First, we will compare similarity/distance measures with respect to each dataset separately.

In AUSLAN, APEdist produces the lowest classification error rate. Statistically testing the differences across the specific subsets, APEdist produces better results than all measures ($p < 0.05$).

Regarding HUMAN GAIT, SimPCA, Eros, kWAS and APEdist seem to provide the best results. Statistically testing their differences across the specific subsets, Sim-

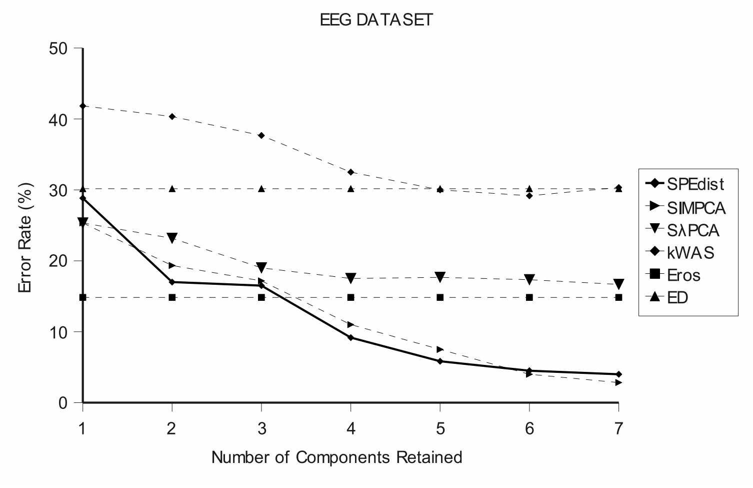

Fig. 5 1-NN Classification Error Rates (EEG dataset)

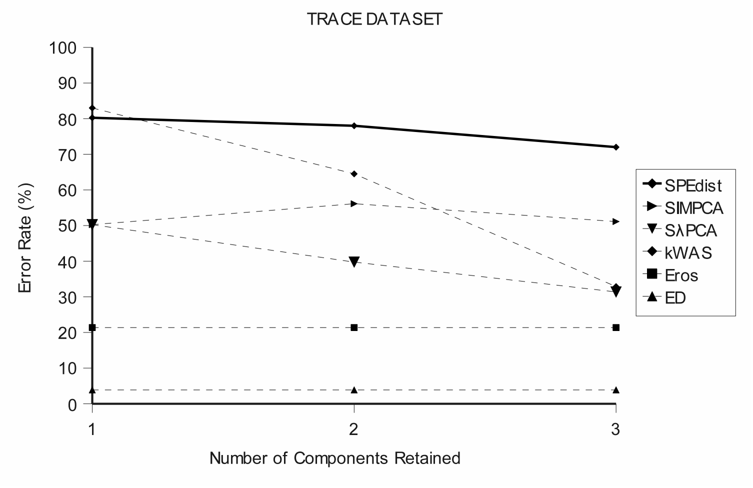

Fig. 6 1-NN Classification Error Rates (TRACE dataset)

PCA produces better results than all ($p < 0.05$), whereas the performances of Eros, kWAS and APEdist are statistically similar ($p > 0.05$).

For EEG, SimPCA and APEdist seem to provide considerably better results than other measures, with classification error rates of 0.00% and 1.83% respectively, when the next best performing measure, Eros, has a classification error rate of 14.83%.

Finally, for TRACE that consists of only 4 variables, Euclidean distance, a non-PCA-based measure, performs essentially better than all measures with 3.9% classification error rate. The next best performing measures are Eros and kWAS with classification error rates of 21.38% and 21.88% respectively.

Table 2 Classification Error Rates (%) [Numbers in parentheses indicate the number of principal components retained. Lack of number indicates measures that exploit all components]

Measure	ASL	HG	EEG	TRC
ED	13.76	5.74	30.17	**3.88**
SimPCA	(1) 12.05	(8) **0.00**	(14) **0.00**	(1) 50.25
SλPCA	(4) 11.46	(3) 17.78	(10) 16.50	(3) 31.38
Eros	9.71	2.96	14.83	21.38
kWAS	(6) 9.24	(12) 2.59	(17) 25.33	(4) 21.88
APEdist	(3) **7.68**	(7) 1.85	(14) 1.83	(3) 72.00

5.2 k-NN Similarity Search

In the following figures, the precision-recall graphs are presented for each dataset separately. The number of retained components is set equal to the one for which the corresponding measure provided the lowest classification error rates (Table 2). Regarding AUSLAN (Fig. 7), all measures seem to perform similarly to each other and better than the Euclidean distance. APEdist provides better results than all, however the differences can not be considered significant.

In HUMAN GAIT (Fig. 8) and EEG (Fig.9), however, SimPCA and APEdist perform better than all. As mentioned in the previous section, these two measures do not take into consideration the explained variance of the retained components. This fact may imply that for these specific datasets, the variance information may not be significant. On the other hand, in AUSLAN, where this information may be important, APEdist provides comparable results to other measures.

In the final dataset, TRACE, Euclidean distance performs better for recall values up to 0.3, whereas kWAS performs better for greater recall values (Fig. 10). Compared to other measures, APEdist seems to improve its performance for recall values greater than 0.6.

5.3 Speeding up the calculation of APEdist

The idea is to apply PAA on each one of the time series that comprise the query object (see Section 3), in order to speed up the calculations of APEdist. We experimented with various degrees of dimensionality reduction by using PAA to retain 10%, 20%, and 30% of the original dimensions of the query object, thus, expecting a 10x, 5x, and 3.33x speed up of the calculations, respectively.

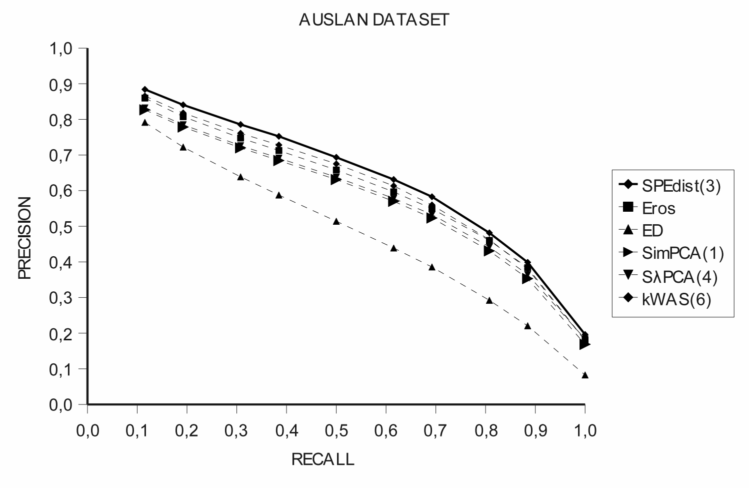

Fig. 7 Precision-Recall Graph for Various Measures (AUSLAN dataset)

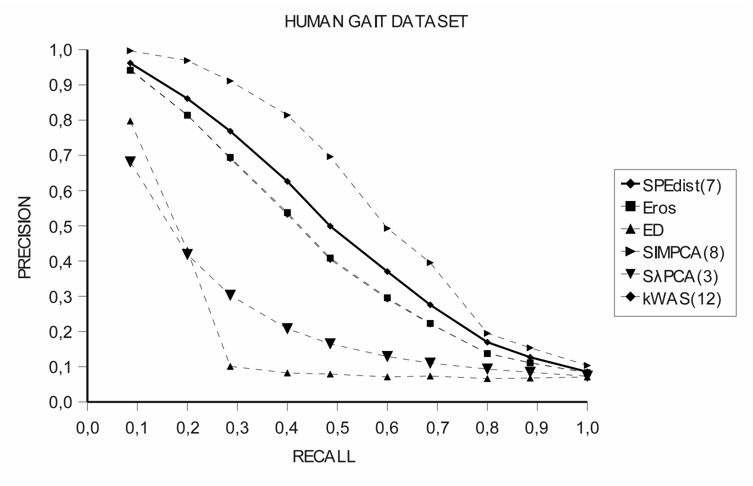

Fig. 8 Precision-Recall Graph for Various Measures (HUMAN GAIT dataset)

Table 3 presents the effect the speed up has on the classification error rate. The number of the retained components is different among datasets and is set equal to the optimal value obtained in Section 5.1 (Table 2).

As it was expected, the classification error rate increases as the speed up increases. Nevertheless, in all datasets, we are able to achieve similar classification error rates by doing at most 20% of the required calculations (a 5x speed up). More specifically, for AUSLAN, even a 5x speed up provides better results than rival measures (Table 2). Regarding HUMAN GAIT, a 10x speed up results into exactly the same classification error rate as the one observed when full calculations were ap-

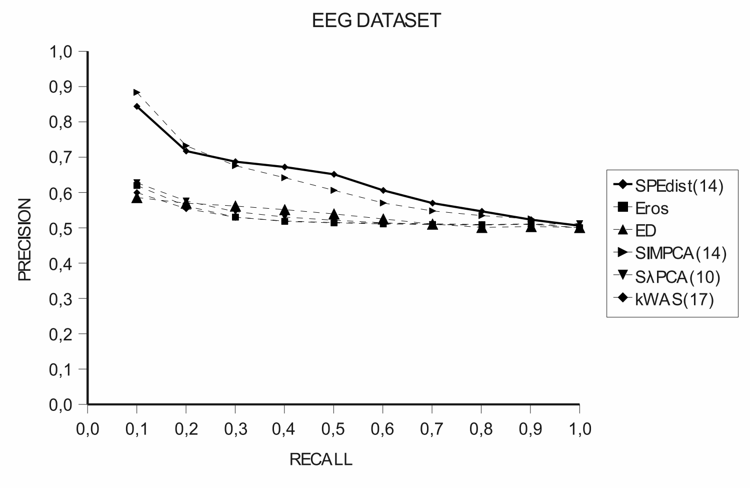

Fig. 9 Precision-Recall Graph for Various Measures (EEG dataset)

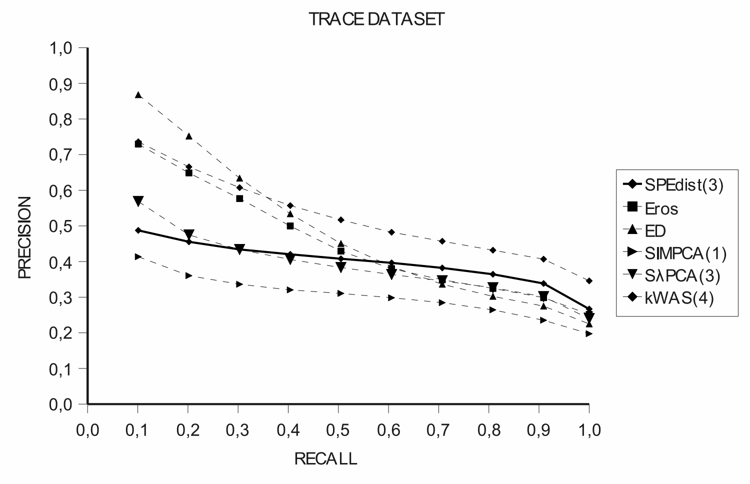

Fig. 10 Precision-Recall Graph for Various Measures (TRACE dataset)

plied. In EEG, although the error rates differ significantly for the various degrees of speed up, the 10x speed up provides lower error rate than rival measures (except from SimPCA). Regarding TRACE, a 5x speed up results into almost the same classification error rate as the one observed when full calculations were applied.

Table 3 1-NN Classification Error Rates for various degrees of dimensionality reduction on the query object

Percentage of Retained Dimensions Speed up	10% 10x	20% 5x	30% 3.33x	100% 1x
AUSLAN	10.80	8.50	8.27	7.68
HUMAN GAIT	1.85	1.85	1.85	1.85
EEG	8.00	3.67	2.33	1.83
TRACE	72.12	74.50	71.38	72.00

6 Conclusion

The main contribution of this paper is the introduction of a novel approach in multivariate time series similarity search for the purpose of improving the efficiency of data mining techniques without affecting the quality of the corresponding results. We investigate the usefulness of our approach, mainly in the context of query by content and 1-NN classification.

Experiments were conducted on four widely utilized datasets and various measures were tested with respect to 1-NN classification and precision/recall. There are three key observations with respect to the results of these experiments. First, there is no measure that can be clearly considered as the most appropriate one for any dataset. Second, in three datasets, our approach provided significantly better results than the Euclidean distance, whereas its performance was at least comparable to the four other PCA-based measures that were tested. Third, there is strong evidence that the application of the proposed approach can be accelerated with little cost in the quality of similarity search. In all datasets, one tenth up to one third of the required calculations was adequate in order to achieve similar results to the full computation case.

A secondary contribution of this paper is the review of several PCA-based similarity/distance measures that have been recently proposed from diverse fields, not necessarily within data mining context. A more general conclusion is that Principal Component Analysis has not been extensively explored in the context of similarity search in multivariate time series and hence, it has the potential to offer more in the Data Mining field.

Future work will focus on improving the speed up of the proposed approach during the pre-processing stage by exploiting the features of other dimensionality reduction techniques. We also intend to conduct experiments on more datasets in order to further validate our approach.

References

1. Agrawal, R., Faloutsos, C., Swami, A.: Efficient similarity search in sequence databases. In: Proc.4th Int. Conf. FODO, Evanston, IL, pp. 69–84, (1993).
2. Bakalov, P., Hadjieleftheriou, M., Keogh, E., Tsotras, V.J.: Efficient trajectory joins using symbolic representations. In: Proc. 6th Int. Conf. on Mobile data management, Ayia Napa, Cyprus, pp. 86–93, (2005).
3. Barbic, J., Safonova, A., Pan, J.Y., Faloutsos, C., Hodgins, J.K., Pollard, N.S.: Segmenting motion capture data into distinct behaviors. In: Proc. Graphics Interface Conf, London, Ontario, Canada, pp. 185–194, (2004).
4. Begleiter, H.: The UCI KDD Archive [http://kdd.ics.uci.edu]. Irvine, CA: University of California, Department of Information and Computer Science, (1999).
5. Buzan, D., Sclaroff, S., Kollios, G.: Extraction and clustering of motion trajectories in video. In: Proc. 17th ICPR, Boston, MA, vol. 2, pp. 521–524, (2004).
6. Cai, Y., Ng, R.: Indexing spatio-temporal trajectories with Chebyshev polynomials. In: Proc. ACM SIGMOD, Paris, France, pp. 599–610, (2004).
7. Chapman, L., Thornes, J.E.: The use of geographical information systems in climatology and meteorology. Progress in Physical Geography, 27(*3*), pp. 313–330, (2003).
8. Chen, L., Ozsu, M.T., Oria, V.: Robust and fast similarity search for moving object trajectories. In: Proc. ACM SIGMOD Int. Conf. on Management of Data, Baltimore, MD, pp. 491–502, (2005).
9. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley & Sons Inc., (1991).
10. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery: An Overview. In: Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, pp. 1–30, (1996).
11. Forbes, K., Fiume, E.: An Efficient Search Algorithm for Motion Data Using Weighted PCA. In: Proc. ACM SIGGRAPH/Eurographics Symp. on Computer Animation, Los Angeles, CA, pp. 67–76, (2005).
12. Gower, J.C.: Multivariate Analysis and Multidimensional Geometry. The Statistician, 17(*1*), pp. 13–28, (1967).
13. Gunopoulos, D., Das, G.: Time series similarity measures. Tutorial notes in: 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, Boston, MA, pp. 243–307, (2000).
14. Hand, D., Mannila, H., Smyth, P.: Principles of Data Mining. Cambridge, Mass., MIT Press, (2001).
15. Johannesmeyer, M.C.: Abnormal situation analysis using pattern recognition techniques and historical data. M.S. thesis, UCSB, Santa Barbara, CA, (1999).
16. Jolliffe, I.T.: Principal Component Analysis. New York, Springer, Chapter 1, (2004).
17. Kadous, M.W.: Temporal Classification: extending the classification paradigm to multivariate time series. Ph.D. Thesis, School of Computer Science and Engineering, University of New South Wales, (2002).
18. Kahveci, T., Singh, A., Gurel, A.: Similarity searching for multi-attribute sequences. In: Proc. 14th SSDBM, Edinburg, Scotland, pp. 175–184, (2002).
19. Kano, M., Nagao, K., Ohno, H., Hasebe, S., Hashimoto, I.: Dissimilarity of process data for statistical process monitoring. In: Proc. IFAC Symp. ADCHEM, Pisa, Italy, vol. I, pp. 231-236, (2000).
20. Kano, M., Nagao, K., Hasebe, S., Hashimoto, I., Ohno, H., Strauss, R., Bakshi, B.R.: Comparison of multivariate statistical process monitoring methods with applications to the Eastman challenge problem. Computers & Chemical Engineering, 26(*2*), pp. 161–174, (2002).
21. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. Knowledge and Information Systems, 3(*3*), pp. 263–286, (2001).
22. Kresta, J., MacGregor, J.F., Marlin, T.E.: Multivariate statistical monitoring of process operating performance. The Canadian Journal of Chemical Engineering, 69, pp. 35–47, (1991).

23. Krzanowski, W.: Between-groups comparison of Principal Components. JASA, 74(*367*), pp. 703–707, (1979).
24. Lee, S.L., Chun, S.J., Kim, D.H., Lee, J.H., Chung, C.W.: Similarity search for multidimensional data sequences. In: Proc. ICDE, San Diego, CA, pp. 599–608, (2000).
25. Li, C., Prabhakaran, B.: A similarity measure for motion stream segmentation and recognition. In: Proc.6th Int. Workshop MDM/KDD, Chicago, IL, pp. 89–94, (2005).
26. Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A symbolic representation of time series, with implications for streaming algorithms. In: Proc. 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, San Diego, CA, pp. 2–11, (2003).
27. Moeslund, T.B., Granum, E.: A survey of computer vision-based human motion capture. Computer Vision and Image Understanding, 81(*3*), pp. 231–268, (2001).
28. Otey, M.E., Parthasarathy, S.: A dissimilarity measure for comparing subsets of data: application to multivariate time series. In: Proc. ICDM Workshop on Temporal Data Mining, Houston, TX, (2005).
29. Quinlan, J.R.: C4.5 - Programs for machine learning. Morgan Kaufmann Publishers, San Mateo, (1993).
30. Ratanamahatana, C.A., Lin, J., Gunopulos, D., Keogh, E., Vlachos, M., Das, G.: Data Mining and Knowledge Discovery Handbook, chapter 51, Mining Time Series Data. Springer US, pp. 1069–1103, (2005).
31. Roverso, D.: Plant diagnostics by transient classification: the Aladdin approach. International Journal of Intelligent Systems, 17(*8*), pp. 767–790, (2002).
32. Singhal, A., Seborg, D.E.: Clustering multivariate time-series data. Journal of Chemometrics, 19(*8*), pp. 427–438, (2005).
33. Tanawongsuwan, R., Bobick, A.: Performance analysis of time-distance gait parameters under different speeds. In: Proc. 4th Int. Conf. AVBPA, Guilford, UK, pp. 715–724, (2003).
34. Valera, M., Velastin, S.A.: Intelligent distributed surveillance systems: a review. In: IEE Proc. Vision Image and Signal Processing, 152(*2*), pp. 192–204, (2005).
35. Vapnik, V.: The nature of statistical learning theory. Springer, New York, (1995)
36. Vlachos, M., Hadjieleftheriou, M., Gunopoulos, D., Keogh, E.: Indexing multidimensional time-series with support for multiple disatance measures. In: Proc. 9th ACM SIGKDD, Washington, D.C., pp. 216–225, (2003).
37. Vlachos, M., Hadjieleftheriou, M., Gunopoulos, D., Keogh, E.: Indexing multidimensional time-series. VLDB Journal, 15(*1*), pp. 1–20, (2006).
38. Yang, K., Shahabi, C.: A PCA-based similarity measure for multivariate time series. In: Proc. 2nd ACM MMDB, Washington, D.C., pp. 65–74, (2004).
39. Yi, B.K., Faloutsos, C.: Fast time sequence indexing for arbitrary Lp Norms. In: Proc. VLDB-2000: Twenty-Sixth International Conference on Very Large Databases, Cairo, Egypt, (2000).