

# Unsupervised Discovery of Abnormal Activity Occurrences in Multi-dimensional Time Series, with Applications in Wearable Systems

Alireza Vahdatpour, Majid Sarrafzadeh  
Computer Science Department  
University of California, Los Angeles  
{alireza, majid}@cs.ucla.edu

## Abstract

We present a method for unsupervised discovery of abnormal occurrences of activities in multi-dimensional time series data. Unsupervised activity discovery approaches differ from traditional supervised methods in that there is no requirement for manually labeled training datasets. In addition, they minimize the need for field experts' knowledge during the setup phases, which makes the deployment phase faster and simpler. We focus our attention on wearable computing systems and their applications in human activity monitoring for health care and medicine. The developed method constructs activity models in multi-dimensional time series based on the frequency and coincidence of activity perceptual primitives in single-dimensional time series data. We study the frequent variations exposed in human activity time series data and leverage physical attributes of the data to classify the activity primitives. A graph clustering approach is used to construct the frequent activity structures. Such structures are used to locate normal and abnormal occurrences of activities in time series. A method is presented to distinguish the abnormal activity occurrences from the normal occurrences. Two state-of-the-art wearable embedded systems (Smartcane and Smartshoe) are used to perform empirical evaluation of the developed methods.

## 1 Introduction

In this paper, we are interested in the discovery of the abnormalities in frequent activities performed by human subjects. Unsupervised discovery of actions and anomalies is specially in the interest of research community for wearable computing applications, where due to the high variety of activities and meaningful patterns, it is difficult (and even impossible) to define all proper activity patterns a priori. There are three main reasons why unsupervised activity discovery might be leveraged in different applications:

- Activity structure models for unconstrained and unknown environments are generally not known a priori. In order to design supervised methods for activity detection, training data has to be used in the training phase, which is not available in such scenarios. In addition, in many applications, the same activity may be performed differently by the subjects. Hence, a supervised method should be

separately tuned for all the subjects, by a field expert, which requires collecting sample training data from the subjects individually. In contrast, one advantage of unsupervised methods is that the discovery and detection can be performed without any training and algorithm tuning phase.

- Field expert knowledge is required for processing and analysis in many critical applications (especially in medicine), however, due to the limited availability of the field experts and high cost of their time, it is desired to accelerate the process. Transforming the low level input to mid level abstraction is one of the major reasons to use unsupervised methods in discovering events, activities, and abnormalities in time series data. Such transformation can be leveraged in visualization, storage, and high level evaluation of the data by the human experts.
- With the expansion of data generation and monitoring applications, unsupervised methods can be used concurrently with supervised methods for validation of the experts' knowledge and also to let them explore the data to expand their understanding of various phenomena in time series.

We have previously studied the problem of unsupervised discovery of frequent activities in multi-dimensional time series data in the presence of noise, scaling, and temporal variation [18]. It has been shown that although using unsupervised methods can lead to highly accurate discovery of frequent activities, depending on the application, supervised methods can achieve better accuracy, since they can be tuned for specific patterns, users, and application conditions. In addition, beside the lower accuracy of activity detection, the notion of activity abnormality is not addressed in unsupervised method.

It should be noted that unlike the most of the current work in unsupervised anomaly and outlier discovery

in time series, activity abnormality discovery is not the practice of locating the most *different* subsequence in a longer time series or a set of short time series. As it will be formally defined later, an abnormal occurrence of an activity is not any subsequence with maximal difference to the activity occurrences subsequences, and it has to follow the general patterns related to legitimate activity models.

It is relatively straight forward to define abnormal activities in supervised methods. As an example, a field expert can define possible abnormalities he/she anticipates to happen during an activity, and based on the defined patterns and features such abnormalities can be detected. For example, it may be defined that if the pressure on toes becomes less than a certain threshold during the walking activity, the activity occurrence is considered as abnormal. However, in an unsupervised activity discovery method, since the action patterns are unknown a priori, the notion of abnormalities is also not clear. In comparison to detection of predefined abnormalities, unsupervised abnormality discovery in activities is interesting in wearable and personalized devices for two main reasons:

- Since activity structures are not known in unconstrained environments, abnormality is also not defined for them. In addition, for each subject, abnormal activity is defined based on the specific normal behavior of the same subject. For example, people tend to walk differently to the point that [6] has used the gait characteristics of people as a method for authentication. Consequently, while a pattern may show abnormality for one subjects walking, it can be the regular gait model of another subject (hence, an application for fall detection should not consider it as a dangerous situation for the subject). Such abnormalities are distinguishable by considering the history of subjects activities and their temporal variation, and cannot be defined a priori by field experts, unless the methods are customized for each subject specifically. On the other hand, the unsupervised abnormality method we develop in this study uses the overall history and trends in activities to distinguish abnormal and normal activities.
- Supervised abnormality detection helps field experts accelerate the data analysis process, especially in very large databases. Experts have to usually setup the abnormality models and determine the patterns for regular and irregular actions, either from a prior knowledge or by using training data sets and the abnormality detection method is then applied to the data to extract interesting pat-

terns. However, a unique advantage of using unsupervised methods is to extend the current knowledge of the experts. More specifically, with the increasing trend in use of medical data collection systems and electronic health record systems, there is the potential to study the variation of known patterns and discover abnormal patterns that were not noticed by experts before. As an example, consider the scenario where the vital signs of numerous subjects performing daily tasks is collected, and based on the activity models which are constructed dynamically, abnormal subjects are discovered.

During the recent years several studies have been conducted to discover both activities and anomalies (discords) in time series. We focus our study on wearable sensing systems that has two main characteristics:

- With the advances in the sensing technology, modern wearable systems usually consist of numerous sensing nodes, and hence generate a set of multi-dimensional time series. Depending on the location and the type of the sensors, most of the actions are projected on a subset of the sensors, and any data analysis method should consider all the time series together to extract information from the data.
- Wearable sensing systems are mostly used to monitor human motions, which naturally show high variability and uncertainty. In addition, since sensing nodes are worn on the body, they are posed to various noise and external forces. As a result, several preprocessing and optimizations should be done on the data and the algorithms respectively, to improve the applicability of well-known time series analysis methods to this application domain.

We present our method for unsupervised discovery of abnormalities along the description the method for activity discovery in multi-dimensional time series in this study. In order to demonstrate the applicability of the methods in real applications, we apply them to the data collected from two state-of-the-art wearable devices, Smartcane [19] and Smartshoe [5]. The structure of the paper is as of the following: in Section 2 we briefly review the related work in the literature and highlight the difference between this study and the state-of-the-art literature in the field. In Section 3 notations and the problem is formally defined. Section 4 introduces the overall approach. In Section 5, we introduce the unsupervised method for activity model extraction and detection of normal and abnormal activity occurrences in multi-dimensional time series. To present the empirical evaluation of the methods, we first introduce two

main testbed systems we used for experiments in Section 6, and then the experimental results will be presented. Finally, we conclude the paper and present the future work in Section 7.

## 2 Background and related work

Design and development of methods for locating anomalies in time series data have been studied widely in the last years. The main purpose of anomaly detection in time series is to locate events and time instances where irregular actions or patterns have happened. Depending on the specifications of different applications, several definitions for anomaly have been proposed in the literature. As an example, [7] defines *discord* in time series to be subsequences (with predefined length) that are maximally different to all the rest of the time series subsequences. The authors suggest a method based on SAX [10] symbolization technique to locate such subsequences. They also present the applications of their discord detection method in analyzing electrocardiogram (ECG) in [9]. A number of other studies address different notions of anomaly in time series. [11] has developed an on-line algorithm for detecting *novelty* in temporal sequences. [3] has defined anomaly as the most infrequent time series pattern and proposed two algorithms for detecting them. In addition, [8] defines a pattern to be *surprising* if the frequency of its occurrences differs substantially from what is expected according to the previous experience. The study takes advantage of linear characteristics of suffix trees to extract such patterns from long time series.

In this study, we are considering the problem of unsupervised discovery of anomalies in activity occurrences discovered in a time series. We consider two main differences between this work and the related work in the literature. First, we are interested in locating abnormalities in frequent activities, meaning that in contrast to previous methods which consider any irregular subsequence as a potential anomaly, we restrict the anomaly candidates to subsequences which are sufficiently similar to the activity model discovered in the time series. Second, due to our applications, we focus on multi-dimensional time series (aka., multivariate time series), and the projection of abnormalities and activities among a set of related time series. Finally, it is worth to mention that unlike some related studies (e.g. [8]), this study doesn't leverage a priori information about the sequences.

Unsupervised discovery of activities has got attention from research community in the recent years. Studies in [14, 18, 13] have explored the use of time series motif detection algorithms in discovering frequent activity occurrences in multi-dimensional time series. Time

series motifs are approximately repeated subsequences in a longer time series data [4]. Furthermore, [13] has shown the deficiency of directly applying motif detection algorithms to time series data collected from wearable sensing systems, and has proposed an approach for efficient discovery of meaningful actions from sensor data representing human activity. In this paper, we apply a new method to single-dimensional time series in order to discover perceptual primitives of activities in the time series and leverage physical attributes of data to cluster these primitives.

## 3 Problem Definition

In this section, first we demonstrate an example of abnormality in frequent activities using a synthetic set of time series and then present the problem formulation.

**3.1 Illustrative example** Since the notion of abnormal activity is broad and undefined, first we depict a synthetic example in Figure 1 to illustrate the general concept of abnormal activity we study in this paper. Assume that time  $t_1$  contains a normal occurrence of an activity which is projected in all four dimensions of the time series. Time instances  $t_2$  and  $t_3$  include two abnormal activity occurrences. At time  $t_2$  similar patterns are projected on the time series, however, the high variation in temporal appearance of the patterns in dimensions denotes an abnormality in the activity occurrence. The activity occurrence at time  $t_3$  is considered an abnormality since the rectangular pattern occurred in second time series has high variation comparing to the pattern in  $t_1$ . Finally, at time  $t_4$ , two of the patterns projected on time series differ than the patterns of  $t_1$ , while two others have similarity. Due to the high variation of the most of the patterns at this time instance, it is unclear whether such occurrence should be considered an abnormality occurrence of the same activity occurred in  $t_1$  or the patterns denote an inherently different activity in the time series. A reasonable solution considers the application and the history of the activity occurrences in the long time series to decide in such situation.

Next we present a formal definition for the problem of finding abnormal activities in multi-dimensional time series, which is applicable to monitoring systems with multiple sensors.

**3.2 Problem formulation** We start by defining the notations used throughout the paper. Note that we avoid repeating the definition of well-known terms such as time series, subsequence, euclidean distance, etc. and refer the reader to numerous related papers for exact definition of such terms.

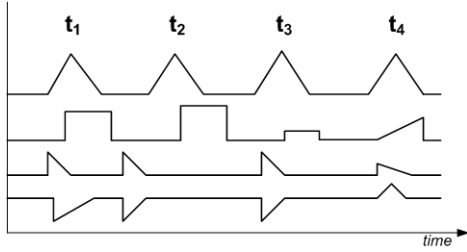


Figure 1: A synthetic set of time series and occurrences of an activity

DEFINITION 3.1. *Multi-dimensional time series:* a multi-dimensional time series  $M = \{T_1, \dots, T_m\}$  consists of  $m$  time series with equal length ( $n$ ).

DEFINITION 3.2. *Multi-dimensional subsequence:* a multi-dimensional subsequence  $S = \{s_1, \dots, s_p\}$  consists of  $p$  subsequences, where there exist at least one subsequence  $s_i$ ,  $1 \leq i \leq p$  such that  $s_i$  has temporal overlap with all other subsequences in  $S$ .

Note that the members of  $S$  are scattered among any subset of dimensions of a multi-dimensional time series  $M$  and they can be different in the shape and length.

DEFINITION 3.3. *Activity:* an activity is a multi-dimensional subsequence  $A = \{a_1, \dots, a_x\}$  with recurrent occurrences in the multi-dimensional time series. The occurrences of the activity  $A$  are noted as  $A^1 \dots A^k$ , when  $A$  has  $k$  occurrences.

DEFINITION 3.4. *Abnormal activity occurrence:* a multi-dimensional subsequence  $R = \{r_1, \dots, r_y\}$  is an abnormal occurrence of the activity  $A$  if:

$$\begin{cases} \alpha < \frac{|R \cap A|}{|A|} < 1 \\ \text{occurrences}(R) < \beta \cdot \text{occurrences}(A) \end{cases}$$

Here,  $A$  is a multi-dimensional subsequence representing an activity.  $|R \cap A|$  is the total number of similar subsequences in  $A$  and  $R$ .  $\text{occurrences}(S)$  denotes the total number of occurrences of the multi-dimensional subsequence  $S$  in a time series.  $\alpha < 1$  is the abnormality similarity constant associated with the application, and denotes the minimum similarity that should exist between the normal and abnormal occurrences of an activity. In addition,  $\beta < 1$  is the abnormality frequency constant and is a factor to decide whether a multi-dimensional subsequence is an independent activity or is an abnormal occurrence of another activity. The last two constants are input parameters for our methods, and are discussed later in the paper.

It is clear from the definition that in order to discover abnormalities, it is first desired to discover the common activity occurred in the time series and locate its occurrences. In the next section, we first review our overall approach, and introduce the main phases of the unsupervised activity and abnormality discovery method. Each phase is described in details in the following sections.

## 4 Overall approach

Figure 2 depicts the general overview of our method for abnormal and normal activity occurrence discovery in multi-dimensional time series. In the first phase, we focus on each dimension of time series separately and extract recurrent patterns and meaningful subsequences from them. These patterns are considered as primitives for activities which are projected in more than one dimensions of a set of time series. We discuss two approaches for this phase; the general time series motif detection, and an activity primitive discovery method based on physical attributes of the data, specifically designed for wearable sensing systems. The single-dimensional activity primitives extracted in this phase are analyzed to discover the structure of the dominant multi-dimensional activities in the set of time series (phase 2). We used a graph clustering approach to construct the activity structure. Thereafter, using the activity structure, all the activity occurrences (both normal and abnormal) in the time series are extracted from the time series (phase 3). Finally, the abnormal activity occurrences are discriminated from the normal occurrences, by leveraging the information from all the previous phases (phase 4).

## 5 Unsupervised activity and abnormality discovery

We study the use of two methods in extracting important subsequences from each dimension of the multi-dimensional time series. Motif detection in time series is a general approach to detect recurrent subsequences in a set of subsequences or in long time series. Based on our previous work in [18], we first study the use of motif detection in activity primitive extraction. While we show the performance of applying this method in the results section, we also developed a domain specific approach for wearable motion sensing systems, that is presented after discussing the motif discovery approach shortcomings in activity discovery applications.

**5.1 Frequent patterns in single-dimensional time series** The first step in unsupervised activity discovery is to extract recurrent patterns in each dimension of the time series. We consider such recurrent patterns

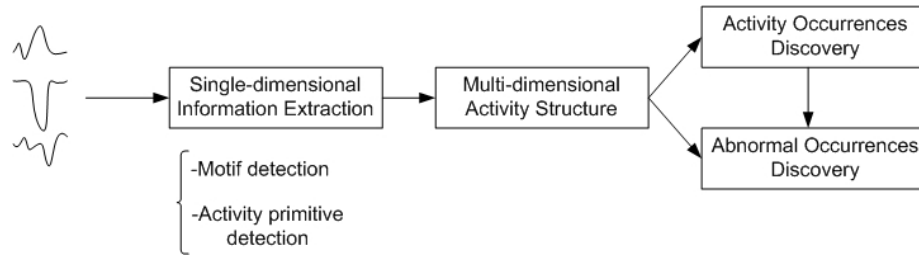


Figure 2: The overall approach for activity discovery and abnormality detection in multi-dimensional time series

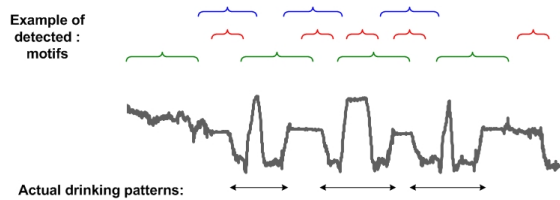


Figure 3: The acceleration of the wrist, in three consecutive occurrences of drinking activity

as the constrictive primitives of the activities that are projected on several dimensions of a set of time series. Motif detection is a common approach for discovering recurrent patterns in time series. Time series motifs are approximately repeated subsequences in a longer time series data [4]. Numerous studies have been done to make motif detection faster, more accurate, and applicable to more applications [16, 20, 17].

In this study we used the probabilistic method of [4] for motif discovery. While the same authors have also developed the *exact* method in [15], the efficiency of the probabilistic method is observed to be sufficient for our application requirements. The above method has been described in many related literature and here we avoid describing it. Applying the motif detection mechanism to the data collected from our wearable systems, we observed the following shortcomings:

- Motif detection algorithms are generally sensitive to the length of the motif subsequences and require the candidate length of the motif as an input parameter. This requires a priori knowledge of the activity primitives length, which is obviously an undesired situation for unsupervised activity discovery process. In addition, human activities tend to be done with variation in the pace and as a result their length and shape differ in each occurrence. Time series in Figure 3 depicts acceleration data, collected from an accelerometer worn on a wrist of a human subject while he repeated the same action of drinking water from a glass. It is clear that oc-

currences of the same activity differ in the length. Therefore, the efficiency of directly applying the motif detection algorithm to cluster them together as one motif will be decreased. Note that the effect of such variations can be reduced by using methods such as Dynamic Time Warping [2], however, the variation in the length of patterns dramatically increases the search space for motif subsequences and decreases the efficiency and the speed of the motif detection process significantly. The figure also depicts the output of applying the motif discovery on the time series with several subsequence lengths as input. Note how due to the sensitivity of the algorithm to the input motif length, different motifs are captured in each execution of the algorithm.

- In many applications, especially those with the purpose of human activity monitoring, the distinctive property of an activity is the deviation it causes in the sensor data comparing to its default or stable value. For example, in the acceleration data of Figure 3, all of the three drinking actions showing similar minimum and maximum values in their peak and valley patterns, since the final orientation of the wrist is similar in all occurrences of the activity. An indirect approach for applying motif detection is to cluster the subsequences that shows transitions between similar stable minimum and maximum values. However, changing the speed of the motion causes the slopes to vary, and hence the motif detection algorithm fails in clustering such short subsequences efficiently (or a high number of false positives will be detected along the valid motif subsequences).

Observing the above shortcomings of applying the general motif discovery method, we developed a technique considering the specific features of time series data in wearable systems, which mostly consists of motion and pressure sensing devices. Note that the study in [13] has designed a motif based approach to extract activity patterns in wearable applications, and evaluated its efficiency. As stated by the author, the mechanism

however relies on many parameters for accurate performance. In the next section, we describe our domain specific method for locating activity primitives in single dimensional time series.

**5.2 Activity primitive discovery in single-dimensional time series** As described earlier, in most of motion monitoring applications, acceleration sensors are the primary source of data. Normally several multi-dimensional accelerometer sensors are placed on body in order to comprehensively monitor the motions. Observing several applications of human motion monitoring systems, we noticed that human motions are projected on single dimensions of sensors as transitive patterns between two stable levels. Assuming the subsequence  $a_1..a_m$  to be an action primitive, the first and the last points in the subsequence ( $a_1$  and  $a_m$ ) are the ending point and the starting point of two stable regions (or regions with the sensor's default value) before and after the subsequence in the time series, which generally denote the situation before and after the action. For example, in accelerometer data, the first and the last values of an activity primitive represent the starting and final orientation of the sensor. Therefore, we define an activity primitive to be a subsequence between two consecutive stable regions (or default sensor value) in the time series. Such subsequences are mostly observed as bell shape patterns (with local minimum or maximum points) or linear slopes between two stable values in human actions, which depending on the speed and nature of the activity have different length and frequency. We classify these primitives based on the physical attribute they represent. For example, the noise free acceleration data can be converted to displacement, using the following equation on a subsequence of length  $m$ .

$$displacement = D \cdot \sum_{k=1}^m \sum_{i=1}^k |a_i - a_1|$$

where  $D$  is the calibration constant factor, which depends on the placement of the sensor and other parameters such as the frequency of data sampling. Algorithm 1 presents the method for extracting primitives of activity subsequences projected on single-dimensional time series. The algorithm extracts the patterns from the time series, classifies them based on the physical attribute of the data and according to the discretization cardinality, and assigns a symbol to them. It also assigns starting time to each symbol. There are variety of methods in literature regarding effective discretization, especially in time series. Here, in order to discretize the calculated *displacement* and assign symbols to it, we consider using the probability distribution of the calcu-

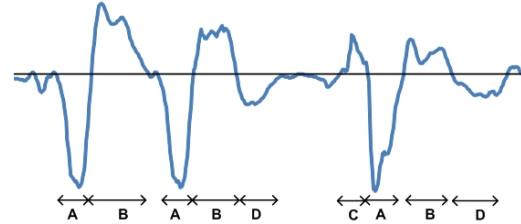


Figure 4: The result of applying algorithm 1 on the Smartcane's rotation data

lated *displacements* and assigning symbols to primitives such that primitives are classified fairly considering the variation of *displacement* in the application runtime. Figure 4 depicts the result of applying this algorithm on a sample time series, collected from Smartcane gyro sensor.

---

**Algorithm 1** Single-dimensional activity primitive extraction

---

```

Input: Time series  $T = t_1..t_n$ 
Output: List of activity primitives along with their
occurrences location  $p\_occurrence(1..j)$ ,  $start(1..j)$ 
 $i \leftarrow 1$ 
while true do
  while  $t_i = default\_value \mid t_i$  is stable region
  do
    increment  $i$ 
  end while
   $start(j) \leftarrow i$ 
  while  $t_i \neq default\_value \ \& \ t_i$  is not stable region
  do
     $d \leftarrow d + \sum_{i=1}^{start(j)} t_i - t_{start(j)}$ 
  end while
   $length(j) \leftarrow i - start(j)$ 
   $displacement(j) \leftarrow D \cdot d$ 
   $d \leftarrow 0$ 
  if  $i = n$  then
     $p\_occurrence(1..j) \leftarrow$ 
       $discretize(displacement(1..j), length(1..j))$ 
    return  $p\_occurrence(1..j), start(1..j)$ 
  end if
   $j \leftarrow j + 1$ 
end while

```

---

Note that in contrast to motif discovery method which we have used in [18] for detecting recurrent patterns in time series, this algorithm leverages domain specific attributes of data (here, being the motion sensing data, which used frequently in wearable systems). Regardless of the method used to discover recurrent activity patterns in the single-dimensional time series, the

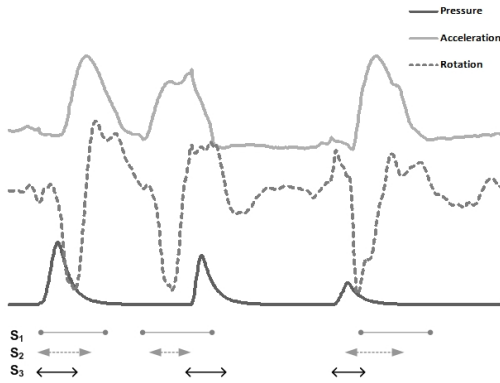


Figure 5: Temporal variation of activity primitives in Smartcane usage

output of this phase is passed to the multi-dimensional activity and abnormality detection algorithm, which is described in the next section.

**5.3 Activity structure in multi-dimensional time series** This section describes the method we developed for discovering the structure of the recurrent activities in multi-dimensional time series. Later, we extend this algorithm to detect abnormal activity occurrences based on the constructed model for normal activities. As mentioned earlier, the first phase of the algorithm is to extract the action primitives from single-dimensional time series. This phase was performed either via using a general approach (motif discovery in single dimensional time series) or via a domain specific method described in the last section. In the second phase, the activity primitives from single-dimensional time series are clustered together, and construct a model for recurrent activities in multi-dimensional time series.

As described earlier, an activity is a recurrent multi-dimensional subsequence. Figure 5, depicts an snapshot of three sensors data during the walking activity in Smartcane device. As it is clear, the activity occurrences are projected on the thee presented time series consistently. However, while in the first occurrence of the activity in Figure 5,  $s_1$ ,  $s_2$ , and  $s_3$  (primitives discovered in each dimension) are synchronized, in the second and third occurrences,  $s_2$  and  $s_3$  appear with temporal variation. While in many critical applications such as aerospace, it is crucial to capture such delays, in wearable monitoring applications, such temporal variations are common. Specifically, in monitoring the human motion data using acceleration and gyro sensor, such variations are common [13], and restricting the activity definition to a set of subsequences with exact temporal relation will result in missing many occurrences. Hence, the general notion of temporal overlap is used to define a set

of primitives as an activity structure. We use statistical information from the primitive occurrences in the time series to construct the structure of the most dominant activities in a multi-dimensional time series.

In order to construct the multi-dimensional activity structures, we first convert the list of discovered activity primitives into a weighted directed graph, so that a graph clustering mechanism can be applied to construct model of activities in multi-dimensional data. In the proposed graph, each vertex represents an activity primitive, and the weight on the vertices represents the number of occurrences of the primitive in the corresponding time series. The weight on each directed edge in the graph is calculated by the following equation:

$$e(i, j) = \frac{\text{coincidence}(i, j)}{\text{total\_occurrences}(j)}$$

The  $\text{coincidence}(i, j)$  denotes the number of times there is temporal overlap between occurrences of primitives  $i$  and  $j$ . As a result,  $e(i, j)$  will be at most 1, when all the occurrences of primitive  $j$  have overlap with occurrences of primitive  $i$ , and it is at least 0, when there is no overlap between occurrences of primitives  $i$  and  $j$ . Figure 6.a and 6.b depict an example of three time series from the Smartcane system, their discovered primitives, and the resulting primitive coincidence graph.

In the next phase, a graph clustering approach is used to construct multi-dimensional activities structure from the primitive coincidence graph. Note that according to the definition presented in Section 3.2, several activity structures may be discovered in a time series. In this study we limit ourselves to the activity structures with the highest number of occurrence in the time series and study abnormal occurrences in respect to them. In addition, according to the definition, eliminating a primitive from the set of primitives that resembles an activity, will still result in an activity model, which is a subset of the original activity structure. Here, we focus on extracting maximal sets representing activities (a maximal activity model is a multi-dimensional subsequence that adding a primitive to it doesn't construct another activity model with multiple occurrences in the time series). Algorithm 2 presents our method for constructing the activity structures from the coincidence graph. This method is similar to the clustering mechanisms proposed in [1]. Clustering the primitives starts with sorting the list of vertices based on the number of occurrences of the primitives. Then, the most frequent primitive is selected as a candidate core for activity structure, and the graph is searched for primitives with high occurrence correlation with the candidate core primitive. If there are several primitives with equal frequency, the one with the larger number of highly cor-

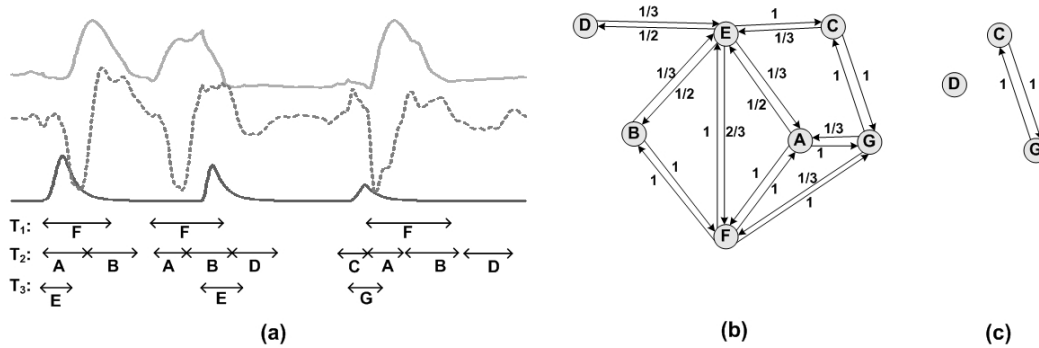


Figure 6: (a) Three time series from the Smartcane and the discovered activity primitives, (b) The resulting coincidence graph (c) The graph after eliminating activity primitives clustered in Algorithm 2

related neighbors is selected as the core primitive. The threshold for the correlation is set to  $1 - \beta$ , where  $\beta$  is the abnormality frequency constant. Upon construction of an activity by clustering correlated primitives, selected primitives are removed from the graph, and the algorithm continues for discovering new activity structures by selecting a new core activity primitive.

---

**Algorithm 2** Activity structure construction from the primitives coincidence graph

---

- 1: Input:  $G(V, E)$  the primitives coincidence graph,  $total\_occurrence(v)$  for all vertices and  $e(v_1, v_2)$  for all directed edges
  - 2: Output:  $S = \{S_i\}$  The set of most recurrent activity structures in the time series, (each  $S_i$  is an activity)
  - 3: Sort the vertices list in  $G$  in descending order of  $total\_occurrence$   
 {if some primitives have equal frequency, then the primitive with the highest number of highly correlated neighbors has the highest priority}
  - 4: **for** each vertex  $v_k$  in the sorted list of vertices **do**
  - 5:   Add  $v_k$  to activity  $S_i$
  - 6:   **for** all neighbors of  $v_k$ , if  $e(v_j, v_k) > 1 - \beta$  **do**
  - 7:     Add  $v_j$  to activity  $S_i$
  - 8:     remove  $v_j$  from the graph
  - 9:   **end for**
  - 10:    $i \leftarrow i + 1$
  - 11:   Update the sorted list of vertices
  - 12: **end for**
  - 13: **return**  $S = \{S_1, \dots, S_{i-1}\}$
- 

Figure 6.c illustrates the results of applying the algorithm to the coincidence graph of previous example in Figure 6.b, when the  $\beta$  is set to 0.4. As it is denoted in the figure, the activity structure extracted contains primitives A, B, F, and E (F is the activity core primitive, due to its high number of occurrences). Upon

eliminating these primitives from the graph, primitives C and G show high correlation, and can be considered as the second activity structure. Note that the effective construction of activity structures highly depends on the frequency of primitives, and in a real world application the frequency and coincidences are significantly higher ( $\beta$  will be set to less than 0.1 in real world applications). The depicted graph is formed from a set of short time series and is selected to be small to better illustrate the logic behind the algorithm.

**5.4 Abnormality discovery in activities** An abnormality is a multi-dimensional subsequence, which has at least one missing primitive comparing to occurrences of activity  $A$ , however, the number of missing primitives is not more than  $(1 - \alpha) \cdot |A|$ . Algorithm 3 presents our method for discovering such multi-dimensional subsequences. The input to the algorithm is the list of all discovered single-dimensional activity primitives in the time series, in addition to the discovered multi-dimensional activity structure. The algorithm first removes the primitive occurrences that are not in the activity structure from the list of primitives (lines 3-7). By scanning the remaining primitives in the list, the algorithm removes primitive occurrences that construct legitimate activity occurrences from the list (lines 9-14). Finally, another scanning is performed on the list of primitive occurrences and any multi-dimensional subsequence with abnormal activity occurrence characteristic of definition 3.4 is returned as output (lines 15-23).

The reason behind the two consecutive scanning iterations of the primitive occurrences list is illustrated in Figure 7. As it is clear in the figure, a primitive occurrence may construct more than one multi-dimensional subsequences, while only one of them being an activity occurrence. Hence, in the first iteration, all the activ-

---

**Algorithm 3** Abnormal activity occurrence discovery
 

---

- 1: Input: Activity structure  $S = \{s_1, \dots, s_m\}$  and list of single-dimensional primitives  $p\_occurrences(1..j)$  and their locations  $start\_time(1..j)$
  - 2: Output: Set of abnormal activity occurrences  $Abnormal = \{abnormal_1, \dots, abnormal_k\}$
  - 3: **for all**  $primitive$  in  $p\_occurrence$  **do**
  - 4:   **if**  $primitive \notin S$  **then**
  - 5:     Remove  $primitive$  from  $p\_occurrences$
  - 6:   **end if**
  - 7: **end for**
  - 8: Sort  $p\_occurrences$  based on  $start\_time$
  - 9: **for all**  $primitive$  in sorted array  $p\_occurrences$  **do**
  - 10:    $temp\_set = \{\text{all the primitive occurrences which have temporal overlap with } primitive\}$
  - 11:   **if**  $\frac{|temp\_set \cap S|}{|S|} = 1$  **then**
  - 12:     Remove  $temp\_set$  members from  $p\_occurrences$
  - 13:   **end if**
  - 14: **end for**
  - 15: **for all** Remained  $primitive$  in  $p\_occurrences$  **do**
  - 16:    $temp\_set = \{\text{all the primitive occurrences which have temporal overlap with } primitive\}$
  - 17:   **if**  $\frac{|temp\_set \cap S|}{|S|} > \alpha$  **then**
  - 18:      $abnormal_i = temp\_set$
  - 19:     Remove  $temp\_set$  members from  $p\_occurrences$
  - 20:     Increment  $i$
  - 21:   **end if**
  - 22: **end for**
  - 23: **return**  $Abnormal = \{abnormal_1, \dots, abnormal_i\}$
- 

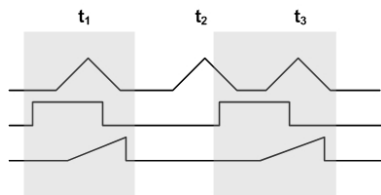


Figure 7: The primitive in  $t_2$  may be detected as abnormality if the activity occurrence in  $t_3$  is not eliminated

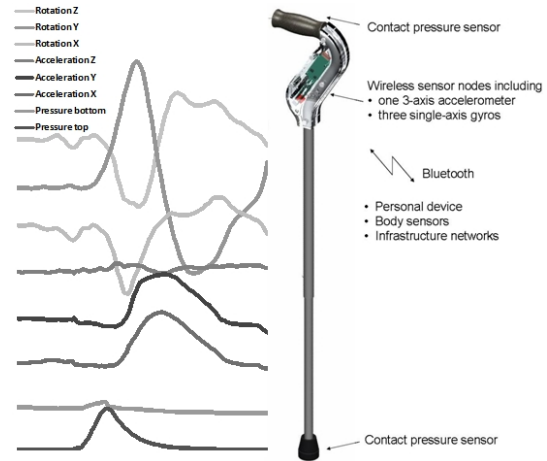


Figure 8: Smartcane system and sample data from walking activity

ity occurrences are discovered and their corresponding primitives are removed from the primitives list, in order to avoid false positives in detecting abnormal occurrences. In the example of Figure 7, while in the first iteration the patterns in  $t_2$  are candidate for abnormality, by eliminating the patterns of activity occurrence at time  $t_3$ , the similarity of patterns in  $t_2$  to the activity model (happened in  $t_1$ ) becomes less than  $\alpha$ , and hence it is not considered as an abnormal occurrence of the activity.

## 6 Empirical evaluation

In this section, we present the empirical evaluation of the presented approach for abnormal activity discovery in multi-dimensional time series. We used two wearable embedded systems for the experimentations. Next, we briefly overview these platforms.

**6.1 Smartcane** The Smartcane system is developed as a device to monitor and train elders and impairs in their assisted walking behavior. Several sensor systems, including a 3-axis accelerometer, a 3-axis gyro, and 2 force sensors are embedded in a normal cane, in order to let the physicians and care-givers completely evaluate and understand the subjects behavior and habits in using the cane [19]. Figure 8 depicts the general structure of the Smartcane system, the placement of the sensors, and a sample set of data collected during walking.

While several guidelines for proper use of cane are available (such as the maximum force that should be applied on the cane), there is no general criteria to discriminate proper usage of the cane. Therefore, the evaluation of the usage is personalized and experts

evaluate each subject behavior separately. On the other hand, early detection of situations that cause irregularity in the walking and cane usage is critical to avoid falls, which are the leading cause of injury and death in elderly. We applied our unsupervised method to discover abnormal occurrences of activities in order to accelerate detection of such conditions.

A great advantage of applying an unsupervised method to this application is that it eliminates training and tuning phase for each subject. Especially, as it was mentioned, the normal walking pattern is subject dependent, and in case of using supervised methods, training and tuning of the system has to be done for each subject individually. In addition, due to variations in users characteristics (e.g. weight), it is anticipated that the same users normal activity patterns vary by time. Hence, a supervised method has to be continuously readjusted.

**6.2 SmartShoe** The Smartshoe system [5] is a regular shoe, utilized with an electronic insole. The insole consists of several sensory circuits, including several pressure sensors to monitor plantar pressure, 3-axis accelerometers and a 3-axis gyro to monitor feet motion. Discovering temporal abnormality in walking patterns is of great interest for medical experts, since not only does it reveal conditions such as alcohol intoxication, drowsiness, or injury, but also walking anomaly is a sign of many critical health situations such as heart attack and stroke. While people walk differently, individual walking patterns also change based on the environmental characteristics. For example, the slope of the surface and the shoe specifications change the plantar pressure and feet motions significantly. Hence, detecting temporal abnormalities should be performed considering the normal activity occurrences in the same time and for the same person. The unsupervised method is therefore a suitable approach for this application, since not only doesn't it require predefining normal and abnormal patterns, but also it detects abnormal activity occurrences by comparing them to the most frequent activity occurrences in the temporal neighborhood. Figure 9 depicts the structure of the Smartshoe and a sample subset of times series captured during the walking activity.

**6.3 Experiments** To evaluate the accuracy of the normal and abnormal activity occurrences discovery, we ran several experiments with Smartcane and Smartshoe. Table 1 presents the result of applying the algorithm on the data collected from three subjects, while walking with the assistance of the Smartcane. We asked the users to exhibit abnormal walking behaviors infrequently during the test. It is clear that changing  $\alpha$ , the

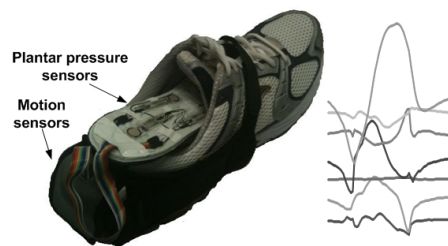


Figure 9: Smartshoe system and sample data from walking activity

Table 2: The result of normal and abnormal activity discovery in Smartshoe for walking and running activities

	Walking		Running	
$\alpha$	.6	.5	.6	.5
Normal	94%		82%	
Abnormal	15%	56%	66%	96%
False positives	2%	6%	13%	37%

abnormality similarity constant, impacts the total number of discovered abnormality occurrences. A greater  $\alpha$  denotes that discovered abnormal occurrences have higher similarity to the normal occurrences. On the other side, while a smaller similarity constant results in detecting more abnormal occurrences, it also increases the false positives in abnormality discovery, since some non-activity subsequences will be considered as abnormalities.

Table 2 presents the result of abnormal and normal activity occurrences discovery for two different activities, both done on the Smartshoe system. Note that in this table, *Normal*, *Abnormal*, and *False Positive* are calculated in a similar way to the results of Table 1. In the running activity, not only are the length of activity subsequences shorter, but also the variation of the patterns is higher, which results in less accuracy in activity occurrences discovery and increased number of false positives. However, the relatively higher impacts of steps and easily recognizable activity primitives in the time series results in easier discovery of abnormal occurrences, since even minor abnormalities cause visible variation in single-dimensional primitives.

Table 3 compares the results of applying the motif detection approach and the domain specific primitive extraction method (described in algorithm 1) in discovering activity occurrences. As it was discussed earlier, motif detection has been shown to be powerful in discovering the recurrent phenomenon in many application domains [15], however, the high variation of patterns in these wearable applications makes them an unsuitable area for motif detection.

Table 1: The result of normal and abnormal activity occurrence discovery in Smartcane

	Subject 1			Subject 2			Subject 3		
$\alpha$	.7	.6	.5	.7	.6	.5	.7	.6	.5
Normal ( $\frac{discovered}{total}$ )	89%			94%			83%		
Abnormal ( $\frac{discovered}{total}$ )	57%	85%	97%	42%	42%	86%	71%	71%	71%
False positives ( $\frac{false\ alarms}{total\ discovered\ abnormalities}$ )	4%	4%	7%	0%	0%	4%	5%	9%	9%

Table 3: The accuracy of using motif detection and primitive extraction in discovering activity occurrences

	Smartcane	Smartshoe
Motif detection	71%	82%
Primitive extraction	89%	91%

Table 4: Abnormal activity detection/discovery accuracy

Method	Abnormal activity detection	Requirement
Decision Tree	~ 98%	Training & setup for each user
Bayes Classifier	~ 95%	Manually labeled data of users
Unsupervised Clustering	~ 85%	No training

Table 4 contains comparison between our unsupervised abnormality discovery method and two well-known supervised techniques. The decision tree method uses thresholds defined by a field expert to discriminate normal and abnormal activity occurrences. The thresholds are specific for individuals, and are set during the setup phase. The method achieves almost perfect resolution in detecting abnormalities in the activities of the subject, if the personal setup phase is done carefully. We also used a Naive Bayes classifier in order to discriminate abnormal activities from normal occurrences. The classifier was trained with a manually labeled data set. While the method achieves high accuracy (~ 95%) in detecting improper usage, the accuracy highly relies on whether the subjects usage patterns are similar to the training data set or not. Our unsupervised method achieves less accuracy in discovering abnormalities, however, it doesn't require any training and setup phase. In addition, unlike the previous work in [13, 18], the method is not limited to a predefined subsequence length for activity primitives, and activity primitives are extracted dynamically from the time series.

## 7 Conclusion and future work

In this paper, we presented the design of an unsupervised technique for discovering abnormal and normal occurrences of activities in multi-dimensional time series data. After discussing several characteristics of abnormal occurrences of activities in multi-dimensional time series, we formally defined the problem and developed a solution to discover abnormalities in frequent activities

projected in multi-dimensional time series. We focused our application domain on wearable embedded systems, mainly used for human activity monitoring. By emphasizing the importance of development of unsupervised techniques beside the frequently used supervised methods in such applications, we showed the application of our method in two state-of-the-art wearable systems, Smartcane and Smartshoe. Our technique has lower accuracy comparing to carefully designed and tuned supervised techniques, however, due to the unknown conditions of several real life scenarios, using an unsupervised technique is the only solution. The future work of this study includes extending the evaluation of the methods in more applications. Finally, we are currently leveraging the concept of frequent episodes in event streams [12] to improve the performance of frequent activity discovery in time series.

## 8 Acknowledgment

The authors wish to thank Professor Eamonn Keogh for providing us with the implementation of the motif detection algorithm. We also thank our colleagues in UCLA school of engineering, UCLA school of medicine, and Los Angeles VA hospital who have contributed in the development of the testbed systems.

To ensure reproducibility of the contents of this paper, we have provided access to all the data and developed techniques proposed in this paper on-line at: <http://cs.ucla.edu/~alireza/SDM2010>

## References

- [1] Javed Aslam, Katya Pelehov, and Daniela Rus. A practical clustering algorithm for static and dynamic information organization. In *SODA '99: Proceedings of the tenth annual ACM-SIAM symposium on Discrete algorithms*, pages 51–60, Philadelphia, PA, USA, 1999. Society for Industrial and Applied Mathematics.
- [2] Donald J. Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *AAAI-94: Workshop on Knowledge Discovery in Databases (KDD-94)*, 1994.
- [3] Xiao-yun Chen and Yan-yan Zhan. Multi-scale anomaly detection algorithm based on infrequent pattern of time series. *Journal of Computational and Applied Mathematics*, 214(1):227–237, 2008.
- [4] Bill Chiu, Eamonn Keogh, and Stefano Lonardi. Probabilistic discovery of time series motifs. In *KDD '03*:

- Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 493–498, NY, USA, 2003. ACM.
- [5] Foad Dabiri, Alireza Vahdatpour, Hyduke Noshadi, Hagop Hagopian, and Majid Sarrafzadeh. Ubiquitous personal assistive system for neuropathy. In *HealthNet '08: The 2nd International Workshop on Systems and Networking Support for Healthcare and Assisted Living Environments in conjunction with ACM MobiSys*, 2008.
  - [6] Amit Kale, Aravind Sundaresan, A. N. Rajagopalan, N. Cuntoor, A. Roy-chowdhury, and V. Krger. Identification of humans using gait. *IEEE Transactions on Image Processing*, 13:1163–1173, 2004.
  - [7] Eamonn Keogh, Jessica Lin, and Ada Fu. Hot sax: Efficiently finding the most unusual time series subsequence. In *ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 226–233, Washington, DC, USA, 2005. IEEE Computer Society.
  - [8] Eamonn Keogh, Stefano Lonardi, and Bill 'Yuan-chi' Chiu. Finding surprising patterns in a time series database in linear time and space. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 550–556, New York, NY, USA, 2002. ACM.
  - [9] Jessica Lin, Eamonn Keogh, Ada Fu, and Helga Van Herle. Approximations to magic: Finding unusual medical time series. In *CBMS '05: Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems*, pages 329–334, Washington, DC, USA, 2005. IEEE Computer Society.
  - [10] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *DMKD '03: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11, New York, NY, USA, 2003. ACM.
  - [11] Junshui Ma and Simon Perkins. Online novelty detection on temporal sequences. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–618, New York, NY, USA, 2003. ACM.
  - [12] Heikki Mannila, Hannu Toivonen, and A. Inkeri Verkamo. Discovery of frequent episodes in event sequences. *Data Mining Knowledge Discovery*, 1(3):259–289, 1997.
  - [13] David Minnen, Thad Starner, Irfan Essa, and Charles Isbell. Discovering characteristic actions from on-body sensor data. In *Wearable Computers, 2006 10th IEEE International Symposium on*, pages 11–18, Oct. 2006.
  - [14] David Minnen, Thad Starner, Irfan A. Essa, and Charles Lee Isbell Jr. Improving activity discovery with automatic neighborhood estimation. In *IJCAI '07: Twenty-first International Joint Conference on Artificial Intelligence*, pages 2814–2819, 2007.
  - [15] Abdullah Mueen, Eamonn J. Keogh, Qiang Zhu, Sydney Cash, and M. Brandon Westover. Exact discovery of time series motifs. In *SDM '09: 2009 Siam International Conference on Data Mining*, pages 473–484, 2009.
  - [16] Yoshiki Tanaka, Kazuhisa Iwamoto, and Kuniaki Uehara. Discovery of time-series motif from multi-dimensional data based on mdl principle. *Mach. Learn.*, 58(2-3):269–300, 2005.
  - [17] Heng Tang and Stephen Shaoyi Liao. Discovering original motifs with different lengths from time series. *Knowledge-Based Systems*, 21(7):666–671, 2008.
  - [18] Alireza Vahdatpour, Navid Amini, and Majid Sarrafzadeh. Toward unsupervised activity discovery using multi dimensional motif detection in time series. In *IJCAI '09: Twenty-first International Joint Conference on Artificial Intelligence*, 2009.
  - [19] Winston Wu, L. Au, B. Jordan, T. Stathopoulos, M. Batalin, W. Kaiser, Alireza Vahdatpour, M. Sarrafzadeh, M. Fang, and J. Chodosh. The smartcane system: an assistive device for geriatrics. In *BodyNets '08: The third International Conference on Body Area Networks*, pages 1–4, Belgium, 2008. ICST.
  - [20] Dragomir Yankov, Eamonn Keogh, Jose Medina, Bill Chiu, and Victor Zordan. Detecting time series motifs under uniform scaling. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 844–853, New York, NY, USA, 2007. ACM.