Detecting Anomalous Activities by Fusion of Accelerometer and Passive Infrared Sensor

Yasutaka Fukumoto  
Sony Corporation  
Tokyo, Japan  
Yasutaka.Fukumoto@jp.sony.com

Federico Castanedo  
Wise Athena & Deusto University  
San Francisco, USA & Bilbao, Spain  
fcastanedo@wiseathena.com

Hamid Aghajan  
Stanford University  
Stanford, USA  
aghajan@stanford.edu

Abstract—In ambient intelligence environment, cameras serve as interface or monitoring devices for user activities. However, the use of cameras is often associated with privacy concerns and resource limitations. In order to alleviate these problems, anomaly detection by fusion of an accelerometer and a passive infrared (PIR) sensor can be employed to trigger the corresponding camera for further analysis only as the need arises. This paper discusses the combination of these off-the-shelf sensors to detect specific anomalous activities. In particular, an example of detecting an irregular user in the work environment is presented. We describe how to extract expressive features from both modalities and combine them to train a classifier on highly imbalanced datasets. The experimental results over real-life data show the effectiveness of our approach in detecting anomalous activities and therefore potentially reducing the use of cameras.

I. INTRODUCTION

Cameras are important components of a smart home or smart office due to the rich information they provide for both users and computers. Several inferences can be made exploiting not only the streams of images to capture user appearances but also the structural settings such as camera positions and orientations. On the other hand, there remain privacy issues especially due to the nature of visual details constantly recorded in the data [1]. Moreover, such data require both large storage and extensive image processing. A reasonable approach to overcome these problems is to reduce the use of a camera by simply turning it off when nothing interesting is happening in the environment. The main question here is when it is necessary to turn back on the camera and gather images for further analysis. Regardless of whether for a security or monitoring purpose, anomalous activities particularly involving an intruder or a visitor can be informative triggers for the camera [2]. In other words, it is useful first to detect an irregular user in the environment without relying on actual images.

In this work, we discuss exploitation of two different types of sensors, accelerometer and passive infrared (PIR) sensor, each capturing different aspects of user activities. Specifically, we focus on a smart office setting, where an accelerometer-equipped smartphone is carried in a user’s pants pocket, and a PIR-based motion sensor is installed near the desk facing the user (see Fig. [1]). Thus, the accelerometer captures postural changes and ambulatory activities such as walking, while the PIR sensor captures upper body motions and transitions between presence and absence of a user. As long as there is only one monitored user working on the desk, these two sensor data should exhibit a regular pattern. In contrast, if another user comes to the desk while the monitored user is away, these data are expected to have a somewhat different pattern. For example, when the monitored user takes a stretch break, induced motions are observed in both data simultaneously, which is not the case for anomalous activities of an irregular user being at the desk. However, even in this apparent situation, detecting an irregular user becomes a challenging task since both of the users may happen to behave in a similar way, resulting in ambiguous patterns.

Our contributions in this paper are twofold. First, we describe a method for detecting anomalous activities by combining data from an accelerometer and a PIR sensor. Second, we show that it can be used to reduce the use of a camera with an example of detecting an irregular user in the work environment. More specifically, several instances of an irregular user being present at the desk are manually labeled in the data collection phase, and they are used to train several supervised classifiers in order to predict an anomalous activity. The feature extraction is explained in detail, particularly regarding how to combine discrete motion sensor outputs with acceleration readings. The accuracy comparison among different classifiers based on post-balanced training datasets is also presented.

The rest of the paper is organized as follows. Section II briefly reviews relevant work. Section III describes the approach for detecting anomalous activities, followed by the detailed methods of data collection, feature extraction, and supervised classification. We present the experimental results in Section IV and conclude the paper with a remark on future work in Section V.

II. RELATED WORK

For a residential or commercial security system, PIR sensors have been commonly utilized, for example, to detect an intruder in the particular area as described in [3]. Using a wireless sensor network of low-cost, low-power PIR sensors, [4] shows a technique to track the motion direction and distance of people in a hallway. In [5], distributed sensor information is exploited to estimate the count of people in an office space. This line of research does not deal with identified individuals, while our goal is to differentiate irregular users from a monitored user.

PIR sensors have been also incorporated into video surveillance systems. The simplest form of integration is to start recording images based on motion detection as described
The camera is useful for image processing per se and human eyes of an anomaly because the visual information captured by the camera. It is natural that the camera is turned on in the event between the accelerometer and the pair of the PIR sensor and usually merely turned off. (4) There is a preset correspondence is installed at the same place as the PIR sensor, which is of a user, and possibly other bodily movements. (3) A camera captures postural changes, ambulatory activities, and possibly other bodily movements. (2) A PIR sensor is installed at a designated place in the environment so that it remotely captures motion signal alone, the number of extracted features is inevitably limited. In this regard, accelerometers are increasingly utilized in activity recognition research due to not only their wide availability on mobile devices today but also their potential energy-efficiency realized by various techniques such as latent Dirichlet allocation (LDA) are used to create an occupancy model over long-term PIR data as in [10]. However, with the binary representation of motion signal alone, the performance of indoor complex activities is considerably limited. In this regard, accelerometers are increasingly utilized in activity recognition research due to not only their wide availability on mobile devices today but also their potential energy-efficiency realized by various techniques such as [11]. In [12], the recognition performance of indoor complex activities is investigated on a real-life dataset by solely using an accelerometer, and the promising results are obtained showing that accelerometers alone can extract quite informative features.

Not much work has been done on the combination of an accelerometer and a PIR sensor to analyze behavioral patterns, for example, of elderly people in order to detect an abnormal status. Unsupervised techniques such as latent Dirichlet allocation (LDA) are used to create an occupancy model over long-term PIR data as in [10]. However, with the binary representation of motion signal alone, the number of extracted features is inevitably limited. In this regard, accelerometers are increasingly utilized in activity recognition research due to not only their wide availability on mobile devices today but also their potential energy-efficiency realized by various techniques such as [11]. In [12], the recognition performance of indoor complex activities is investigated on a real-life dataset by solely using an accelerometer, and the promising results are obtained showing that accelerometers alone can extract quite informative features.

Our method comprises two stages: (1) Learn a model for anomaly detection based on various features calculated from continuous acceleration and motion data. (2) Use the model to detect an anomalous activity, and in which case, turn on the camera for further analysis of this anomaly. Here we define an anomalous activity as something that involves a different user than the monitored one. In case of an irregular user (not necessarily carrying an accelerometer) being present at the designated place, the accelerometer and PIR sensor assigned for the monitored user evidently indicate different patterns since the users have different behavior in most of the time. By adopting a machine learning framework, it becomes possible to detect such a situation. A multiplicity of those sensors is conceivable, but it is out of the scope of this paper.

A. Data Collection

There are many accelerometer-equipped smartphones and many residential or commercial PIR-based motion sensors in the market today. We choose to use the off-the-shelf sensors embedded in these products. The monitored user carries a smartphone (Sony XPERIA U ST25A) in his pants pocket. Note that the phone is used here for its convenience in collecting data, so any other accelerometer-equipped device can be substituted. A motion sensor (HomeSeer HSM100 S2) and a camera (Axis M1011 W) are installed near the desk facing the user. Note that PIR sensors are already implemented on some cameras, but we used separate products just for convenience sake. The sampling rate of three-dimensional acceleration is configured to be 50 Hz with a slight variation due to some system limitations. The PIR sensor is configured to report every motion (i.e. event-driven), but the maximum throughput is experimentally determined to be either 2 or 4 seconds (see Fig. 3). The camera is used to record videos for ground-truth annotation in the data collection phase during the first stage and is to be triggered in the event of an anomaly for further analysis during the second stage.
data indicating whether there is a motion or no motion. Labeled as positives and negatives are illustrated with motion positives and 99.15% are negatives. In Fig. 4, the whole data that we end up with highly imbalanced datasets; 0.85% are 112 minutes are manually labeled as positives (anomalies) period, a total of 221 hours of data are collected, out of which meanwhile, one of the co-workers is asked to be at the monitored user’s desk and do whatever he or she wants to do for bathroom, running some errands, or going out to lunch. The longest interval which is not in the graph turns out to be 96 minutes.

As for acceleration-related features, we can not rely on raw data just as usual, sometimes stretching, taking a short break for bathroom, running some errands, or going out to lunch. Meanwhile, one of the co-workers is asked to be at the monitored user’s desk and do whatever he or she wants to do while the monitored user is not there. Within a three-month period, a total of 221 hours of data are collected, out of which 112 minutes are manually labeled as positives (anomalies) observing 5 different co-workers in the video recordings. Note that we end up with highly imbalanced datasets; 0.85% are positives and 99.15% are negatives. In Fig. 4, the whole data labeled as positives and negatives are illustrated with motion data indicating whether there is a motion or no motion.

### B. Feature Extraction

As a preprocessing step, all acceleration data are filtered using a median filter and a low-pass Butterworth filter with 20 Hz cutoff frequency in order to remove some noise as explained in [13]. The acceleration and motion data together are then divided into a sequence of overlapping frames. Features are calculated only based on the data sample in each frame. The longer the frame length, the more computation for each frame required, but also the more information the features can be constructed from. The longer the overlapping length, the more computation for each sequence required, but also the more frequently the features can be calculated. This is an important factor for realizing a real-time anomaly detection. In this work, we explore such trade-offs by varying the frame length from 5 to 60 seconds while fixing the overlapping length to 5 seconds. Note that the first and the last frames in each set of data contain invalid data since the smartphone needs to be operated in hand at the start and the end of recordings. To sanitize data, those frames corresponding to the first and the last 90 seconds of each set are discarded from our final datasets.

For each frame, a total of 41 different features are extracted, out of which 14 features are related to only acceleration data, 3 to only motion, and 24 to both acceleration and motion. As for acceleration-related features, there have been a number of signal processing techniques developed in the context of activity recognition. Those techniques are broadly classified into three domains, namely time, frequency, and discrete representation domains as well summarized in [14]. In our office setting, basic activities such as walking and transitions from sitting to standing as well as other small bodily motions in the chair are significant. We consider only orientation-independent features since the smartphone can be positioned in any way in the pocket. We calculate the features in time and frequency domains as listed in the first part of Table I.

As for motion-related features, we can not rely on raw motion data due to the fact that the motion sensor sometimes

<table>
<thead>
<tr>
<th>TABLE I. SUMMARY OF FEATURES CALCULATED FROM ACCELERATION AND MOTION DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>F.1. Mean of magnitude of acceleration</td>
</tr>
<tr>
<td>F.2. Variance of magnitude of acceleration</td>
</tr>
<tr>
<td>F.3. Range of magnitude of acceleration</td>
</tr>
<tr>
<td>F.4. Zero-crossing of magnitude of acceleration</td>
</tr>
<tr>
<td>F.5. Maximum variance of acceleration of all axes</td>
</tr>
<tr>
<td>F.6. Maximum range of acceleration of all axes</td>
</tr>
<tr>
<td>F.7. Maximum zero-crossing of acceleration of all axes</td>
</tr>
<tr>
<td>F.8. Mean of magnitude of difference in acceleration</td>
</tr>
<tr>
<td>F.9. Variance of magnitude of difference in acceleration</td>
</tr>
<tr>
<td>F.10. Range of magnitude of difference in acceleration</td>
</tr>
<tr>
<td>F.11. Maximum variance of difference in acceleration of all axes</td>
</tr>
<tr>
<td>F.12. Maximum range of difference in acceleration of all axes</td>
</tr>
<tr>
<td>F.13. Centroid of one-sided power spectrum for magnitude of acceleration</td>
</tr>
<tr>
<td>F.14. Entropy of one-sided power spectrum for magnitude of acceleration</td>
</tr>
<tr>
<td>F.15. Area under curve of continuous motion data</td>
</tr>
<tr>
<td>F.16. Minimum interval time of discrete motion data</td>
</tr>
<tr>
<td>F.17. Maximum interval time of discrete motion data</td>
</tr>
<tr>
<td>F.18-21. Weighted version of F.1-4 using continuous motion data</td>
</tr>
<tr>
<td>F.22-24. Weighted version of F.5-7 using continuous motion data</td>
</tr>
<tr>
<td>F.25-27. Weighted version of F.8-10 using continuous motion data</td>
</tr>
<tr>
<td>F.28-29. Weighted version of F.11,12 using continuous motion data</td>
</tr>
<tr>
<td>F.30-33. Same as F.18-21 but using cont. mot. data per event and taking max. among them</td>
</tr>
<tr>
<td>F.34-36. Same as F.22-24 but using cont. mot. data per event and taking max. among them</td>
</tr>
<tr>
<td>F.37-39. Same as F.25-27 but using cont. mot. data per event and taking max. among them</td>
</tr>
<tr>
<td>F.40,41. Same as F.28,29 but using cont. mot. data per event and taking max. among them</td>
</tr>
</tbody>
</table>
Fig. 5. This graph illustrates an example of generating continuous motion data. Suppose there are two motion data (events) 4 seconds apart. The interpolation function is made of a product of two sigmoidal functions in such a way that it allows more uncertainty (longer tail) after an actual sensor output than before. By adding the two interpolation functions, we can fill in the gap between the sensor outputs. In this case, it is well modeled that we are not sure whether there is any motion in such a gap because it coincides with the conceivable maximum throughput of 4 seconds.

fails to report any motion for a while because of unknown hardware behavior. In addition, the fact that it can only generate motion data in discrete manner at relatively low throughput also results in imprecise feature extraction. Therefore, we design an interpolation function and replace each motion datum with it to interpret motion data as continuous signal. The function values are basically added up to generate final continuous motion data as described in Fig. 5. We also design a conversion function so as to compress interval time of motion data to a smaller range as described in Fig. 6. We then calculate a few features using these modified motion data (see the second part of Table I). Lastly, we take the continuous motion data as a weight function and apply it in various ways to all the previously calculated acceleration-related features. This produces many more features (e.g. weighted mean of magnitude of acceleration as to F18 in Table I) which are meant to extract correlations between acceleration and motion data (see Fig. 7).

C. Supervised Classification

Since having an irregular user in the workplace may be an indicator of an alarm, early identification of such a situation is important. Our goal is to predict at each time frame on the basis of the combined features if the sensor data are generated by the monitored user or by a different one.

If anomalous activities are annotated in advance, the problem here can be viewed as a supervised learning problem. Supervised classification methods can be roughly divided into white-box and black-box methods. White-box methods are those where the space is split up into disjoint sets, and the resulting model explains classification boundaries. On the other hand, black-box methods lack a clear explanation of the learned function.

Our collected data give a set of $N$ features $X = (x_1, \ldots, x_N)$ as input to a classifier and a label as its output which is a single binary variable we want to predict for each instance. Suppose the output variable is denoted by $Y \in \{y_0, y_1\}$ where $Y = y_1$ indicates the presence of an irregular user, and $Y = y_0$ indicates that of the monitored user. Our task is to learn the function $f(X) = E[Y|X]$ which is the expected value of the output variable $Y$ given a particular numerical combination of the input variables $X$.

However, the ratio of positive instances ($y_1$) to negative instances ($y_0$) is extremely small. This type of setting brings about the problem of training a classifier with highly imbalanced datasets. In order to overcome such a problem, at the data level, we use the SMOTE algorithm [15] and generate
artificial instances of the positive class to obtain a balanced training set.

1) Decision Trees: We apply decision trees as an example of a white-box classification method and in particular CART method \[16\]. Decision tree classifiers can be categorized by the style of partition. The CART method follows a divide and conquer approach and uses linear partitions perpendicular to the axes in building a model. It constructs a tree with only binary splits, and the resulting model can be represented as a binary tree. For the splitting criterion, the CART method depends on Gini diversity index. Let \( RF(y_j, S) \) denote the relative frequency of instances in a set \( S \) that belong to class \( y_j \). The Gini diversity index in case of \( k \) classes is generally defined as follows:

\[
I_{\text{Gini}}(S) = 1 - \sum_{j=0}^{k-1} \left( RF(y_j, S) \right)^2
\]

When \( S \) is partitioned into \( t \) subsets, the information gained due to the splits \( S_i \) is computed by the following equation:

\[
\text{Gain}(S) = I_{\text{Gini}}(S) - \sum_{i=1}^{t} \left( \frac{|S_i|}{|S|} I_{\text{Gini}}(S_i) \right)
\]

The CART uses minimal cost-complexity pruning technique, which assumes that the bias in the resubstitution error of a tree increases linearly with the number of leaf nodes.

2) Neural Networks: We also apply neural networks as an example of a black-box classification method. Neural networks are well-known techniques that mimic the human brain by creating a structure of neurons interacting with each other. The weights for connections between neurons are learned by an iterative process of refinement. Specifically, we adopt a single-hidden-layer feedforward neural network to learn the nonlinear function using BFGS optimization method, with which we can classify the instances into respective classes.

IV. EXPERIMENTAL RESULTS

The collected data are grouped into a test set (20%) and a learning set (80%) which is subsequently divided into subsets (training/validation sets) via 4-fold cross-validation. The training phase includes oversampling the training data with the SMOTE algorithm, so positive and negative instances become about equally likely. Several decision tree and neural network (with 10 neurons in the hidden layer and decay of 0.1) classifiers using different frame lengths are trained in order to predict rare events in the post-balanced training set. Regarding frame lengths, the experiment for 20-second frames reports the best sensitivity results with a mean of 81% over the test set and 69% over validation sets.

Fig. 8 shows the mean and standard error of the obtained sensitivity results for different frame lengths over validation and test sets. The results suggest that decision trees outperform neural networks in both validation and test sets. On the other hand, Fig. 9 shows the mean and standard error of the obtained specificity results. In this case, neural networks outperform decision trees, and the results over validation and test sets similarly improve in proportion to the frame length after that of 20 seconds. However, positive predictive value (PPV) results are not satisfactory because we have difficulty in avoiding false positives.

Our first contribution in this paper is verified by the sensitivity result, i.e. detection rate of 81% over the test set using the decision tree with 20-second frame length. With the same model, the second contribution is verified by the reduction of the number of images required to process, while 7,828 out of a total of 33,045 negative instances get reported as false positives, i.e. reduction rate of 76% over the test set.

Regarding the importance of the features to build the decision tree with 20-second frame length, we find that they are contributive in the following order; F36, F17, F6, F1, F2, F37, and F11 (cf. Table I).

V. CONCLUSIONS AND FUTURE WORK

This paper presented an insight into the combination of physical (mobile) and environmental (stationary) sensors to enhance user behavior analytics. We discussed an approach to reduce the use of a camera by limiting it to those situations where anomalous activities take place. Our method relies on fusing two other types of sensors, namely accelerometer and PIR sensor, in order to predict an anomalous activity.

We demonstrated our approach with realistic sensor data collected over a three-month period (221 hours), in which the instances (0.85%) of an irregular user being present at the monitored user's desk were manually labeled. These data were used to train decision trees and neural networks with different frame lengths in order to predict the labeled activities. Even with those off-the-shelf sensors and traditional supervised learning algorithms, we achieved the sensitivity of 81% and specificity of 76% using the decision tree with 20-second frame length. Compared to other classifiers, decision tree models are simple to interpret and easy to understand, and they typically perform well even with large datasets. However, a more sophisticated method needs to be contrived to achieve a higher PPV, so that the camera does not have to be triggered by false positives.

One direction for future research is to utilize continuous PIR sensor data instead of discrete data. Our preliminary experiments imply that such fine-grained data would provide more expressive features, from which we can expect to obtain better results for our purpose. We are also interested in applying unsupervised methods such as LDA to discover user behavioral patterns in general. In this case, non-predefined anomalous activities might be discovered to trigger the camera as well.

\[1^{\text{sensitivity}} = TP/(TP + FN)\]
\[2^{\text{specificity}} = TN/(TN + FP)\]

Fig. 8. Sensitivity results over validation and test sets when varying the frame length from 5 to 60 seconds while fixing the overlapping length to 5 seconds. Each point in the graph represents the mean and standard error computed based on the four configurations of training and validation/test sets.

Fig. 9. Specificity results over validation and test sets when varying the frame length from 5 to 60 seconds while fixing the overlapping length to 5 seconds. Each point in the graph represents the mean and standard error computed based on the four configurations of training and validation/test sets.


