Chapter 10
Data Fusion with a Dense Sensor Network for Anomaly Detection in Smart Homes

Kevin Bing-Yung Wong, Tongda Zhang and Hamid Aghajan

Abstract Research into assistive technologies for the elderly has been increasingly driven by the rapidly expanding population of older adults in many developed countries. One area of particular interest is technologies that enable aging-in-place, which allows older adults to remain in their own homes and live an independent life. Our work in this space is based on using a network of motion detectors in a smart home to extract patterns of behavior and classify them as either typical or atypical. Knowledge of these patterns can help caregivers and medical professionals in the study of any behavioral changes and enable better planning of care for their patients. Once we define and extract these patterns, we can construct behavioral feature vectors that will be the basis of our behavioral change detection system. These feature vectors can be further refined through traditional machine learning approaches such as K-means to extract any structure and reduce the dimensionality of the data. We can then use these behavioral features to identify significant variations across time, which could indicate atypical behavior. We validated our approach against features generated from human labeled activity annotations, and found that patterns derived from raw motion sensor data can be used as proxies for these higher level annotations. We observed that our machine learning-based feature vectors show a high correlation with the feature vectors derived from the higher level activity annotations and show a high classification accuracy in detecting potentially atypical behavior.

K.B.-Y. Wong (✉) · T. Zhang · H. Aghajan
Department of Electrical Engineering, Stanford University, Stanford, CA, USA
e-mail: kbw5@stanford.edu

T. Zhang
e-mail: tdzhang@stanford.edu

H. Aghajan
e-mail: hamid@icdsc.org

P. Spagnolo et al. (eds.), *Human Behavior Understanding in Networked Sensing*,
DOI 10.1007/978-3-319-10807-0_10

10.1 Introduction

The World Health Organization (WHO) and the US Department of Health have both predicted a rapid rise in the proportion of older adults that make up the populations of developed countries. Globally, it is predicted that the number of people older than 60 will triple between 2000 and 2050 from 600 million to 2 billion, while considering just the United States alone, the number of citizens older than 65 is expected to grow from 40.3 to 72.1 million in 20 years [1, 2]. This rapidly growing and aging population segment is of great concern in terms of their healthcare and management, since eldercare is traditionally very labor intensive and costly.

In response to the concern of how to care for an increasing elderly population, many assistive and monitoring technologies are being examined to reduce the need for caretakers and to enable the aging population to retain a measure of independence [3–7]. These efforts are mostly intended to enable aging-in-place, which is defined by the Center for Disease Control as "The ability to live in one's own home and community safely, independently, and comfortably, regardless of age, income, or ability level" [8]. This would reduce the infrastructure costs of eldercare, and would align better with the wishes of older adults. A survey conducted by AARP, an American advocacy group that addresses issues concerning older adults, indicated that 84 % of people over the age of 50 wish to remain in their current homes, with the percentage increasing to 95 % for respondents over the age of 75 [9].

The use of long-term health and wellbeing indicators is also beneficial for self-reflection as a motivational tool to change possibly unhealthy trends in lifestyle. These systems could be used by healthcare professionals and caregivers to augment existing doctor visits to provide doctors with a more continuous view of a patient's heath and wellbeing, as well as alert them if any sudden changes occur. The "Quantified Self" movement is an example of this recent trend for self reflection and health monitoring. Using environmental sensors would allow a monitoring system to unobtrusively monitor a patient continuously while they are in the home with no active compliance from the patient [10].

10.2 Related Work

There has been a great deal of work in developing assistive technologies for the elderly, with the goal of enabling so-called "aging in place." A common type of assistive technologies are those that monitor for anomalous activities or events, one such system developed by Shin et al., used a network of 5 PIR(Passive InfraRed) sensors to track the mobility of 9 elderly occupants of government sponsored housing [11]. This system first derived indicators of mobility, such as the percentage of time the motion sensors were triggered and how often the user would move between motion sensors, to look for changes in these indicators over time. The authors detected these changes by using 24 different SVDD (Support Vector Data Descriptors) based

classifiers, one for each hour of the day, to classify normal activities [12]. Since the SVDD classifiers would generate warnings for any abnormal activity, their system would often generate false warnings due to the irregular behavior patterns of the users, such as waking up late and performing cleaning activities during different times of day. Other work by O'Brien et al. also used PIR sensors in the home as a primary input, but their work was focused on visualization techniques to potentially identify movement disorders for the older adults [13]. Cuddihy et al. used the same PIR-based motion sensors with the goal of identifying periods of unusually long inactivity of occupants in their homes to generate alerts to caregivers [14]. More general work by Fine et al. used location-based information to build feature vectors to determine classes of normal or abnormal activities using a clustering-based approach [15].

The CASAS project, led by Diane Cook at Washington State University, has been studying a related field of identifying Activities of Daily Life or ADL, which are defined as common activities that a person performs to care for themselves, using data mining and machine learning to automatically identify important activities through finding the most common pattern in motion sensor data [16–19]. ADL's could be useful to describe the behavior of a smart environment's occupant and identify abnormalities if certain ADL's are not completed. Some of the recent work by Jakkula et al., focuses on using one class SVM's to classify anomalous behavior using an annotated dataset based on motion and door sensors in a home setting [20]. Anomaly detection by monitoring drifts and outliers of detected parameters were also studied by Jain et al., however that research focused less on ambient sensors and more on wearable health monitors [21]. More theoretical approaches to activity detection, such as work by Kalra et al. have focused on machine learning and statistical models for ADL detection [22].

There have been many other proposed approaches toward detecting, representing, and analyzing activity and behavioral patterns in a home setting. One such approach by Lymberopoulos et al. uses a home sensor network to track a user's motion and presence throughout the home [23]. Using region occupancy and the associated occupancy time, they create a set of symbols that encode the location, duration of the user's presence, and the time of day the user was present in a specific area in the home. From these symbols, they discover frequent sequences of symbols and their likelihoods to extract the users activity patterns form their 30 day dataset. Vision centric approaches by Gómez-Conde el al. [24], focus on detecting abnormal behavior using cameras and computer vision techniques such as motion detection and object segmentation to develop tele-assistance applications for the elderly. Their system mainly focuses on the sensing and classification of events, not recognizing longer term patterns or behaviors. Other research projects such as [25, 26] also focus more on shorter term monitoring and on techniques for person tracking and fall detection. Due to the general lack of long-term monitoring data and privacy issues with data collection, there has not been a significant body of work dedicated to long-term behavioral and health monitoring based on cameras.

10.3 Methodology

We examine the task of detecting atypical behaviors using two different approaches, using region-based occupancy patterns, derived from measuring how an occupant spends time in specific regions in their home, and concurrency-based models, which find important areas of a home automatically based on how an occupant moves through their home. For both approaches to atypical behavior detection, we use long-term datasets from WSU's CASAS project [16]. CASAS was intended as a smart home testbed, to evaluate algorithms for activity recognition and home automation. Common to the three datasets that we used, are a fairly dense deployment of 20+ PIR motion sensors, configured to have a small field of view. Judging from the documentation, we estimated that the motion sensors had a detection area of 8 m [27]. These motion sensor deployments do not offer the rich data available from cameras, but they sidestep several privacy issues inherent with placing cameras in people's homes. Since motion sensors can only detect the presence of motion in their fields of view, they will be better tolerated, even if data is recorded and stored for later processing or evaluation. Archiving video data for later processing would have significant privacy concerns for occupants, not to mention the logistics of storing months or years of video from multiple cameras.

Our initial experiments centered on a CASAS dataset featuring a single occupant for a 220 day experiment, codenamed "Aruba" by CASAS. The apartment layout and sensor placements can be seen in Fig. 10.1. We wanted to start with a single occupant, since motion sensors cannot differentiate between multiple people. For our region occupancy-based atypical behavior detector, we later extended our work to include a

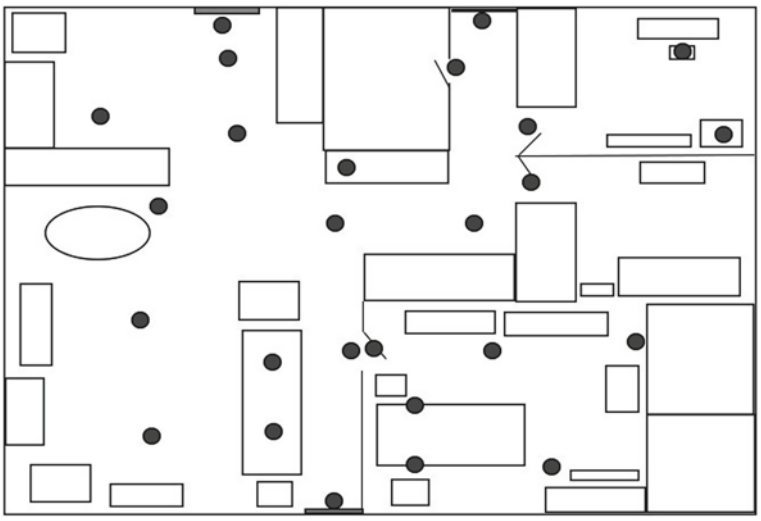


Fig. 10.1 "Aruba" apartment sensor layout

shorter 30 day dataset that contained an occupant with a pet, codenamed "Milan," and finally a 180 day dataset that contained a couple living together, codenamed "Tulim." All of the datasets that we evaluated had activities that were human annotated, and we used this as a baseline input to our atypical behavior system to compare against features derived directly from the motion sensors.

The following sections describes how we used the CASAS motion sensor-based data for both region occupancy and concurrent activation-based atypical behavior detection.

10.3.1 Region Occupancy-Based Atypical Behavior Detection

In order to identify atypical behaviors of a smart home's occupant, we need to have a representation for the occupant's behaviors. One clear choice is to use occupant activities to form a behavioral model, however, identifying activities using PIR's is a complicated task, since some important activities, such as sleeping, is marked by an absence of motion. We then hypothesized that activities in a home are closely tied to specific regions of a home, for example, cooking mostly takes place in the kitchen, and sleeping takes place mostly in the bedroom. Using that observation, we decided to use region occupancy as a proxy for occupant activities to generate daily patterns of behavior, which we can then compare across days to identify outliers, which we treat as atypical. These observations are illustrated in Fig. 10.2, in which we show both region occupancy and annotated activity patterns for each 30 minute window across 220 days of data collection, with each horizontal line in the image representing the occupant's occupied regions or activities for one day. The coloring for the patterns is different, since there is a different number of regions and activities, but there is still a clear visual similarity between the two. The following sections describe our steps to derive these region occupancies and atypical detections from the raw sensor data.

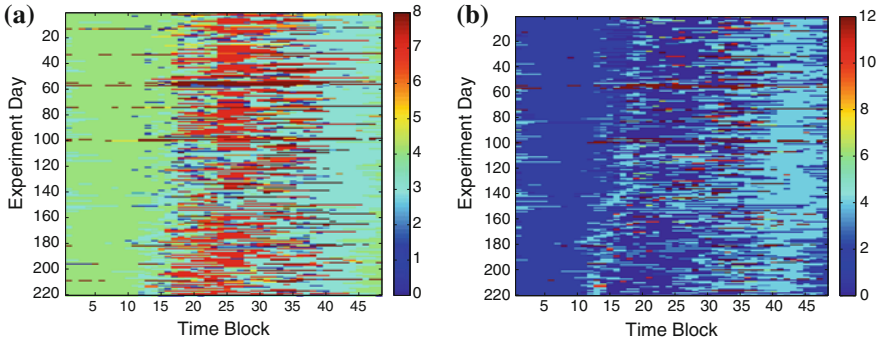


Fig. 10.2 Comparisons of region occupancy patterns to annotated activity patterns, based on most frequent region/activity in each 30 min time window. **a** Region occupancy patterns. **b** Annotated activity patterns

10.3.1.1 Initial Processing

The CASAS data consists of a series of PIR motion sensor activation events, which represent times in which an occupant or a guest moved within the detection area of one of the PIR's used in the deployment. These PIR's are located in Fig. 10.1 as the dark circles, and were mainly concentrated in the kitchen, living, and bedroom areas of the testbeds.

We first took these activation events, and examined them manually to look for sensor malfunctions. For example, for one day in the "Aruba" dataset, all of the motion sensors were triggered simultaneously for several hours. After we identified these days, we removed them from the dataset so they would not skew our results.

Next, we estimated the position of the occupant within the apartment based on PIR activation events. Since the user can activate many PIR's at any one moment, using the locations of the activated PIR's as the user's position would be quite noisy and subject to rapid changes. Thus, we decided to use a relatively simple approach to estimate the occupant's position by averaging the locations of the activated motion sensors within a small time window. Most of the PIR sensors used in the instrumented apartment were ceiling mounted units with a relatively small detection radius of 8 ft. With these pieces of information, we are able to determine the position and time of all motion detection events. This result was then temporally smoothed and used to assign a region label to every point in the occupant's estimated trajectory, as described in the next section.

10.3.1.2 Region Labeling

As a first step in forming the region occupancy patterns that we will later use for atypical behavior detection, we assign each point in the occupant's trajectory with a region label that corresponds to the room or semantic region where the occupant is currently located. For the "Aruba" dataset, the region labels include: *Kitchen*, *Dining Area*, *Living Room*, *Bedroom*, *Guest Bedroom*, *Office*, *Hallway* and *Outside*. The current scheme to map region labels is based on a lookup table that maps the current position of the user to one region. Figure 10.3 describes the different regions as different colored boxes overlaid on the deployment map for the dataset. Table 10.1 lists the region identifiers used. The time series of region labels is then post processed to filter out region changes that last less than 3 s. This acts to remove rapid region label oscillations that would occur if the user walked along the boundary of two regions. Similar region maps were generated for the "Milan" and "Tulim" datasets, as shown in Fig. 10.4a, b, respectively.

10.3.1.3 Region Occupancy Histograms

We decided to represent the time series of region occupancies, described in the previous section, as a set of region occupancy histograms for consecutive

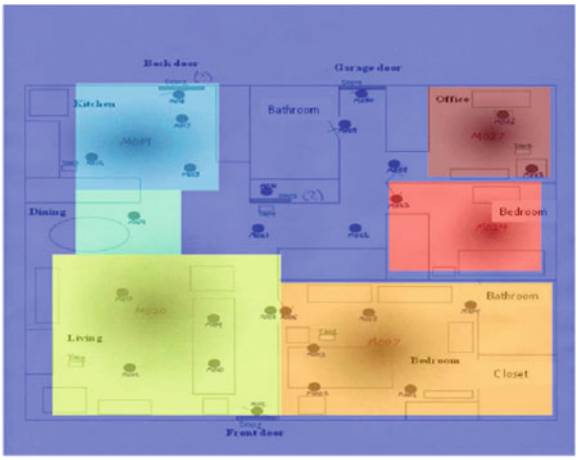


Fig. 10.3 Region map

Table 10.1 List of region labels used for processing

Identifier	Region description
0	Hallway
1	Kitchen
2	Dining area
3	Living room
4	Main bedroom
5	Guest bedroom
6	Office
7	Outside apartment
8	Inside with visitors

non-overlapping regions in time. These region occupancy histograms are more compact than the region time series, and can be interpreted more quickly than a long sequences of region occupancies. To form the actual region occupancy histograms that we used as a proxy for the occupant's behaviors, we first subdivided each day's data into smaller 30 minute chunks, and calculated the region occupancy histogram for each chunk. The size for the time chunk was chosen so that it would still be a reasonably compact size, to limit processing time for our experiments.

The result was 48 region histograms that represented the daily region occupancy patterns for the occupant of the apartment. We initially used several methods to summarize these 48 histograms to form even more compact representations, including using just the region with the highest occupancy for every time chunk. Other alternatives included using a Bag of Words representation along with a PCA to reduce the dimensionality of the data. We ultimately settled on simply concatenating these 48 histograms into a single feature vector to represent the occupant's region

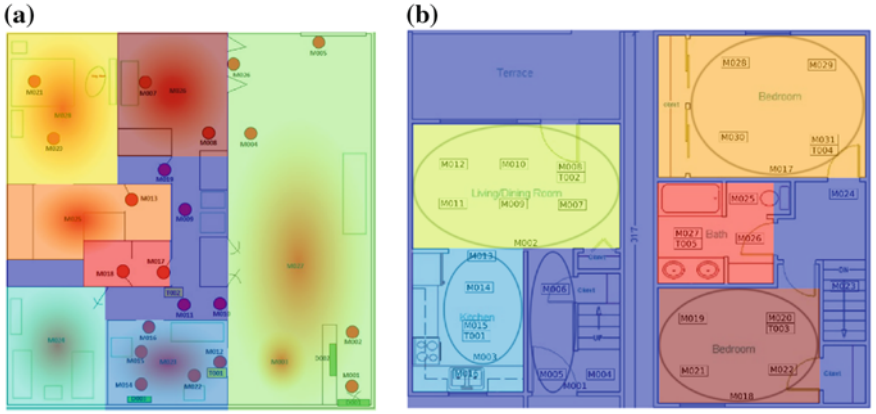


Fig. 10.4 Sensor layout and region maps of "Tulim" and "Milan" testbeds. **a** Milan testbed with region overlay. **b** Tulim testbed with region overlay

occupancy for each day, which we refer to as "Complete Region Histograms," since the compact representations did not perform significantly better, and were difficult to interpret. We later augmented these concatenated region histograms with region-specificmotion detection histograms, which represented the number of motion events in each region during the same time chunk, which formed "Complete Region and Motion Histograms" features.

For the sake of completeness, we experimented with using only the motion histograms, formed by histograms of the region a motion sensor was activated per time period, and an even finer histogram of which motion sensor was activated per time period. However, our results for these last sets of histograms showed poor correlation with our ground truth-based activity classifier. Table 10.2 briefly describes some of the different features that we evaluated.

10.3.1.4 Identifying Atypical Daily Patterns

After mapping the occupant's position to regions, and forming histogram-based patterns using region occupancy, we sought to identify region occupancy patterns that were atypical. We started by calculating how dissimilar each day's pattern was to every other day in our dataset. For our "Complete Region and Motion Histograms," we used the Euclidian distance between each pattern of concatenated histograms as our dissimilarity metric.

We can compactly represent the dissimilarity between all of the days of a dataset in the form of a distance matrix, where every entry (i, j) represents the dissimilarity between day i and day j, which is shown in Fig. 10.5 for both region occupancy and annotated activity-based patterns. We note that days that were very dissimilar to every other day, the red bands in the distance matrix, appear to be identical for both

Table 10.2 Different region/occupancy features considered for anomaly detection

Region occupancy feature	Associated distance metric	Description of descriptor
Most occupied region	Edit distance	The most frequently occupied region is used to represent each time period
"Bag of words feature"	Edit distance	The region occupancy histogram for a time period is mapped to the closest K-means based cluster for each time period
"Complete Region Histograms"	Euclidian distance	A concatenated vector of region occupancy histograms is used to represent each time period
"Complete Region and Motion Histograms"	Euclidian distance	The concatenated region occupancy histograms are appended with a histogram of region-based motion detection events
"Motion Only"	Euclidian distance	A histogram of region-based motion detection events for each time period

region occupancy and annotated activity-based patterns. Initially, we summed these dissimilarity of a single day i to every other day together to form a global dissimilarity metric for day i. Since lower values indicate a lower dissimilarity of a day's patterns to every other day in the dataset, we can use this global dissimilarity metric to classify days as atypical. To do this, we set a threshold on the global dissimilarity of each day to classify days with a high dissimilarity as atypical. A histogram of these global dissimilarity for every day can be seen in Fig. 10.6, we note that the histogram seems normally distributed, except that the distribution has a very long tail. We decided to find the average and standard deviation of these global dissimilarity values, and classify values that deviated from more than one standard deviation from the mean as atypical.

Our first approach used data from every day in the dataset in order to form global dissimilarity metrics to classify days as atypical, however, this is not a practical solution, since it assumes that we have access to data in the future. We wanted to also evaluate a sliding window approach to classification, where we would step through one day at a time and only classify the days based on the current and past data only. This approach may not identify the current day or past days as atypical if there is not enough historical data, for example, if the system just stared to collect data. So as each day is added to the sliding window, we reclassify all days using dissimilarity for all of the data up to the current day. We also experimented with different finite sliding window sizes, but we felt that the size of our daily region occupancy patterns was quite small, on the order of 10 KB, so we could reasonable store a lifetime's worth of occupancy data in less than 1 GB of memory.

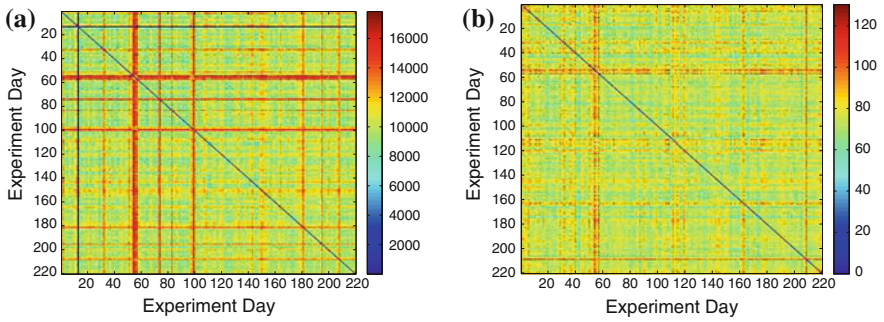


Fig. 10.5 Comparisons of distance matrices of the "Complete Region and Motion Histograms" patterns and CASAS annotated activities. **a** Region occupancy. **b** Annotated activity

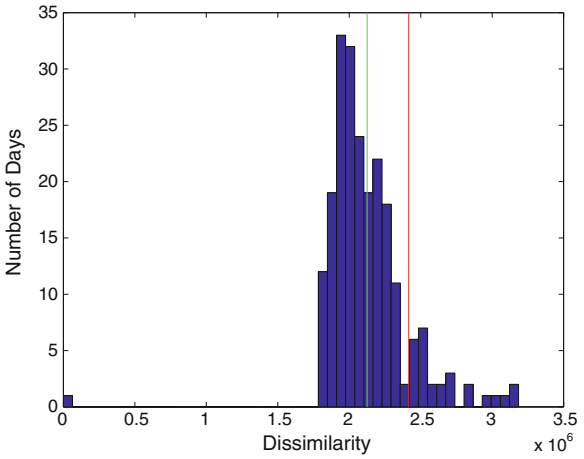


Fig. 10.6 Histogram of global dissimilarity values for the "Aruba" dataset

Lastly, we evaluated an alternative method of calculating dissimilarity other than our global dissimilarity metric. We decided to use the Local Outlier Factor or LOF, original proposed by Breunig et al. as a method to identify outliers in a dataset by comparing the local density around each point in the dataset [28]. The LOF algorithm also works on distance matrices, so we used our existing matrices calculated using our region occupancy patterns as an input and calculated the LOF for every day of our "Aruba" dataset. LOF has one parameter k, which is used to determine the size of the local neighborhood used to calculate the local density, we choose to use a value of 30, determined empirically. The LOF algorithm returns a value that represents the degree that a particular point is an outlier, but with no fixed threshold, we used the same method of choosing a threshold as with our global dissimilarly approach, after plotting a histogram of the LOF values as seeking a similar distribution as with dissimilarity.

10.3.1.5 Evaluating Region-Based Atypical Behavior Detector

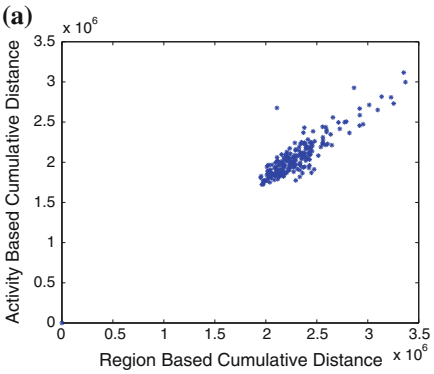
To evaluate the performance of using region occupancy-based behavior patterns to identify atypical behaviors, we leveraged the activity annotations included with the CASAS dataset, the activities included in the "Aruba" dataset are shown in Table 10.3. Using the same procedure as with the region occupancy patterns, we instead used activity patterns segmented into the same 30 minute time chunks. These activity patterns were histograms of the amount of time the occupant performed one of the annotated activities in each time chunk. With these activity histograms we used the same methods as with the region occupancy histograms to form daily descriptors and compared each day to every other day.

After using these annotated activity patterns to classify days as atypical we treated these classifications as a ground truth and compared them against the classification results from our region occupancy-based classifier and presented the resulting comparison as a confusion matrix. We also compared how the region-based dissimilarity compared to the activity-based dissimilarity for each day in the dataset. When both the region-based and activity-based dissimilarities are plotted against each other, we can evaluate how well the two relate to each other visually as well as calculate a correlation coefficient. A high degree of correlation can be used to determine if the region occupancy can be used as an effective proxy for annotated activities.

For the original "Aruba" dataset, we noticed that the "Complete Region and Motion Histograms," as described in Table 10.2 had the highest correlation coefficient at 0.916115, which indicates that region occupancy is a very good proxy for annotated activities for the purposes of atypical behavior detection. Figure 10.7a, shows the correlation plot and Fig. 10.7b shows a confusion matrix of the region occupancy-based classifier output when compared to the output of an annotated activity-based classifier. This approach yielded an accuracy of over 95 % and a precision of over 90 % for identifying atypical days.

Table 10.3 List of annotated activities in CASAS dataset

Identifier	Activity
1	Sleeping
2	Bed to toilet
3	Meal preparation
4	Relax
5	Housekeeping
6	Eating
7	Wash dishes
8	Leave home
9	Enter home
10	Work
11	Resperate



(b)

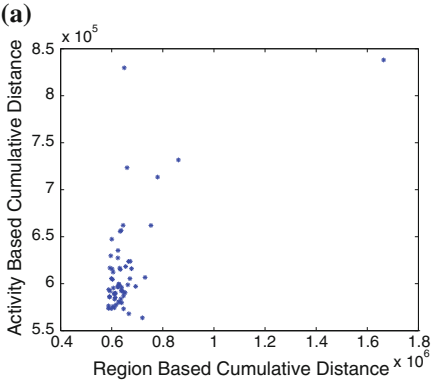
		Activity Based	
		Normal	Abnormal
Region	Normal	191	8
	Abnormal	2	19

Fig. 10.7 Correlation between region-and activity-based daily dissimilarity and a classification confusion matrix for the "Aruba" dataset. **a** Plot of daily total dissimilarity of the "Complete Region and Motion Histogram" patterns versus the CASAS annotated activities, correlation coefficient: 0.916115. **b** Confusion matrix of region-based abnormality detector of the "Complete Region and Motion Histograms" based classifier

We further wanted to verify that our approach was suitable for other homes and living situations, which is why we evaluated out approach for the "Milan" and "Tulim" datasets, which featured different apartment layouts and more than one occupant. Figure 10.4a, b show the layouts and regions, which are highlighted in different colors, for the "Milan" and "Tulim" testbeds, respectively. For these two datasets, we found that adding in the raw motion data to form the "Complete Region and Motion Histograms" yielded slightly worse performance compared to using just region occupancy. This can be attributed to the continued presence of multiple occupants, who would have generated many more motion events compared to the"Aruba" testbed, which just had a single occupant. From the correlation plots and confusion matrices of the "Milan" testbed, shown in Fig. 10.8a, b, we note that the correlation was quite lower than "Aruba" at 0.638051, this is probably due to the short duration of the "Milan" test set, which was considerable shorter at only 30 days. The "Tulim" testbeds correlation and confusion matrices, shown in Fig. 10.9a, b, show a better correlation of 0.801184, which is promising, considering "Tulim" had two occupants, which could imply that the occupants were frequently in the same region and performed the same activities, which for a coupling living together, seems highly likely.

We also wanted to study what effect using a sliding window approach rather than a global approach for atypical behavior detection would have on our classification accuracy. We found that using a sliding window, in which daily region occupancies patterns were added one at a time, did increase the number of false positives, which can be seen in Fig. 10.10a when compared to the global approach, which is reproduced in Fig. 10.10b. Most of these false positives occurred early in the dataset, when there was not enough region occupancy histogram to adequately separate out atypical from normal behavior.

(a)

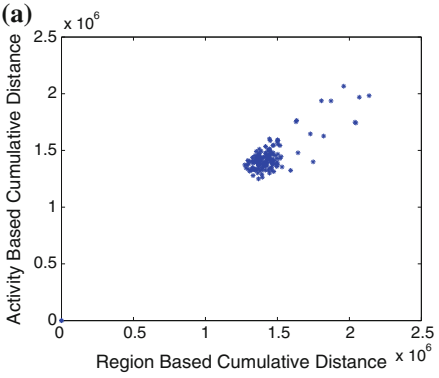


(b)

		Activity Based	
		Normal	Abnormal
Region	Normal	80	1
	Abnormal	0	2

Fig. 10.8 Correlation between region and activity-based daily dissimilarity and a classification confusion matrix for the "Milan" dataset. **a** Plot of daily total dissimilarity of the "Complete Region Histograms" patterns versus the CASAS annotated activities for the "Milan" dataset, correlation coefficient: 0.638051. **b** Confusion matrix of region-based abnormality detector of the "Complete Region Histograms" based classifier for the "Milan" dataset

(a)



(b)

		Activity Based	
		Normal	Abnormal
Region	Normal	174	2
	Abnormal	2	7

Fig. 10.9 Correlation between region-and activity-based daily dissimilarity and a classification confusion matrix for the "Tulim" dataset. **a** Plot of daily total dissimilarity of the "Complete Region Histograms" patterns versus the CASAS annotated activities for the "Tulim" dataset, correlation coefficient: 0.801184. **b** Confusion matrix of region-based abnormality detector of the "Complete Region Histograms" based classifier for the "Tulim" dataset

Lastly, we compared the performance of a LOF-based classify compared to a classify based on global dissimilarities of annotated activities, to evaluate other methods of identifying outliers. We found good correlation between the two, 0.86167, as shown in Fig. 10.11a. The confusion matrix of classification results, as shown in Fig. 10.11b , were not as good as with the global or sliding window approaches, with more false positive and negative classifications. This increase in false classifications could be due to an overly conservative threshold that we applied to the LOF value to determine if a day was an outlier.

(a)

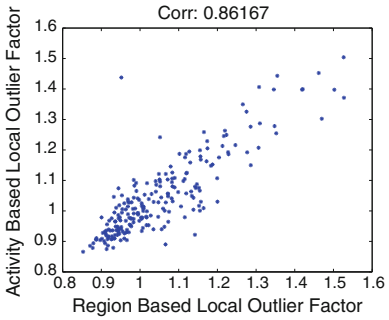
		Activity Based	
		Normal	Abnormal
Region Based	Normal	184	6
	Abnormal	7	23

(b)

		Activity Based	
		Normal	Abnormal
Region Based	Normal	187	6
	Abnormal	4	23

Fig. 10.10 Comparing a sliding window approach to a global approach for atypical behavior detection. **a** Sliding window approach, 220 day maximum size. **b** Global approach

(a)



(b)

		Activity Based	
		Normal	Abnormal
LOF Region Based	Normal	182	8
	Abnormal	9	21

Fig. 10.11 Correlation between region-and activity-based LoF scores and a classification confusion matrix for the "Aruba" dataset. **a** Plot of daily LoF of the "Complete Region and Motion Histograms" patterns versus the CASAS annotated activities for the "Aruba" dataset, correlation coefficient: 0.86167. **b** Confusion matrix of region-based abnormality detector using a LoF- based classifier for the "Aruba" dataset

10.3.1.6 Interpreting Atypical Patterns

The previous section primarily evaluated how well region-based atypical behavior compared to a similar classification process using human annotated activities. In this section, we examine the results of the atypical behavioral detector by looking at the region occupancy patterns of the days classified as normal or atypical.

Figure 10.12 shows features that correspond to normal days, with each horizontal line corresponding to a day's region occupancy patterns. Each one is stacked on top of each other, so looking at vertical columns would represent the region occupancy at the same time across days. These histogram patterns have been resorted, so that the components of the histograms that represent the same region are adjacent to each other, with highlighted boxes separating the different region's occupancy. The labels above each block represent the semantic region that the block represents. Each region block is broken up into 48 columns, representing the 48 time chunks in which the histograms were originally calculated, so each day's occupancy of that region is encoded left to right, with the far left side representing midnight. We can then examine each block to determine how the occupancy of a specific region varies within a day, and throughout the dataset by looking horizontally and vertically respectively. When examining the occupancy patterns for the bedroom, we see a high occupancy for the

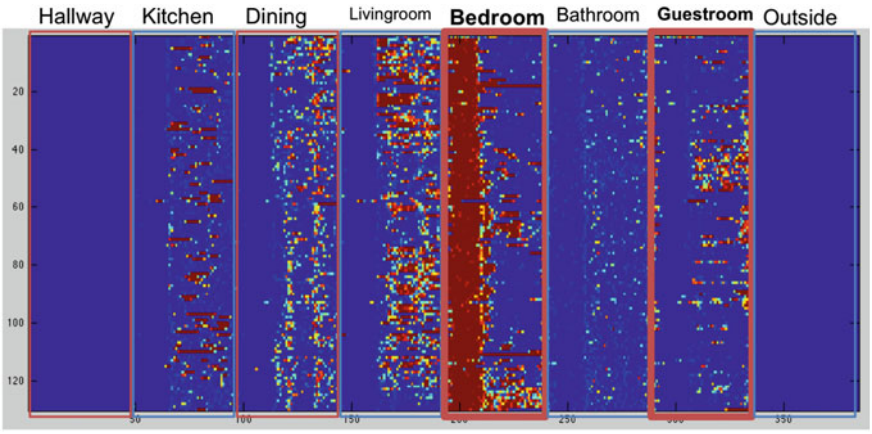


Fig. 10.12 Region occupancy histograms for normal days

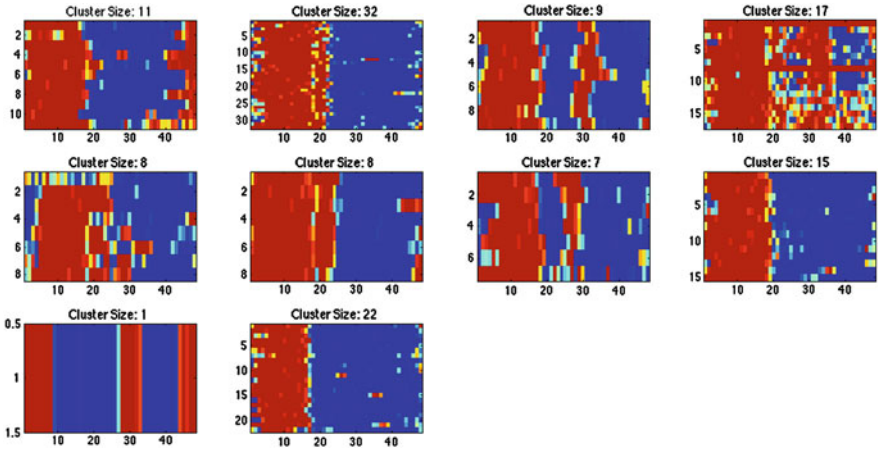


Fig. 10.13 Cluster analysis of bedroom occupancy patterns for normal days

early hours, most likely representing sleep. Also note that the normal days do not have significant Guest room occupancy during the same early hours time period.

Using the bedroom as an example, we sought to understand the variations of the bedroom region occupancy for our "Normal" days. We performed a clustering analysis of the bedroom occupancies using k-means, which yielded 10 unique activity patterns for normally classified behavior. These bedroom occupancy clusters can be seen in Fig. 10.13. We note that these cluster represent different sleep and wake times, and that these clusters of bedroom occupancy could be used to monitor sleep habits be used to specifically monitor for atypical sleep patterns by comparing the bedroom occupancy patterns with these common "Normal" bedroom occupancy clusters.

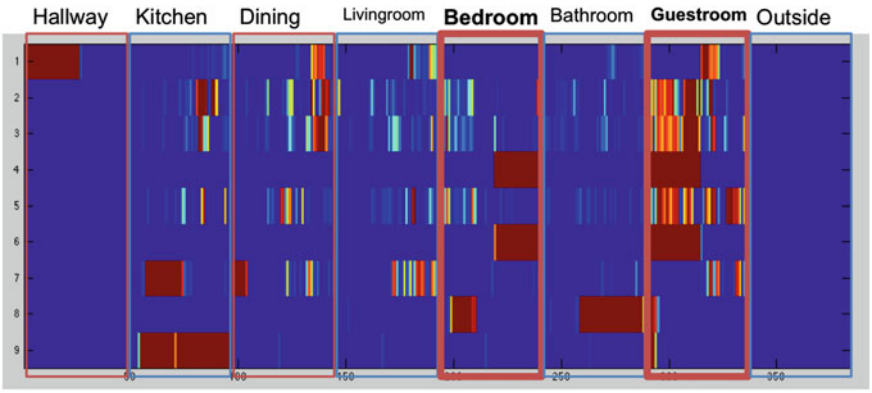


Fig. 10.14 Region occupancy histograms for atypical days

Figure 10.14 shows the region occupancy patterns that correspond to atypical days. These patterns appear quite different from the "Normal" days show in Fig. 10.12. If we compare just the Bedroom and Guestroom regions, it seems like there is much higher overnight Guestroom occupancy compared to the "Normal" days. This could be why these days were statistically different from the "Normal" days.

Caregivers and medical professions could also view these occupancy patterns for both normal an atypical days to look for causes of the atypical classification, for example, they could compare the occupancy patterns of specific regions across both the "Normal" and "Atypical" classes to see of the classification was due to changes in a specific region.

10.3.2 Atypical Behavior Detection Using Concurrent Hotspot

In a motion sensor network, a concurrent activation event is when a moving object enters an area that overlaps with multiple sensors' detecting region and generates one or more motion detection events. In this section, we will develop and use a concurrent activation model to find concurrent hotspots (the location where occupants move frequently). Then, an occupant's daily behavior at these hotspots will be used to estimate how likely that occupant is displaying anomalous behavior.

The basic intuition behind this analysis is that people tend to maintain regular daily routines and the behavior at the concurrent hotspots are able to summarize the occupant's daily activity. So any deviation of a occupant's behavior at these concurrent hotspots might indicate atypical behavior.

10.3.2.1 Finding Concurrent Hotspots

Since a concurrent hotspot is the location where occupants move frequently, we need to first define a concurrent event and the location of such events.

Assuming that a motion sensor network consists of N binary motion sensors, defined as sensors only have two possible states $\{0, 1\}$. In the cause of motion sensors, a 0 state means that no motion is detected, and 1 indicates the sensor is being triggered by motion at this time.

The model of a single sensor's activation can be demonstrated with Fig. 10.15. To simplify the activation model, the detection region of a motion sensor is considered a circular area. If an object is moving inside the detection region, the sensor will be activated; conversely, if the moving object is outside the detection boundary, the sensor will not be triggered. However, some exceptions could generate false detections, PIR-based motion sensors can be triggered by sudden local changes in temperature, such has air moving through a heating or cooling vent. We define the probability that a sensor is activated with no moving object present as false positive rate β. Our previous work shows that this model fits the CASAS Aruba data set quit well, with an estimated false positive rate of $\widetilde{\beta} = 0.0406$, was calculated by a Genetic Optimization algorithm.

With this single sensor's activation model, we can now turn to the concurrent activation case. Suppose a moving object is at the location with coordinates (x_u, y_u) with k sensors being activated at the same time. Intuitively, that object is somewhere in the overlapping detection areas of those k sensors. Assuming that all of sensors have the same detection area size and have detection centers located at $\{(x_{\lambda 1}, y_{\lambda 1}), (x_{\lambda 2}, y_{\lambda 2}), \ldots, (x_{\lambda k}, y_{\lambda k})\}$, the most likely estimation of user's location $(\hat{x}_u, \hat{y}_u)$ is given by the following equation:

$$\hat{x}_u = \frac{1}{k}\sum_{i=1}^{k} x_{\lambda i} \tag{10.1}$$

$$\hat{y}_u = \frac{1}{k}\sum_{i=1}^{k} y_{\lambda i}$$

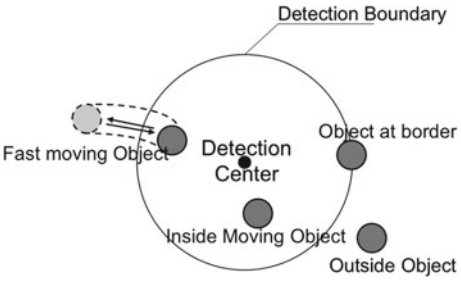


Fig. 10.15 Model of a single sensor's activation

Equation 10.1 is simply averages of locations of all the activated sensors. This approach is easy to implement but it ignores the fact that some of the sensors are activated by false positive events as we mentioned earlier. For example, if some of the active sensors amount those k activated sensors are activated by environmental temperature changes or some other accidental factors (false positive), the above approach described by Eq. 10.1 might give a result far away from the occupant's actual location.

In order to limit of the influence caused by false positive motion detection events, we leverage a Concurrency Graph that contains prior knowledge of overlapping areas of sensors' detection regions. The Concurrency Graph is defined as a weighted graph $G = (V, E)$, which consists of a set of motion sensors V as nodes and a set of edges E. A concurrent activation event will increase the weight of edges between any pair of activated sensors by one. Therefore, the higher the weight of a edge, more likely those two involved sensors will be activated at the same time.

After filtering out the edges with weight less than a certain threshold $T = \lambda \cdot \widetilde{\beta}^2$, where λ is the total number of concurrent events, and $\widetilde{\beta}$ is the false positive rate, most of the edges created by false positive activation can be removed.

Figure 10.16 shows the concurrency graph built on the "Aruba" dataset from 2010-11-01 to 2010-12-01. The red circles in the figure are the sensor nodes; the links are the edges between nodes; and numbers on the edges are the weights of corresponding edges. Figure 10.17 shows the concurrency graph after filtering out edges using the threshold mentioned above.

With the Concurrency Graph, if we have k activated sensor, instead of calculating the average location directly, we mark the corresponding k nodes in the concurrency graph Fig. 10.17. This will generate one or more sub graphs, where we choose the subgraph, G_u, with largest number of nodes to estimate the occupant's location with Eq. 10.2.

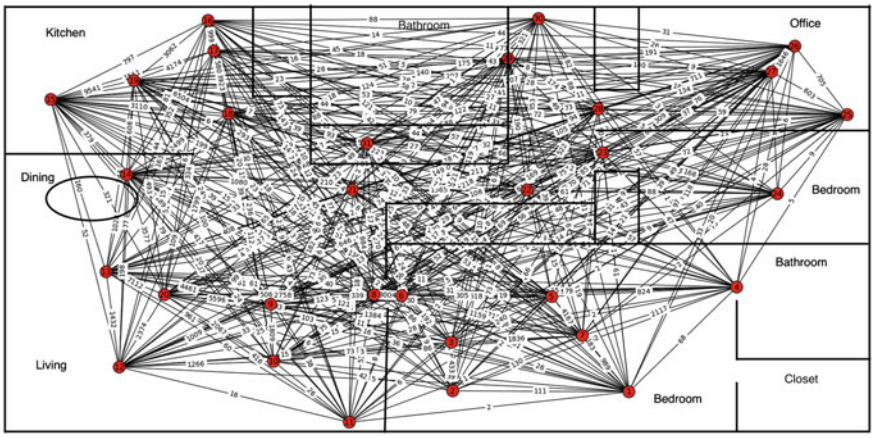


Fig. 10.16 Concurrency graph for 2010–11

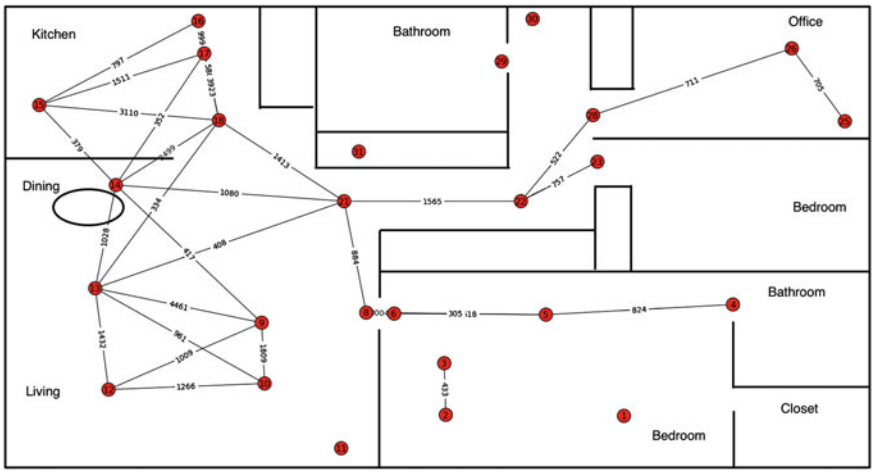


Fig. 10.17 Concurrency graph for 2010–11 with thresholding and area sensor deleted

$$\hat{x_u} = \frac{1}{|G_u|} \sum_{i:s_i \in G_u} x_i,$$
$$\hat{y_u} = \frac{1}{|G_u|} \sum_{i:s_i \in G_u} y_i \tag{10.2}$$

We define the concurrent hotspots as center of areas where an occupant frequently moves. Every time the activation state of any sensor changes, an estimation of the user's location can be calculated by Eq. 10.2. To find the concurrent hotspots we just defined, we go through the whole "Aruba" dataset and get a sequence of estimated user's active locations $\{(\hat{x}_{u1}, \hat{y}_{u1}), \ldots, (\hat{x}_{un}, \hat{y}_{un})\}$. If we can cluster the sequence into several groups, then the centroids of these clusters will be the hotspots that we are looking for.

We choose to use K-Means as our clustering algorithm, to decide how many clusters, k to use, we plotted the clustering error or cost, defined as the distance of every point to their assigned cluster, versus the number of clusters, as shown in Fig. 10.18. As can be seen from the figure, a value $k = 10$ is a good choice since it is the turning point of the cost curve. While K-means has a component of randomness, as initial cluster centroids are randomly selected, which may yield a different result, multiple experiments resulted in very similar centroid sets. We picked one of these sets shown in Fig. 10.19, where the red circles are the hotspots found, and the size of the circle indicates the size of the cluster, which represents how many concurrency events it encompasses.

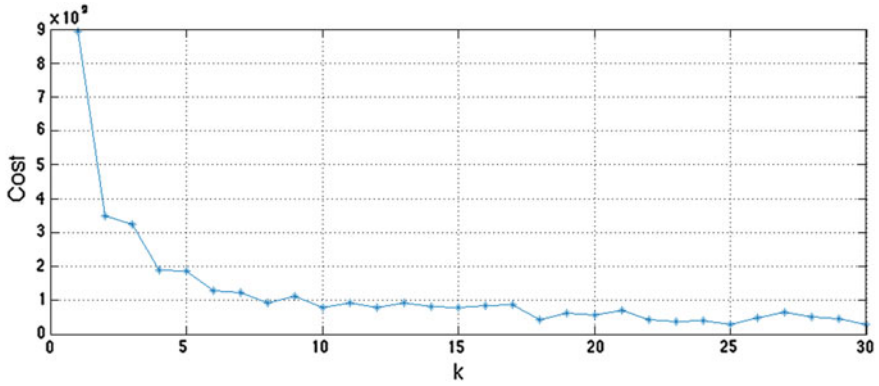


Fig. 10.18 Kmeans: cost versus cluster number

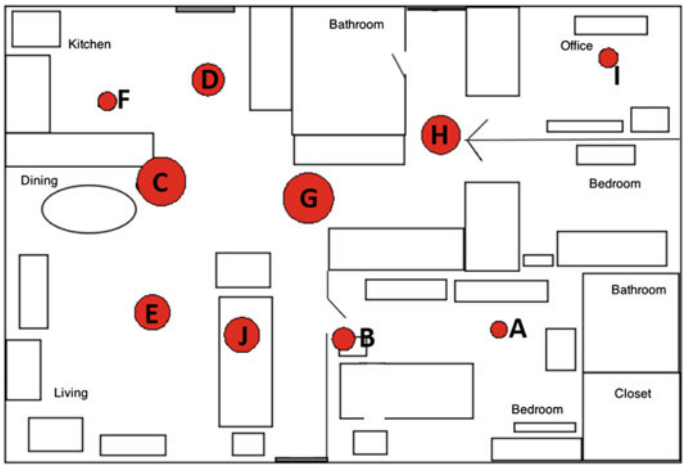


Fig. 10.19 Hotspot distribution

10.3.2.2 Hotspot-Based Occupant's Behavior Modeling Selection

The hotspots we found in the previous section are calculated by clustering locations where the occupant frequently moves. As a result, different clusters represents the centers of the occupant's activities to some extent. For examples, clusters *A* and *B* in Fig. 10.19 are related to the occupant's activities in the bedroom (sleeping, walking, and so on); clusters *D*, *G*, and *H* represent walking through hallway; cluster *E* reflects occupant's movements inside the living room; cluster *J* is for the sofa related activity; cluster *F* acts as the indicator of the occupant's cooking activity; and cluster *I* is the occupant's behavior in the office.

With the help of concurrent hotspots, a occupant's daily activities can be divided into different clusters according to their relative locations to the hotspots. To

summarize the occupant's daily active level at each concurrent hotspot, we use Algorithm 1 that can summarize the occupant's active level of one day into a 1×10 vector $V = [v_1, v_2, \ldots, v_{10}]$. The algorithm is described as follows:

Algorithm 1: Active Level Summary with Hotspots

Data: $H = \{h_1, h_2, \ldots, h_{10}\}$
$E = \{(\hat{x}_{u1}, \hat{y}_{u1}), (\hat{x}_{u2}, \hat{y}_{u2}), \ldots, (\hat{x}_{un}, \hat{y}_{un})\}$
Result: $V = [v_1, v_2, \ldots, v_{10}]$
0.1 initialize V to a zero vector ;
0.2 **foreach** $(\hat{x}_u, \hat{y}_u) \in E$ **do**
0.3 $\quad \hat{i} = \arg\min_{i:h_i \in H} \left(distance(h, (\hat{x}_u, \hat{y}_u))\right)$;
0.4 $\quad wcore = ScoreMapping(t)$;
0.5 $\quad v_{\hat{i}} = v_{\hat{i}} + wcore$;
0.6 **end**

In Algorithm 1, H is the Hotspots Set, where that h_i is the location of the ith concurrent hotspot; E is a sequence of user's active locations estimated by Eq. 10.2 for a day; $ScoreMapping(t)$ is a function that map time duration t into a active level score. Therefore, the algorithm calculates the active level score of the occupant and accumulates the score for each concurrent hotspot.

The $ScoreMapping(t)$ function in the above algorithm defines how we evaluate a occupant's active level from the amount of time they are moving at one location. For example, if the $ScoreMapping()$ is linear, the algorithm just accumulates the time an occupant spends in a certain hotspot. In this chapter, a super linear function is chosen as $ScoreMapping()$ which indicates that a long-time continuous activity has more impact than several short-time activities together.

For the whole CASAS "Aruba" dataset, we apply Algorithm 1 for each day and get a $N \times 10$ result matrix, where N is the total number of days, and each row is the output result of the algorithm 1 for the corresponding day. Therefore, a column $C_j = [v_j^{(1)}, v_j^{(2)}, \ldots, v_j^{(N)}]^T$ $(j = 1, 2, \ldots, 10)$ of the result matrix is the occupant's active level scores at the jth hotspot for N days.

Now we can use this resulting model derived from time spend in each hotspot to detect abnormal activity of the occupant. Our assumption is that the occupant's activity level at each concurrent hotspot satisfies is a specific distribution.

Therefore, we set up a distribution library which contains Normal, Rayleigh, Rician, t location-scale, Weibull, and Generalized extreme value distributions. For each concurrent hotspot, we fit the occupant's active level scores to every distributions in our distribution library. And we choose the distribution with the highest log likelihood as the behavior model for the occupant at that hotspot. Table 10.4 shows the result for each concurrent hotspot.

The histogram of occupant's active level scores at hotspots A and B, and the result curves of corresponding fitted distribution are showed in Fig. 10.20a, b. It can be seen that some of the curves fit the active level scores well. The Table 10.5 summaries the best probabilistic models and parameters for all the concurrent hotspots.

For at location-scale distribution with parameter μ, σ, ν, the probabilistic density function is

Table 10.4 Concurrent hotspot model fitting

HotSpot	Normal	Rayleigh	Rician	t location-scale	Weibull	Generalized extreme value
A	−1232.68	−1256.35	−1230.19	−1199.89	−1236.12	−1205.71
B	−1500.01	−1585.86	−1500.38	−1499.56	−1500.57	−1498.22
C	−1316	−1287.56	−1287.56	−1300.66	−1286.7	−1260.19
D	−1299.57	−1279.22	−1279.22	−1275.59	−1278.62	−1243.49
E	−1326.36	−1308.94	−1308.94	−1314	−1307.84	−1283.66
F	−1287.33	−1259.48	−1259.48	−1268.75	−1258.69	−1235.75
G	−1139.27	−1117.72	−1117.7	−1129.07	−1117.58	−1094.88
H	−1534.13	−1534.16	−1534.16	−1533.94	−1531.83	−1530.7
I	−1339.2	NA	NA	−1214.28	NA	−1112.02
J	−1577.44	−1608.62	−1576.43	−1572.18	−1579.38	−1574.47

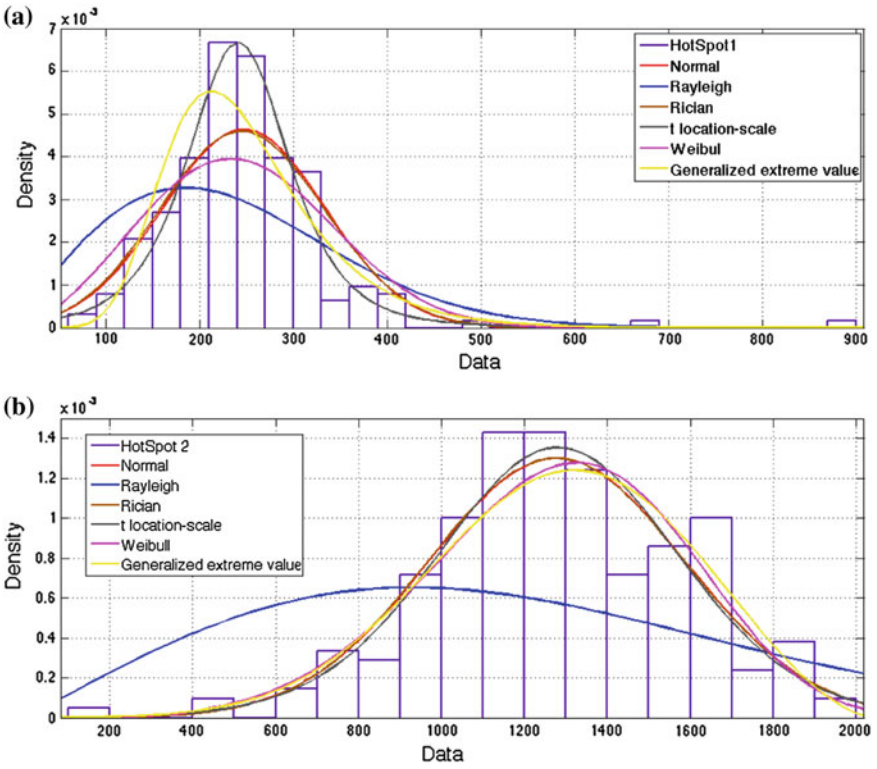


Fig. 10.20 Histogram and model fitting for hotspot A and B. **a** Hotspot A. **b** Hotspot B

Table 10.5 Hotspot model fitting

Hotspot	Probabilistic model	Parameters
A	t location-scale	$\mu = 240.839\ \sigma = 56.1281\ \nu = 3.93679$
B	Generalized extreme value	$k = -0.371485\ \sigma = 321.032\ \nu = 1183.27$
C	Generalized extreme value	$k = 0.294019\ \sigma = 70.3057\ \nu = 154.99$
D	Generalized extreme value	$k = 0.222946\ \sigma = 67.6701\ \nu = 178.837$
E	Generalized extreme value	$k = 0.175264\ \sigma = 84.0357\ \nu = 205.644$
F	Generalized extreme value	$k = 0.245482\ \sigma = 64.3452\ \nu = 137.352$
G	Generalized extreme value	$k = 0.201744\ \sigma = 33.6826\ \nu = 78.0005$
H	Generalized extreme value	$k = -0.152369\ \sigma = 330.334\ \nu = 509.017$
I	Generalized extreme value	$k = 1.35126\ \sigma = 20.4538\ \nu = 12.7242$
J	t location-scale	$\mu = 1297.43\ \sigma = 368.005\ \nu = 6.48294$

$$p(x;\mu,\sigma,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{(\frac{x-\mu}{\sigma})^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

For a generalized extreme value distribution with parameter k, σ, ν, the probabilistic density function is

$$p(x;k,\sigma,\nu) = \frac{1}{\sigma}\left[1+\nu(\frac{x-k}{\sigma})\right]^{-\frac{1}{\nu}-1}\exp\left(-\left[1+\nu(\frac{x-k}{\sigma})\right]^{-\frac{1}{\nu}}\right)$$

With fitted probabilistic model of each concurrent hotspot, we propose a multivariable distribution model PM that models the occupant's daily active level score vector (the 1 by 10 vector that contains active score for each hotspot):

$$V = [v_1, v_2, \ldots, v_{10}] \sim \mathrm{PM}(p_1, p_2, \ldots, p_{10}) \tag{10.3}$$

where p_i $(i = 1, 2, \ldots, 10)$ is the probabilistic distribution model for hotspot i showed in Table 10.5. The model (Eq. 10.3) is the representation random distribution model for user's daily active level at each concurrent hotspot.

10.3.2.3 Abnormal and Typical Active Level Detection

We define the normalness of a day as log likelihood the occupant's active level score vector which is calculated by our proposed PM model (Eq. 10.3).

Given the active score vector $V = [v_1, v_2, \ldots, v_{10}]$ and the distribution model is $\mathrm{PM}(p_1, p_2, \ldots, p_{10})$, the likelihood of the active level scores can be calculated by the following equation:

$$L = \prod_{i=1}^{10} p_i(v_i) \tag{10.4}$$

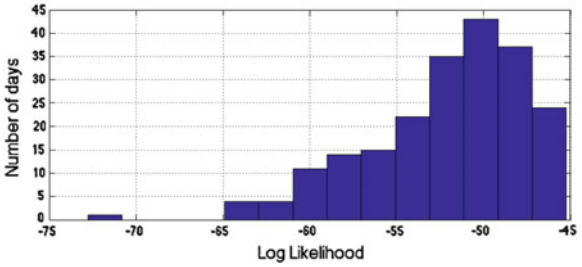


Fig. 10.21 The histogram of log probability score

Then, the log likelihood's calculation becomes straightforward:

$$\text{Log}(L) = \sum_{i=1}^{10} \log(p_i(v_i)) \tag{10.5}$$

Using Eq. 10.5, the higher the value of $\text{Log}(L)$ indicates that a specific day is more likely to be a normal day. Similarly, the lower the value, the more likely that day is a atypical day. The histogram of the log likelihood of all the days is shown in Fig. 10.21.

Setting the lower threshold $T_{\text{low}} = -57$ and the higher threshold $T_{\text{high}} = -47$, days with log likelihood greater than T_{high} are marked as typical, days with log likelihood less than T_{low} are considered as abnormal days. Using this rule, we have the result that {1, 2, 6, 32, 54, 56, 57, 73, 74, 76, 83, 90, 99, 100, 103, 107, 108, 109, 111, 132, 142, 143, 155, 167, 170, 181, 200, 201, 207} are classified as be atypical, and days {7, 20, 23, 28, 30, 31, 35, 38, 42, 43, 53, 58, 65, 69, 71, 72, 78, 85, 88, 89, 91, 94, 05, 06, 13, 17, 22, 24, 26, 38, 48, 64, 68, 76, 77, 78, 89, 93, 94, 96, 103, 105, 210} are marked as typical, all the other days are normal.

In summary, the whole process of detecting the occupant's abnormal behavior is very straightforward. It starts from location estimation based on concurrency graph, representing user's daily active level using a score vector, to calculating the log likelihood of the score vector. Therefore, The proposed method offered us a simple method to interpret the information from binary motion sensors directly into a user daily behavior summary.

10.4 Future Work

The two approaches that we have used so far, region occupancy-based and concurrency-based show promise in identifying whole days as"Normal" or "Atypical," however, it is not clear why a specific day was classified as "Atypical." Our future work seeks to address this issue by incorporating features that more closely

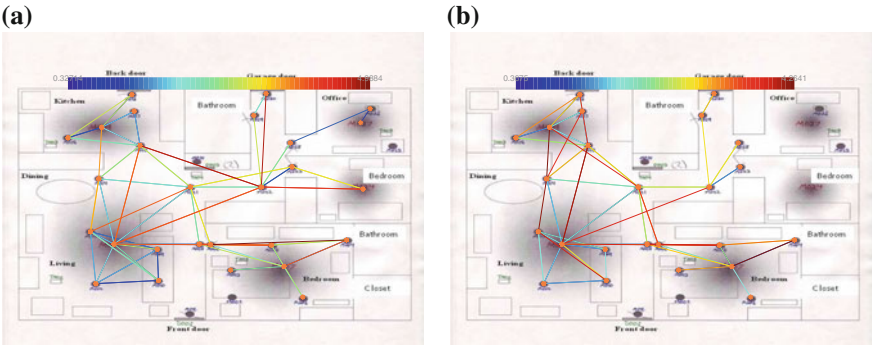


Fig. 10.22 Variation of occupant mobility between two months. **a** Occupant apartment transversals, month 2. **b** Occupant apartment transversals, month 5

correlates with an occupants health and wellbeing. For example, Fig. 10.22 shows an occupant's motion between PIR sensors for the "Aruba" dataset between two one moth periods. Most of the upper left and lower portions are similar between the two months, but on the left graph, the occupant shows more movement into and out of the office and guest bedrooms and less movement in and around the kitchen. If we used these compact representations of how an occupant moves around a home in our classification system, the atypically classified day's mobility patterns could directly aid caretakers/healthcare professionals in diagnosing mobility difficulties if movement around the apartment decrease or significantly changes.

One other extensions that we are evaluating is to determine if the motion activity in any one region or hotspot is atypical, which would future give healthcare professionals an indication as to where these atypical events are occurring. Similar work is also being planned to look for long-term shifts or changes in an occupants mobility and behavior, in order to look for more subtle changes in mobility over time.

10.5 Conclusion

Longitudinal data collection and interpretation is now becoming more practical with ambient and wearable sensors technologies. With this reality, our approach of using region occupancy patterns can be used to model behavior and to identify common patterns and atypical daily patterns of smart home occupants. We show that both region occupancy-and concurrency-based hotspots derived from simple motion sensors can be used to identify "Atypical" days and that patterns from these days can be further analyzed to look for specific causes of the "Atypical" classification. This has the potential of reducing the work load of caretakers and healthcare professionals if the occupants are elderly or otherwise need monitoring, since they only would have to evaluate the atypical days closely. Additionally, common behavioral patterns and

atypical patterns could be useful as diagnostic tools to indicate lifestyle changes for all smart home occupants, since comparisons could be made across time to see the evolution of behavior patterns.

References

1. Administration for Community Living (2012) A profile of older Americans
2. World Health Organization (2012) What are the public health implications of global ageing?
3. Hashimoto H, Matsunaga T, Tsuboi T, Ohyama Y, She JH, Amano N, Yokota S, Kobayashi H (2007) Comfortable life space for elderly—using supporting systems based on technology. In: SICE, 2007 annual conference, Sept 2007, pp 3037–3042
4. O'Brien A, Ruairi RM (2009) Survey of assistive technology devices and applications for aging in place. In: Second international conference on advances in human-oriented and personalized mechanisms, technologies, and services, CENTRIC'09, pp 7–12, Sept 2009
5. Pollack ME (2005) Intelligent technology for an aging population: the use of ai to assist elders with cognitive impairment. AI Mag 26(2):9–24
6. Tran BQ (2002) Home care technologies for promoting successful aging in elderly populations. In: Engineering in medicine and biology, 2002, 24th annual conference and the annual fall meeting of the biomedical engineering society EMBS/BMES conference, 2002, proceedings of the second joint, vol 3, Oct 2002, pp 1898–1899
7. Wang L, Wang Z, He Z, Gao X (2010) Research of physical condition monitoring system for the elderly based on zigbee wireless network technology. In: 2010 international conference on E-Health networking, digital ecosystems and technologies (EDT), vol 1, pp 32–35, Apr 2010
8. Cdc—healthy places—healthy places terminology, Mar 2014
9. Kochera A, Straight AK, Guterbock TM, AARP (Organization) (2005) Beyond 50.05: a report to the nation on livable communities: creating environments for successful aging, AARP, 2005
10. Quantified self: self knowledge through numbers, Mar 2012
11. Shi JH, Lee B, Park KS (2011) Detection of abnormal living patterns for elderly living alone using support vector data description. IEEE Trans Inf Technol Biomed 15(3):438–448
12. Tax DMJ, Duin RPW (2004) Support vector data description. Mach Learn 54(1):45–66
13. O'Brien A, McDaid K, Loane J, Doyle J, O'Mullane B (2012) Visualisation of movement of older adults within their homes based on pir sensor data. In: 2012 6th international conference on pervasive computing technologies for healthcare (PervasiveHealth), May 2012, pp 252–259
14. Cuddihy P, Weisenberg J, Graichen C, Ganesh M (2007) Algorithm to automatically detect abnormally long periods of inactivity in a home. In: Proceedings of the 1st ACM SIGMOBILE international workshop on systems and networking support for healthcare and assisted living environments, HealthNet'07, ACM, New York, pp 89–94
15. Fine BT (2009) Unsupervised anomaly detection with minimal sensing. In: Proceedings of the 47th annual southeast regional conference, ACM-SE 47, ACM, New York, pp 60:1–60:5
16. Cook DJ (2012) Learning setting-generalized activity models for smart spaces. Intell Syst IEEE 27(1):32–38
17. Rashidi P, Cook DJ (2010) Mining and monitoring patterns of daily routines for assisted living in real world settings. In: Proceedings of the 1st ACM international health informatics symposium, IHI'10, ACM, New York, pp 336–345
18. Rashidi P, Cook DJ, Holder LB, Schmitter-Edgecombe M (2011) Discovering activities to recognize and track in a smart environment. IEEE Trans Knowl Data Eng 23(4):527–539
19. Roley SS, DeLany JV, Barrows CJ, Brownrigg S, Honaker D, Sava DI, Talley V, Voelkerding K, Amini DA, Smith E, Toto P, King S, Lieberman D, Baum MC, Cohen ES, Cleveland PA, Youngstrom MJ (2008) Occupational therapy practice framework: domain & practice, 2nd edn. Am J Occup Ther 62(6):625–683

20. Jakkula VR, Cook DJ (2011) Detecting anomalous sensor events in smart home data for enhancing the living experience. In: Artificial intelligence and smarter living, volume WS-11-07 of AAAI workshops, AAAI
21. Jain G, Cook D, Jakkula V (2006) Monitoring health by detecting drifts and outliers for a smart environment inhabitant. In: Proceedings of the international conference on smart homes and health telematics
22. Kalra L, Zhao X, Soto AJ, Milios AE (2012) A two-stage corrective markov model for activities of daily living detection. In: ISAmI'12, pp 171–179
23. Lymberopoulos D, Bamis A, Savvides A (2008) Extracting spatiotemporal human activity patterns in assisted living using a home sensor network. In: Proceedings of the 1st international conference on PErvasive technologies related to assistive environments, PETRA'08, ACM, New York, pp 29:1–29:8
24. Gómez-Conde I, Olivieri DN, Vila XA, Rodriguez-Liñares L (2010) Smart telecare video monitoring for anomalous event detection. In: 2010 5th Iberian conference on information systems and technologies (CISTI), June 2010, pp 1–6
25. Cardile F, Iannizzotto G, La Rosa F (2010) A vision-based system for elderly patients monitoring. In: 2010 3rd conference on human system interactions (HSI), May 2010, pp 195–202
26. Rougier C, Meunier J, St-Arnaud A, Rousseau J (2011) Robust video surveillance for fall detection based on human shape deformation. IEEE Trans Circuits Syst Video Technol 21(5):611–622
27. Sahaf Y (2011) Comparing sensor modalities for activity recognition. PhD thesis, Washington State University
28. Breunig MM, Kriegel H-P, Ng RT, Sander J (2000) Lof: Identifying density-based local outliers. Sigmod Rec 29(2):93–104