Noname manuscript No.
(will be inserted by the editor)

Exploring Zero-Training Algorithms for Occupancy Detection based on Smart Meter Measurements

Vincent Becker · Wilhelm Kleiminger

Received: date / Accepted: date

Abstract Detecting the occupancy in households is becoming increasingly important for enabling context-aware applications in smart homes. For example, smart heating systems, which aim at optimising the heating energy, often use the occupancy to determine when to heat the home. The occupancy schedule of a household can be inferred from the electricity consumption, as its changes indicate the presence or absence of inhabitants. As smart meters become more widespread, the real-time electricity consumption of households is often available in digital form. For such data, supervised classifiers are typically employed as occupancy detection mechanisms. However, these have to be trained on data labelled with the occupancy ground truth. Labelling occupancy data requires a high effort, sometimes it even may be impossible, making it difficult to apply these methods in real-world settings. Alternatively, one could use unsupervised classifiers, which do not require any labelled data for training. In this work, we introduce and explain several unsupervised occupancy detection algorithms. We evaluate these algorithms by applying them to three publicly available datasets with ground truth occupancy data, and compare them to one existing unsupervised classifier and several supervised classifiers. Two unsupervised algorithms perform the best and we find that the unsupervised classifiers outperform the supervised ones we compared to. Interestingly, we achieve a similar classification performance on coarse-grained aggregated datasets and their fine-grained counterparts.

Keywords Occupancy detection · Smart meters · Unsupervised classification

Vincent Becker, Wilhelm Kleiminger
Department of Computer Science, ETH Zurich, Switzerland
E-mail: vincent.becker@inf.ethz.ch
V. Becker ORCiD: 0000-0003-0522-0312

1 Introduction

Occupancy, i.e. whether the inhabitants of a dwelling are at home or not, is one of the major contextual features used in smart home applications. Determining occupancy patterns as shown by the examples in figures 1a and 1b could be used to control the heating, electronic devices, the burglar alarm, etc. There are several ways to determine the presence or absence of inhabitants (cf. Section 4). A common approach is installing sensors in the dwelling, such as reed switches on the main doors or motion detectors indoors. However, these approaches are relatively obtrusive and require the installation of dedicated hardware. Another possibility is to use location-connected services on smartphones, such as the inhabitants' GPS location or the Wi-Fi networks their smartphones are connected to. However, the inhabitants would have to carry their smartphone with them at all times for this to work.

A different and promising possibility relies on monitoring the electricity consumption of the household. Previous research has shown that it is possible to detect occupancy from electrical load data using machine learning algorithms with sufficiently high accuracy [25,27,28]. Indeed, electrical load data is a good proxy for a household's occupancy since its magnitude and changes in the power consumption are indicators for human activity (i.e. interactions with appliances) in the household. At the same time, smart electricity meters, which continuously measure the electrical power demand of a household, are becoming more and more ubiquitous. In sixteen EU member countries, a smart meter penetration rate of 95% is expected by 2020 [18]. This large-scale deployment of smart meters makes it increasingly viable to use their measurements for purposes like occupancy detection.

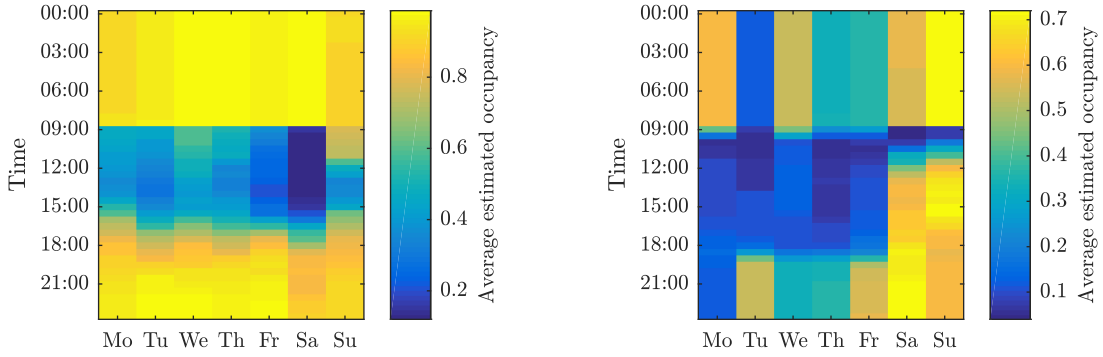

Fig. 1: Average weekly schedules for two different households. The higher the value (as displayed by the colour) in a time slot, the likelier the home is occupied during that time slot.

The task of occupancy detection in our context can be defined as follows: Given a time series of electricity consumption measurements, determine (i.e. estimate with a sufficiently high probability) whether the home is occupied or not for each time interval. We thus face a classification problem with two classes, *occupied* and *unoccupied*.

In machine learning, there are two main types of classifiers: supervised and unsupervised. Supervised classifiers have to be trained on data labelled with ground truth in order to learn the relevant patterns. In terms of our setting, electricity consumption samples labelled with the true occupancy class of the specific household would have to be provided. This labelling process entails a great effort, yielding supervised algorithms difficult to apply in a real-world scenario, where labelled data is not easily available. By contrast, unsupervised algorithms do not require any labelled samples, hence the term zero-training.

In this work we address two challenges: First, creating and exploring different unsupervised classifiers for occupancy detection, and second, coping with coarse-grained electricity consumption data with a sampling interval of half an hour. The classifiers presented are very lightweight and could easily be run locally, i.e. within the home without having to disclose information to the outside. In more detail, our contributions are:

- Developing unsupervised classifiers for occupancy detection from generally available electricity consumption data (only energy measurements, no voltages or currents).
- Being able to handle coarsely grained data at a sampling interval of 30 minutes.
- Validating and evaluating the algorithms on three publicly available datasets containing electrical energy consumption and ground truth occupancy values; further, comparing them to previous algorithms including supervised classifiers.

The remainder of the paper is structured as follows: In Section 2, we show the design of our occupancy detection algorithms. We evaluate them in Section 3. In Section 4, we discuss related work done on occupancy detection. Finally, we draw conclusions in Section 5.

2 Occupancy Detection from Electricity Consumption Data

Our aim is to determine the occupancy state of a household by analysing its electricity consumption. We assume a coarse-grained sampling interval of 30 minutes. First, we pre-process the data. We take the logarithm of all power values (to the base 10) and use a moving average filter with a window size of 5 to smooth the data. Then, we perform the classification, using one of the unsupervised classifiers mentioned below. The classifier assigns a label (either occupied or unoccupied) to each sample (i.e. 30 minutes interval). The sequence of the resulting labels is the occupancy schedule. After classification, we post-process the schedule by again performing moving average smoothing on the schedule. The whole process is depicted in Figure 2.

In the following we will detail on the three unsupervised occupancy detection algorithms we developed: a Hidden Markov Model, a model using a geometric moving average, and one using a Page-Hinkley test. We face two challenges: firstly, in a real-world scenario, there are no labels available, i.e. the electricity data is not annotated with the occupancy ground truth. Hence, the use of supervised classifiers is not possible. Secondly, we want to be able to deal with coarse-grained electricity consumption data. For a large sampling interval

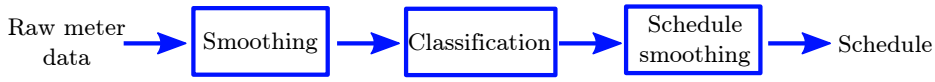

Fig. 2: The pipeline for occupancy detection.

the raw data already gives a very aggregated view of the household's electricity consumptions. Therefore, the possibilities to calculate features over time windows are limited since one feature would aggregate a long period of time. To compare the algorithms and measure their quality, we apply each of them to three datasets for which occupancy ground truth is available, i.e. it is possible to determine how well they perform (cf. Section 3).

2.1 First algorithm: Hidden Markov Model (HMM)

The occupancy of a household can be modelled as a probabilistic model with two states - *occupied* and *unoccupied*. At any point in time, there is a certain probability of the household changing from one to the other. This kind of system can be represented by a Hidden Markov Model (HMM), a statistical model, through which a process is modelled by a Markov chain with hidden states. This means that the model randomly changes from one state to another with certain transition probabilities depending only on the current state. The states cannot be observed themselves, they are hidden and instead only the states' emissions are observable which emerge with certain probabilities depending on the emitting state. In our case the HMM models the binary occupancy of a home and hence has only two states, *unoccupied* and *occupied*. The emissions are power consumption values which are drawn from emission distributions. The model is depicted in Figure 3.

Figure 4 shows an example for a household from 4 a.m. to 12 a.m. The only information we can observe are the power consumption values for each 30 minute time slot. The goal is to find the most probable state sequence to a given sequence of observations (also known as decoding), which is solved by the Viterbi algorithm [48]. Usually, the parameters of the model would be learnt using a training algorithm, e.g. the Baum-Welch algorithm [38] following an expectation-maximisation approach. HMMs have been used for occupancy detection from electricity data in this supervised manner before [25, 27, 28]. For that however, training data would have to be available. Thus, we determine the emission distribution and transition probability on basic assumptions we make, which we detail on in the following.

Transition probabilities The transition probabilities define how probable it is in each time slot for the state to change from unoccupied to occupied or from occupied to unoccupied, respectively. In our method a single day has 48 half-hour time slots. The transition probabilities depend on the expectation how long the home is unoccupied or occupied, respectively and how many "leave" and "return" events there are. We calculate $\alpha = \frac{\#return}{\#unoccupied}$ and $\beta = \frac{\#leave}{\#occupied} = \frac{\#leave}{48-\#unoccupied}$. Since it is the most common, we assume a typical working day schedule in which the home is unoccupied nine continuous hours a day with a single "leave" event and a single "return" event. Hence there are 30 occupied and 18 unoccupied slots out of the 48 slots per day in total. Thus we calculate the transition probabilities by $\alpha = \frac{1}{18}$ and $\beta = \frac{1}{30}$. If further knowledge such as a rough estimation of the schedule of the household was available, this could be easily adjusted.

Emission probabilities For the emission probabilities we have to find a set of samples which we assume to belong to the unoccupied and the occupied state, respectively. To do so we use the mean over all power values from the household's data as threshold and assume that all samples below the mean belong to the unoccupied state and all above or equal to the occupied state. In our experiments the mean has proven to be a good heuristic for a threshold to separate occupied from unoccupied emissions. For each sample set we fit a normal distribution and use this as the emission distribution.

2.2 Second algorithm: Geometric Moving Average (GeoMA)

The motivation behind this strategy is that periods of absence will decrease the moving average of the electricity consumption. As soon as the inhabitants are home again the consumption will increase. The average will too, but naturally not as fast. Hence, the momentary electrical consumption will rise above the average and we will signal occupancy.

Implementing this idea, the algorithm using the geometric moving average follows a simple strategy. In each time step we calculate the geometric moving average. If the current sample is greater than the average, we set the schedule for the current time slot to *occupied*, otherwise to *unoccupied*. The procedure is shown in Appendix A and Figure 5 depicts an example for a single day.

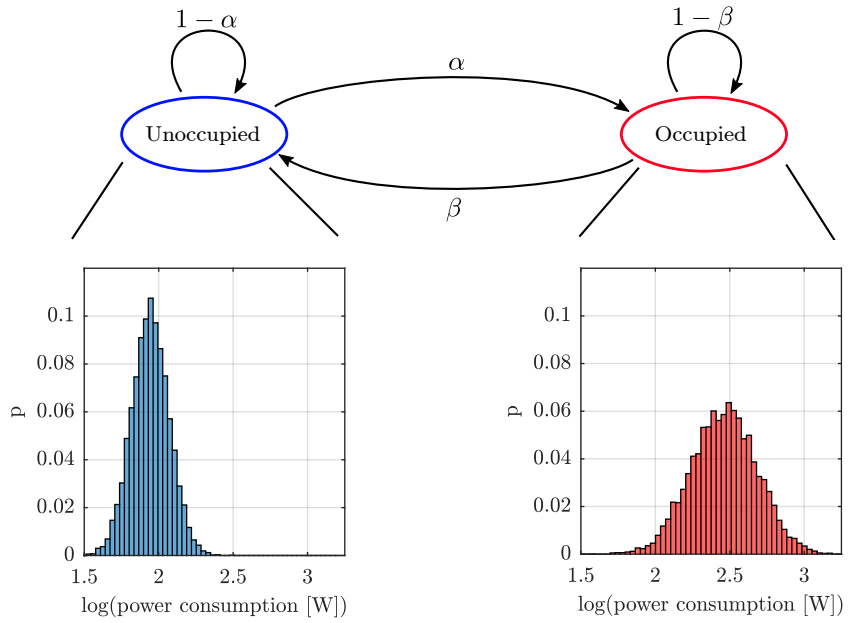

Fig. 3: The HMM for occupancy detection. Each state emits power values from a certain emission probability distribution and the transition from one state to the other takes place with a certain probability in each step.

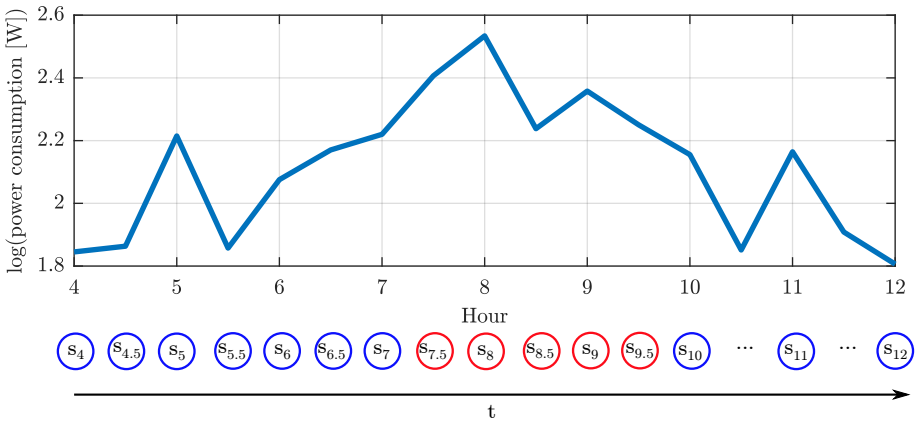

Fig. 4: An example of the 30 minutes aggregated power consumption for a specific household from 4 a.m. to 12 a.m. The states are either occupied or unoccupied. The colours (red implying occupied and blue unoccupied) are a guess for a schedule and depict the non-trivial task of estimating the state sequence.

2.3 Third algorithm: Page-Hinkley Test (PHT)

The Page-Hinkley test [34] is an unsupervised concept change detection algorithm. In the area of data streams (a potentially unbounded sequence of data points, such as our power consumption values), the concept is considered as the probability distribution generating the stream data. In our case we can imagine two concepts, the unoccupied and occupied home, which incorporate two different distributions emitting the data. The aim is to find the changes, i.e. when the stream process moves from one conept (i.e. distribution) to the other, which corresponds to the change from unoccupied to occupied or vice versa. The Page-Hinkley test detects changes in signals by observing the difference of cumulative variables from adapting averages. Basically, it is a more sophisticated version of the geometric moving average explained above. The procedure, fitted to our application, is shown in Appendix A. The Page-Hinkley test can detect increasing and decreasing changes. If we find an increasing change, we set the schedule for that slot to 1, i.e. occupied, for a decreasing change to 0, i.e. unoccupied. In case we detect no change we set the slot to the state in the previous slot.

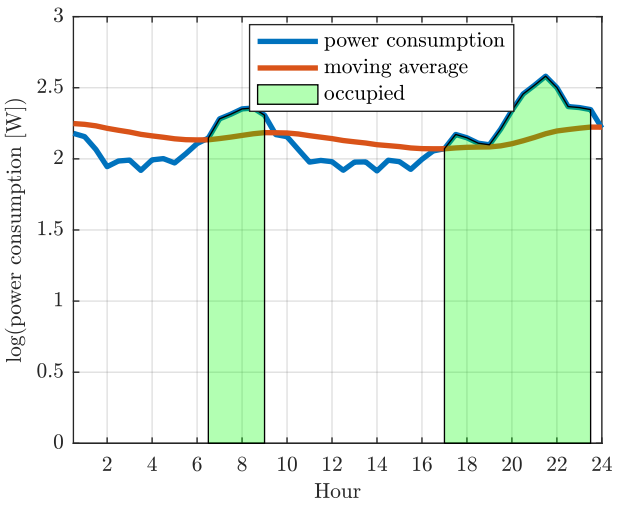

Fig. 5: An example for the geometric moving average, marked as the red line. Whenever the momentary electricity consumption (shown as the blue line) is higher than the geometric moving average, the household is considered to be occupied. The estimation for the given data is shown by the green areas.

2.4 NIOM (Non-Intrusive Occupancy Monitoring)

NIOM [14] was presented by Chen et al. (cf. Section 4). We include it in our comparison of the three previously mentioned methods (and the supervised algorithms). NIOM calculates three features over a time window. These are the average, the standard deviation, and the maximum range of values in the window. The home is considered to be *occupied* in a specific time slot if one or more of the features are above a certain threshold. The thresholds are determined by the maxima of the features during the previous night. Thereby, the thresholds are dynamic. If any two time slots are detected to be *occupied* and are within a certain window $\tau_{cluster}$, then all slots in between are set to *occupied*. If the home was *occupied* in the evening, then it is also considered to be *occupied* during the night.

2.5 Adding a Nightly Schedule

Detecting nightly occupancy merely from electricity consumption data is not a simple task, since during sleep people do not interact with electrical devices and most of them are turned off or in standby mode. Similar to the authors of [14] we resort to an additional simple rule-based approach by adding a nightly schedule when applying any of our algorithms. If we detect occupancy with a duration of at least one hour from 8 p.m. to midnight, we set the state of each time slot for the following night (until 9 a.m.) to occupied, beginning with the slot which is the last to be detected as occupied.

3 Validation and Evaluation

We test our three unsupervised algorithms HMM, GeoMA, and PHT on three publicly available labelled datasets to be able to asses their performance. Due to the high effort and costs of annotating the data, such datasets are relatively rare. We downsampled each dataset by averaging to a sampling interval of half an hour to show we can indeed handle coarse-grained data. This downsampling does not impair the performance of the classification in most cases. Rather, it is often improved (cf. Section 3.5).

For comparison we also apply NIOM and three supervised algorithms, k-Nearest Neighbours ($k = 5$) (KNN), a Support Vector Machine (SVM) with an RBF-kernel, and a random forest (RF). Additionally we show a baseline, which assumes that the home was occupied in every time slot. The baseline is a lower bound for the performance the other classifiers should achieve. For all supervised classifiers we use the standard implementations in MATLAB. For evaluation, we use 10-fold cross-validation (90% training and 10% testing). The features we use for the supervised classifiers are the mean, the standard deviation, the sum of absolute differences, and the maximum of the difference in a time window of two observations. Naturally, these supervised algorithms cannot be applied in real-world scenarios without any prior training. For GeoMA we set the adaptation rate $\lambda = 0.05$. For PHT we use 0.05 for the magnitude threshold and 0.3 for the detection threshold.

As metrics for the performance of a classifier we use the accuracy (ACC) and the Matthews correlation coefficient (MCC, [32]), which are defined in equations 1 and 2. The bounds for the accuracy are 0 and 1, for the MCC -1 and 1. In both cases a higher number indicates a better result. The ACC result may be misleading in case of an unbalanced class distribution (as it is in our case, since the homes are occupied more than they are unoccupied), whereas the MCC has the advantage that it compensates for skewed classes.

$$ACC = \frac{\mathrm{tp} + \mathrm{tn}}{\mathrm{tp} + \mathrm{tn} + \mathrm{fp} + \mathrm{fn}} \tag{1}$$

$$MCC = \frac{\mathrm{tp} * \mathrm{tn} - \mathrm{fp} * \mathrm{fn}}{\sqrt{(\mathrm{tp} + \mathrm{fp})(\mathrm{tp} + \mathrm{fn})(\mathrm{tn} + \mathrm{fp})(\mathrm{tn} + \mathrm{fn})}} \tag{2}$$

The arguments are the number of true positives (tp), false positives (fp), true negatives (tn), and false negatives (fn). If the classifier assigns all samples to one

class, the *MCC* cannot be calculated and thus will not be shown in the results in such cases.

3.1 Dataset A: Tang et al.

The dataset is presented in [43]. It contains the electricity and occupancy data over one month for a single household in Victoria, BC, Canada. The consumption data was collected using off-the-shelf measuring devices and the occupancy information was derived from the GPS traces of the inhabitants' mobile phones. The collection period was from the 23rd February 2015 to 23rd March 2015. The original sampling frequency was 0.1 Hz. Table 1 displays the results for the detection algorithms. The HMM, the PHT, and the GeoMA perform

Table 1: The average results of Tang's dataset.

Algorithm	ACC	MCC
Baseline	0.65	-
HMM	0.89	0.76
GeoMA	0.90	0.78
PHT	0.89	0.75
NIOM	0.85	0.67
KNN	0.70	0.33
SVM	0.65	-
RF	0.70	0.33

the best, even better than the supervised algorithms. The baseline assumes the house to be occupied all the time. Thus, there are no true and false negatives and the *MCC* cannot be calculated. Since the occupancy is relatively high (65%), the SVM seems to have learnt to always classify a slot as occupied, i.e. it behaves just like the baseline.

3.2 Dataset B: Chen et al.

Chen's dataset [14] is part of the Smart* dataset [6] augmented with occupancy information, which are again obtained from the GPS traces of the inhabitants' smartphones. It contains three parts, spring (1st April 2013 to 7th April 2013) and summer (8th July 2013 to 14th July 2013) measurements for one house, and only summer measurements for another. The original sampling interval is 1 minute. The households both are in Western Massachusetts, US. Figure 6 show the results on this dataset and the averages are in Table 2. In terms of accuracy, all algorithms perform similarly well. For the MCC metric, however, the HMM and GeoMA are the best and NIOM, the algorithm which was evaluated on this dataset in [14], is outperformed.

Table 2: The average results over the three parts of Chen's datast.

Algorithm	ACC	MCC
Baseline	0.77	-
HMM	0.90	0.73
GeoMA	0.91	0.73
PHT	0.89	0.68
NIOM	0.88	0.55
KNN	0.88	0.63
SVM	0.85	-
RF	0.90	0.69

3.3 Dataset C: ECO Dataset

The ECO dataset presented in [7] contains the data of six households in Thun, Switzerland. It was collected over a period of eight months from June 2012 to January 2013. The data is split into two periods, summer and winter. The sixth household did not provide any occupancy information so we omit it here. To create occupancy ground truth the inhabitants manually registered presence and absence with a tablet. Additionally, a PIR sensor near the main door and several smart plugs were deployed in each household, connected to devices such as PCs, etc., to enhance the annotation. The original sampling rate is 1 Hz. Except household two, all households are occupied nearly all the time, hence it makes it difficult to perform better than the baseline. Figure 7 shows the results on this dataset and the averages are in Table 3.

Table 3: The average results over the ECO dataset.

Algorithm	ACC	MCC
Baseline	0.82	-
HMM	0.69	0.20
GeoMA	0.70	0.20
PHT	0.68	0.14
NIOM	0.76	0.17
KNN	0.81	-
SVM	0.83	-
RF	0.82	-

The authors of [28] have already shown the difficulty of beating the baseline for this dataset, even with 1 Hz data and supervised learning algorithms. Here we work on 30 minute sampling intervals and our methods do not compute features. The supervised algorithms perform the best, but often make predictions similar to the baseline, estimating occupancy for nearly all time slots. Among the unsupervised algorithms, the HMM and the GeoMA perform the best in terms of MCC.

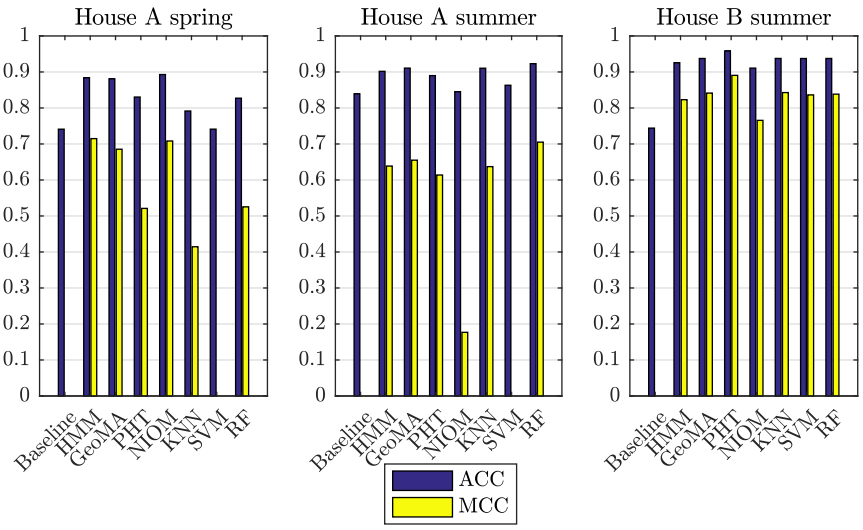

Fig. 6: The results on the dataset of Chen et al.

3.4 Overall Results

In terms of MCC performance, which is a more appropriate performance measure than ACC, the HMM and the GeoMA perform the best with the PHT as close runner-up. Note that the disadvantage of the HMM is that we need all the data prior to detection, while the GeoMA and the PHT work online. Whether the achieved classification performance is sufficient clearly depends on the application. We believe it is sufficient for many non-critical systems, such as an occupancy controlled heating or lighting system, which could be easily overruled by humans in case of false classification. In addition, since our algorithms achieve a significantly better performance than a random guess, they certainly would be beneficial when combined with with other occupancy techniques to create an ensemble which achieves higher performance.

3.5 Performance with Higher Sampling Rates

As mentioned before, we downsampled the datasets to a common sampling interval of 30 minutes. To show that this did not strongly decrease the performance we also evaluated the original datasets. The original sampling rates were 0.1 Hz for the dataset of Tang et al., once per minute for the dataset of Chen et al., and 1 Hz for the ECO dataset. We run the HMM on the original datasets to be able to compare to the downsampled version. The average classification performance for each dataset is shown in Table 4.

The results for the original and the downsampled version are similar for Tang's dataset. For Chen's dataset the results are better in the downsampled version. Only for the ECO dataset the results are significantly better using the original dataset.

Table 4: The results of the HMM on the original datasets with higher sampling rates. The values show the average performance for each of the datasets.

Dataset	ACC	MCC
A (Tang)	0.74	0.78
B (Chen)	0.78	0.68
C (ECO)	0.75	0.38

4 Related Work on Occupancy Detection

In this section we discuss the related work that has been done in the effort of detecting the occupancy in households. There are several approaches for occupancy detection, differing both conceptually and technologically.

4.1 Using the Inhabitants' Smartphones

One way to detect the occupancy of a household is to augment its members with devices which sense their location. This location information can then be used to check whether the inhabitants are at home or not. A device with localisation capabilities which many people already possess is a smartphone.

Gupta et al. use GPS (Global Positioning System) information to calculate the inhabitants' distance to their

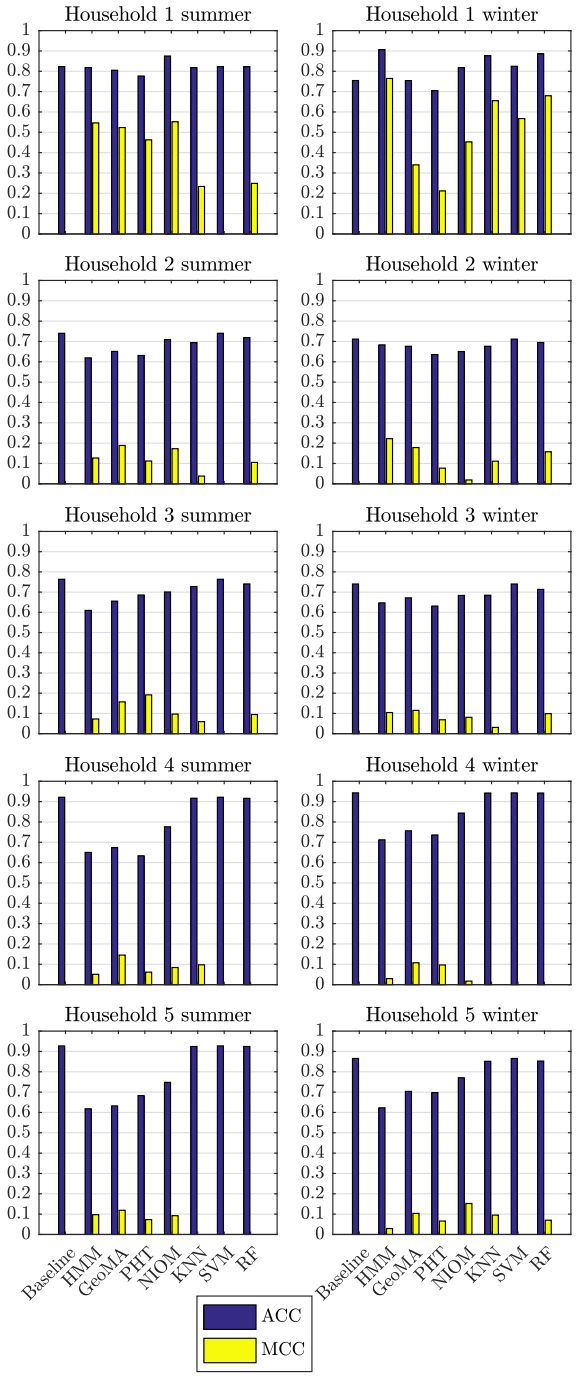

Fig. 7: The results on the ECO dataset.

home and employ a GPS-enabled thermostat to control the home's heating based on that distance [22]. Thereby, the home can be reheated prior to the inhabitants' expected return. In that sense the systems actually applies occupancy *prediction* which naturally can be used for occupancy *detection*.

In the homeset algorithm, Kleiminger et al. exploit the WLAN information sensed with an inhabitant's smartphone [26]. In an initial stage a set of Wi-Fi networks, which are reachable from the home, is created. Whenever one of the networks in the set is reachable from the inhabitant's smartphone, the inhabitant is considered to be at home, otherwise not.

Smith et al. use RFID tags on objects and a user-worn wristband to identify which activity a human is performing [40].

A disadvantage of these approaches is that the inhabitants have to carry their smartphones or other devices with them and that the location service has to be turned on at all times.

4.2 Using Sensors Inside the Home

Often, a variety of sensors inside the home is used to directly detect the occupancy of a building or even single rooms.

In their review, Guo et al. examine different occupancy detection sensors such as PIR sensors, ultra-sonic sensors, micro-wave sensors, light barriers, and video cameras in order to be able to control the lighting of a building [21]. In their approach "Smart Thermostat", Lu et al. use PIR sensors in rooms and magnetic reed switches on the main door of the home to create features [31]. With these features, a Hidden Markov model is applied to infer the occupancy states, either being Active, Away, or Sleep. Using the occupancy information, the HVAC system of the house is controlled. To be able to preheat the home in time before the occupants arrive home again in order to avoid a loss in comfort, an occupancy schedule is set up incorporating historical occupancy data to predict arrival times. Hence the method is a mixture of occupancy detection and prediction.

Soltanaghaei et al. present WalkSense, a system consisting of motion sensors distributed along the walkways of a home [41]. Brown et al. use an ultra-wideband radar module to sense occupancy [10]. Woodstock et al. employ time-of-flight sensors to detect occupancy [50]. Wang et al. measure the indoor CO_2 levels to detect if someone is present [49]. Sensing changes in the CO_2 levels is a common approach in many other works [11, 15, 23, 42], also in combination with other environmental values, such as temperature [16,17], light and humidity [12] and others [4, 37]. Zikos et al. explore conditional random fields for fusing different combination of sensors including CO_2, motion, and acoustic sensors [52]. Amayri et al. calculate information gains to determine the most useful measurements [3]. Teixeira et al. present an approach to determine the number of people in a room

using a camera sensor network [45]. Gao et al. present a self-programmable thermostat [20]. The leaving and arrival time-points of the inhabitants are registered using simple sensors and with that data an automatic heating schedule is defined. Similarly, Barbato et al. create user profiles from data gathered with a wireless sensor network in order to optimize energy consumption [5]. Brackney et al. present an image processing occupancy sensor which analyses video data aiming to detect humans in the pictures [9]. It overcomes several disadvantages of PIR and ultra-sonic based motion sensors such as detecting humans who are not in motion. Furthermore it is possible to determine the number of people present. However, a video analysis system has severe privacy issues.

Moreover there are many other approaches which do not aim at occupancy detection directly but could be applied for that purpose. Patel et al. sense differences in air pressure when doors are opened and closed using only a single sensor in a HVAC unit [35]. The authors in [36] detect and classify electrical events by their pattern in the power line as occupants trigger switches using plug-in sensors. Froehlich et al. sense the water activity, e.g. the use of the kitchen sink, shower, or toilet with a single-point sensor [19]. Besides, there are more approaches for human activity recognition in homes using several types of sensors [44, 47]. There also are commercial providers for energy management systems using sensors for occupancy detection, such as motion or body heat sensors, in order to control the temperature of the home [46].

However, all these approaches entail the installation of sensors in the home.

4.3 Using Electricity Data form Smart Meters

As smart meters become more wide-spread, it is attractive to use their generated electricity data to infer the household occupancy. Basically, the smart meter is employed as a sensor in this approach, however, the advantage is that the sensor is not installed solely for the purpose of occupancy detection and the inhabitants are not required to carry any devices with them.

When employing algorithms utilising energy consumption data, a common approach is to use supervised classification to determine the occupancy states. This means that labelled training data is necessary, i.e. the occupancy ground truth needs to be obtained. Usually, an individual classifier is trained for each household. Yang et al. use conditional random fields to determine the number of people present in a household [51]. As features the peak, mean, and variance for short time intervals were chosen. Moreover, they use supervised classifiers such as a random forest, a decision tree, KNN, NaiveBayes, and an MLP for the binary occupancy case and are able to reach accuaries up to 98% against a baseline of 88%. Akbar et al. apply supervised machine learning methods such as KNN and SVM to detect if office desks are occupied [1]. For this, they use energy meters at each desk. As features they use the real power, the root mean square of voltage and current, and the phase angle between them. They extend the binary setting and add a standby state to model short breaks of the people. Kleiminger et al. also use supervised classifiers such as SVM, KNN, and a HMM to detect occupancy on the ECO dataset (cf. Section 3) [25, 27, 28]. They achieve accuracies of more than 80%. Chaney et al. use an HMM as well, but combine electricity with sensed CO_2 levels and dew point temperature utilising the Dempster-Shafer theory for sensor fusion [13]. Boait et al. exploit the electricity load and hot water usage data to infer occupancy [8]. Additionally, they incorporate occupancy data from the previous week to set a prior probability of occupancy. This information is combined using the Bayes rule to infer the a-posteriori probability of the home being occupied and hence to control the heating system. Jin et al. try to reduce the need for supervision by applying transfer learning, i.e. using supervised classifiers trained on similar households [23].

Another method of analysing electrical data is Non Intrusive Load Monitoring (NILM, [7]). The aim is to determine when single appliances are turned on or off from the aggregate electricity data. From the appliances' states, the occupancy of a home could be derived. However, NILM methods usually rely on high sampling rates and need additional information about the appliances. Alhamoud et al. use appliance-level power consumption to derive human behavioural patterns [2].

The requirement of having to obtain ground truth data for supervised classification is a problem for the application of occupancy detection in practice. However, unsupervised classification is in principle more difficult, since the household's consumption patterns are unknown. Chen et al. present their threshold-based algorithm NIOM (Non-Intrusive Occupancy Monitoring, cf. Section 2.4) which signals occupancy as soon as one of the features exceeds its corresponding threshold [14]. As features they use the mean, the standard deviation, and the maximum range of the electricity power data. The thresholds are calculated as the maximum of the features during the previous night.

Jin et al. present PresenceSense [24], a zero-training algorithm for occupancy detection in office buildings based on plug loads. The algorithm uses a vague working schedule of the participant as initial estimate and iteratively refines the assessment through classifiers learning

from the predictions in the previous iteration. As features they use the discrete power levels, the maximum absolute change, the mean of absolute difference, the mean of the length of changes, and the standard deviation. They state that a sampling interval of one minute is sufficient, a relatively high rate compared to half an hour in our case.

Tang et al. present a framework named SHARK which requires no training. It models a household's appliances' states [43]. The mode states are decoded by solving an optimization problem. From the decoded state, the occupancy of the household is inferred. Although the algorithm needs no training, knowledge about the appliances within the house is necessary for the decoding step. Also the approach cannot be used online, since the optimisation process takes place over periods of time in the past.

A comparative study on occupancy prediction algorithms using electricity data can be found in [29]. Further occupancy prediction algorithms based on other principles among others are Preheat [39], Neurothermostat [33], or Presence Probabilities [30].

5 Conclusions

Our aim in this work was to advocate unsupervised classification algorithms for occupancy detection in private households which use the electricity consumption data measured by smart meters. One specific objective we approached was to examine the suitability of coarse-grained consumption data relative to fine-grained consumption data. We evaluated the performance of our algorithms on three datasets containing ground truth occupancy information. Among the algorithms we presented are also online algorithms, which are ready to be used in a real scenario. Besides, all algorithms we showed require little computational power and can easily be run inside the home. The best performing algorithms showed an accuracy of 69% to 90%, or an MCC of 0.20 to 0.78. In general we found that our unsupervised (i.e. zero-training) algorithms compare favourably to supervised algorithms and that the use of coarse-grained data is comparable to the use of fine-grained data to the performance in two out of three test cases.

A Algorithms

Algorithm GeoMA: t represents the timesteps, x is an array containing the electricity consumption values. Each entry represents a 30 minute window of data. Accordingly, *schedule* is an array containing the resulting occupancy estimations for each time slot. The parameter λ determines the importance of recent values over older values, i.e. $(1 - \lambda)$ determines the decay of the average.

```
function GeoMA(x, λ)
    average ← x(1)
    for all t do
        average ← λ ∗ x(t) + (1 − λ)∗ average
        if x(t) ≥ average then
            schedule(t) ← 1
        else
            schedule(t) ← 0
        end if
    end for
    return schedule
end function
```

Algorithm PHT: t represents the timesteps, x is an array containing the electricity consumption values. Each entry represents a 30 minute window of data. Accordingly, *schedule* is an array containing the resulting occupancy estimations for each time slot.

```
function PHT(x, magThreshold, detectThreshold)
    currentState ← 0
    for all t do
        deviation ← x(t) − x̄− magThreshold
        mt ← mt + deviation
        increasingMT ← min(increasingMT, mt)
        decreasingMT ← max(decreasingMT, mt)
        increasingPHT ← mt − increasingMT
        decreasingPHT ← decreasingMT − mt
        if increasingPHT > detectThreshold then
            schedule(t) ← 1
            currentState ← 1
            mt ← 0
        else if decreasingPHT > detectThreshold then
            schedule(t) ← 0
            currentState ← 0
            mt ← 0
        else
            schedule(t) ← currentState
        end if
    end for
    return schedule
end function
```

References

1. A. Akbar, M. Nati, F. Carrez, and K. Moessner. Contextual occupancy detection for smart office by pattern recognition of electricity consumption data. In *2015 IEEE Int. Conf. on Communications (ICC)*, pages 561–566, June 2015. doi:10.1109/ICC.2015.7248381.
2. A. Alhamoud, P. Xu, F. Englert, A. Reinhardt, P. Scholl, D. Boehnstedt, and R. Steinmetz. Extracting human behavior patterns from appliance-level power consumption data. In *Proc. 12th European Conf. on Wireless Sensor Networks, EWSN 2015, Porto, Portugal*, pages 52–67, 2015. doi:10.1007/978-3-319-15582-1_4.
3. M. Amayri, A. Arora, S. Ploix, S. Bandhyopadyay, Q.-D. Ngo, and V. R. Badarla. Estimating occupancy in heterogeneous sensor environment. *Energy and Buildings*, 129:46 – 58, 2016. doi:10.1016/j.enbuild.2016.07.026.
4. O. Ardakanian, A. Bhattacharya, and D. Culler. Non-intrusive techniques for establishing occupancy related energy savings in commercial buildings. In *Proc. 3rd ACM Int. Conf. on Systems for Energy-Efficient Built Environments*, BuildSys '16, pages 21–30, 2016. doi:10.1145/2993422.2993574.
5. A. Barbato, L. Borsani, A. Capone, and S. Melzi. Home energy saving through a user profiling system based on wireless sensors. In *Proc. 1st ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, BuildSys '09, pages 49–54, New York, NY, USA, 2009. ACM. doi:10.1145/1810279.1810291.
6. S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy, and J. Albrecht. Smart*: An open data set and tools for enabling research in sustainable homes. In *Proc. 2012 Workshop on Data Mining Applications in Sustainability (SustKDD 2012)*, Beijing, China, 2012.
7. C. Beckel, W. Kleiminger, R. Cicchetti, T. Staake, and S. Santini. The ECO data set and the performance of non-intrusive load monitoring algorithms. In *Proc. 1st ACM Conf. on Embedded Systems for Energy-Efficient Buildings*, BuildSys '14, pages 80–89, New York, NY, USA, 2014. ACM. doi:10.1145/2674061.2674064.
8. P. J. Boait and R. M. Rylatt. A method for fully automatic operation of domestic heating. *Energy and Buildings*, 42(1):11–16, 2010. doi:10.1016/j.enbuild.2009.07.005.
9. L. J. Brackney, A. R. Florita, A. C. Swindler, L. G. Polese, and G. A. Brunemann. Design and performance of an image processing occupancy sensor. In *The Second Int. Conf. on Building Energy and Environment 2012*, pages 987–994, Boulder, Colorado, USA, 2012.
10. R. Brown, N. Ghavami, H.-U.-R. Siddiqui, M. Adjrad, M. Ghavami, and S. Dudley. Occupancy based household energy disaggregation using ultra wideband radar and electrical signature profiles. *Energy and Buildings*, 141:134–141, 2017. doi:10.1016/j.enbuild.2017.02.004.
11. D. Calì, P. Matthes, K. Huchtemann, R. Streblow, and D. Müller. CO2 based occupancy detection algorithm: Experimental analysis and validation for office and residential buildings. *Building and Environment*, 86:39–49, 2015. doi:10.1016/j.buildenv.2014.12.011.
12. L. M. Candanedo and V. Feldheim. Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models. *Energy and Buildings*, 112:28–39, 2016. doi:10.1016/j.enbuild.2015.11.071.
13. J. Chaney, E. H. Owens, and A. D. Peacock. An evidence based approach to determining residential occupancy and its role in demand response management. *Energy and Buildings*, 125:254–266, 2016. doi:10.1016/j.enbuild.2016.04.060.
14. D. Chen, S. Barker, A. Subbaswamy, D. Irwin, and P. Shenoy. Non-intrusive occupancy monitoring using smart meters. In *Proc. 5th ACM Workshop on Embedded Systems for Energy-Efficient Buildings*, BuildSys'13, pages 9:1–9:8, New York, NY, USA, 2013. ACM. doi:10.1145/2528282.2528294.
15. A. Ebadat, G. Bottegal, M. Molinari, D. Varagnolo, B. Wahlberg, H. Hjalmarsson, and K. H. Johansson. Multi-room occupancy estimation through adaptive gray-box models. In *Proc. 4th IEEE Conf. on Decision and Control (CDC)*, pages 3705–3711, 2015. doi:10.1109/CDC.2015.7402794.
16. A. Ebadat, G. Bottegal, D. Varagnolo, B. Wahlberg, H. Hjalmarsson, and K. H. Johansson. Blind identification strategies for room occupancy estimation. In *2015 European Control Conf. (ECC)*, pages 1315–1320, 2015. doi:10.1109/ECC.2015.7330720.
17. A. Ebadat, G. Bottegal, D. Varagnolo, B. Wahlberg, and K. H. Johansson. Regularized deconvolution-based approaches for estimating room occupancies. *IEEE Transactions on Automation Science and Engineering*, 12(4):1157–1168, 2015. doi:10.1109/TASE.2015.2471305.
18. European Commission. Cost-benefit analyses & state of play of smart metering deployment in the EU-27. Technical Report 52014SC0189, EU Commission, Brussels, 2014. URL: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52014SC0189.
19. J. Froehlich, E. C. Larson, T. Campbell, C. Haggerty, J. Fogarty, and S. N. Patel. Hydrosense: infrastructure-mediated single-point sensing of whole-home water activity. In *Proc. 11th Int. Conf. on Ubiquitous Computing, UbiComp 2009, Orlando, Florida, USA*, pages 235–244, 2009. doi:10.1145/1620545.1620581.
20. G. Gao and K. Whitehouse. The self-programming thermostat: Optimizing setback schedules based on home occupancy patterns. In *Proc. 1st ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, BuildSys '09, pages 67–72, New York, NY, USA, 2009. ACM. doi:10.1145/1810279.1810294.
21. X. Guo, D. Tiller, G. Henze, and C. Waters. The performance of occupancy-based lighting control systems: A review. *Lighting Research and Technology*, 42(4):415–431, 2010. doi:0.1177/1477153510376225.
22. M. Gupta, S. S. Intille, and K. Larson. Adding GPS-control to traditional thermostats: An exploration of potential energy savings and design challenges. In *Proc. 7th Int. Conf. on Pervasive Computing, Pervasive 2009, Nara, Japan*, pages 95–114, 2009. doi:10.1007/978-3-642-01516-8_8.
23. M. Jin, N. Bekiaris-Liberis, K. Weekly, C. J. Spanos, and A. M. Bayen. Occupancy detection via environmental sensing. *IEEE Transactions on Automation Science and Engineering*, PP(99):1–13, 2016. doi:10.1109/TASE.2016.2619720.
24. M. Jin, R. Jia, Z. Kang, I. C. Konstantakopoulos, and C. J. Spanos. PresenceSense: Zero-training algorithm for individual presence detection based on power monitoring. In *Proc. 1st ACM Conf. on Embedded Systems for Energy-Efficient Buildings*, BuildSys '14, pages 1–10, New York, NY, USA, 2014. ACM. doi:10.1145/2674061.2674073.
25. W. Kleiminger. *Occupancy Sensing and Prediction for Automated Energy Savings*. PhD thesis, ETH Zurich, 2015. doi:10.3929/ethz-a-010450096.
26. W. Kleiminger, C. Beckel, A. K. Dey, and S. Santini. Using unlabeled Wi-Fi scan data to discover occupancy patterns of private households. In *Proc. 11th ACM Conf. on Embedded Network Sensor Systems, SenSys '13, Rome, Italy*, pages 47:1–47:2, 2013. doi:10.1145/2517351.2517421.

27. W. Kleiminger, C. Beckel, and S. Santini. Household occupancy monitoring using electricity meters. In *Proc. 2015 ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 975–986, 2015. doi:10.1145/2750858.2807538.
28. W. Kleiminger, C. Beckel, T. Staake, and S. Santini. Occupancy detection from electricity consumption data. In *Proc. 5th ACM Workshop on Embedded Systems for Energy-Efficient Buildings*, BuildSys'13, pages 10:1–10:8, New York, NY, USA, 2013. ACM. doi:10.1145/2528282.2528295.
29. W. Kleiminger, F. Mattern, and S. Santini. Predicting household occupancy for smart heating control: A comparative performance analysis of state-of-the-art approaches. *Energy and Buildings*, 85:493–505, 2014. doi:10.1016/j.enbuild.2014.09.046.
30. J. Krumm and A. J. B. Brush. Learning time-based presence probabilities. In *Proc. 9th Int. Conf. on Pervasive Computing, Pervasive 2011, San Francisco, CA, USA*, pages 79–96, 2011. doi:10.1007/978-3-642-21726-5_6.
31. J. Lu, T. Sookoor, V. Srinivasan, G. Gao, B. Holben, J. Stankovic, E. Field, and K. Whitehouse. The smart thermostat: Using occupancy sensors to save energy in homes. In *Proc. 8th ACM Conf. on Embedded Networked Sensor Systems*, SenSys '10, pages 211–224, New York, NY, USA, 2010. ACM. doi:10.1145/1869983.1870005.
32. B. W. Matthews. Comparison predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405(2):442–451, 1975. doi:10.1016/0005-2795(75)90109-9.
33. M. Mozer, L. Vidmar, and R. H. Dodier. The neurothermostat: Predictive optimal control of residential heating systems. In *Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA*, pages 953–959, 1996.
34. E. Page. Continuous inspection schemes. *Biometrika*, 41(1-2):100–115, 1954. doi:10.1093/biomet/41.1-2.100.
35. S. N. Patel, M. S. Reynolds, and G. D. Abowd. Detecting human movement by differential air pressure sensing in HVAC system ductwork: An exploration in infrastructure mediated sensing. In *Proc. 6th Int. Conf. on Pervasive Computing, Pervasive '08*, pages 1–18, 2008. doi:10.1007/978-3-540-79576-6_1.
36. S. N. Patel, T. Robertson, J. A. Kientz, M. S. Reynolds, and G. D. Abowd. At the flick of a switch: Detecting and classifying unique electrical events on the residential power line. In *Proc. 9th Int. Conf. on Ubiquitous Computing, UbiComp '07, Innsbruck, Austria*, pages 271–288, 2007. doi:10.1007/978-3-540-74853-3_16.
37. T. H. Pedersen, K. U. Nielsen, and S. Petersen. Method for room occupancy detection based on trajectory of indoor climate sensor data. *Building and Environment*, 115:147–156, 2017. doi:10.1016/j.buildenv.2017.01.023.
38. L. Rabiner and B. Juang. An introduction to Hidden Markov Models. *IEEE ASSP Magazine*, 3:Appendix 3A, 1986. doi:10.1109/MASSP.1986.1165342.
39. J. Scott, A. Bernheim Brush, J. Krumm, B. Meyers, M. Hazas, S. Hodges, and N. Villar. Preheat: Controlling home heating using occupancy prediction. In *Proc. 13th Int. Conf. on Ubiquitous Computing*, UbiComp '11, pages 281–290, New York, NY, USA, 2011. ACM. doi:10.1145/2030112.2030151.
40. J. R. Smith, K. P. Fishkin, B. Jiang, A. Mamishev, M. Philipose, A. D. Rea, S. Roy, and K. Sundara-Rajan. RFID-based techniques for human-activity detection. *Communications ACM*, 48(9):39–44, 2005. doi:10.1145/1081992.1082018.
41. E. Soltanaghaei and K. Whitehouse. Walksense: Classifying home occupancy states using walkway sensing. In *Proc. 3rd ACM Int. Conf. on Systems for Energy-Efficient Built Environments*, BuildSys '16, pages 167–176, 2016. doi:10.1145/2993422.2993576.
42. A. Szczurek and M. Maciejewska. Detection of occupancy events from indoor air monitoring data. In *3rd Int. Conf. on Mathematics and Computers in Sciences and in Industry (MCSI)*, pages 229–234, 2016. doi:10.1109/MCSI.2016.050.
43. G. Tang, K. Wu, J. Lei, and W. Xiao. The meter tells you are at home! Non-intrusive occupancy detection via load curve data. In *2015 IEEE Int. Conf. on Smart Grid Communications (SmartGridComm)*, pages 897–902, Miami, FL, USA, Nov 2015. doi:10.1109/SmartGridComm.2015.7436415.
44. E. M. Tapia, S. S. Intille, and K. Larson. Activity recognition in the home using simple and ubiquitous sensors. In *Proc. 2nd Int. Conf. on Pervasive Computing, Pervasive '04, Vienna, Austria*, pages 158–175, 2004. doi:10.1007/978-3-540-24646-6_10.
45. T. Teixeira and A. Savvides. Lightweight people counting and localizing for easily deployable indoors WSNs. *J. Sel. Topics Signal Processing*, 2(4):493–502, 2008. doi:10.1109/JSTSP.2008.2001426.
46. Telkonet. Telkonet products. URL: http://www.telkonet.com [cited 18.05.2017].
47. T. van Kasteren, A. K. Noulas, G. Englebienne, and B. J. A. Kröse. Accurate activity recognition in a home setting. In *Proc. 10th Int. Conf. on Ubiquitous Computing, UbiComp 2008, Seoul, Korea*, pages 1–9, 2008. doi:10.1145/1409635.1409637.
48. A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Information Theory*, 13(2):260–269, 1967. doi:10.1109/TIT.1967.1054010.
49. S. Wang. CO_2-based occupancy detection for on-line outdoor air flow. *Indoor Built Environment*, pages 165–181, 1989.
50. T. K. Woodstock, R. J. Radke, and A. C. Sanderson. Sensor fusion for occupancy detection and activity recognition using time-of-flight sensors. In *Proc. 19th Int. Conf. on Information Fusion (FUSION)*, pages 1695–1701, 2016.
51. L. Yang, K. Ting, and M. B. Srivastava. Inferring occupancy from opportunistically available sensor data. In *Proc. IEEE Int. Conf. on Pervasive Computing and Communications, PerCom 2014*, pages 60–68, Los Alamitos, CA, USA, 2014. doi:10.1109/PerCom.2014.6813945.
52. S. Zikos, A. Tsolakis, D. Meskos, A. Tryferidis, and D. Tzovaras. Conditional random fields - based approach for real-time building occupancy estimation with multi-sensory networks. *Automation in Construction*, 68:128 – 145, 2016. doi:10.1016/j.autcon.2016.05.005.