Look Twice: Uncover Hidden Information in Room Climate Sensor Data

Dominic Wörner*, Thomas von Bomhard†, Marc Röschlin*and Felix Wortmann†
*Department of Management, Technology and Economics, ETH Zurich, Switzerland
Email: dwoerner@ethz.ch, romarc@student.ethz.ch
†Institute of Technology Management, University of St. Gallen, Switzerland
Email: thomas.vonbomhard@unisg.ch, felix.wortmann@unisg.ch

***Abstract*—Connected sensors are on the march to become pervasive. While they are often deployed for a single purpose it is worth to take a second look. In this study, we show that the widespread Netatmo weather station which is intended to monitor and improve indoor climate can be used to estimate binary occupancy of individual rooms. We collected data from 11 rooms in 3 apartments including binary occupancy for several days. We show that CO_2 measurements and derivatives thereof qualify as observables to be used in Hidden Markov Models and achieve accuracies well above 75% in most cases. However, we see that the accuracy metric is often misleading for such timeseries data and consider additional performance metrics as well which show varying results depending on the respective occupancy patterns of a room.**

I. Introduction

The *Internet of Things* (IOT) promises to change the world and our lives. Billions of connected devices will be deployed which deliver zettabytes of data. However, so far most so-called IOT applications are vertically integrated and the generated data is captured in silos [1]. Before a *Future Internet* [2] is able to break those silos, an escape is given by a growing number of APIs which allow to find new means for data and enable developers to build mash-ups revealing innovative use-cases and applications. One prominent area of application for the IOT is the smart home. Contrary to intelligent fridges, smart thermostats and room climate monitoring solutions are gaining traction. Therefore, it is worth to take a second look at the arising data. In this study, we show that measurement data from the Netatmo weather station[1], a commercial indoor and outdoor climate sensor with a cloud API, can be used for binary occupancy estimation of individual rooms. This information in turn may be useful for several smart home application, e.g. to control heating systems more efficiently. The structure of this work is as follows. First, the Netatmo weather station is presented in Section II. Thereafter, in Section III previous work on environmental sensor-based occupancy detection is reviewed. In Section IV, the observational setup and data acquisition shown. Next, in Section V the Hidden Markov Model (HMM) is introduced and the feature identification is discussed. In Section VI the performance of the occupancy estimation using the HMMs is presented along several performance metrics. Before concluding in Section VIII, limitations and the application of the estimation for heating control is discussed in Section VII.

[1]http://www.netatmo.com

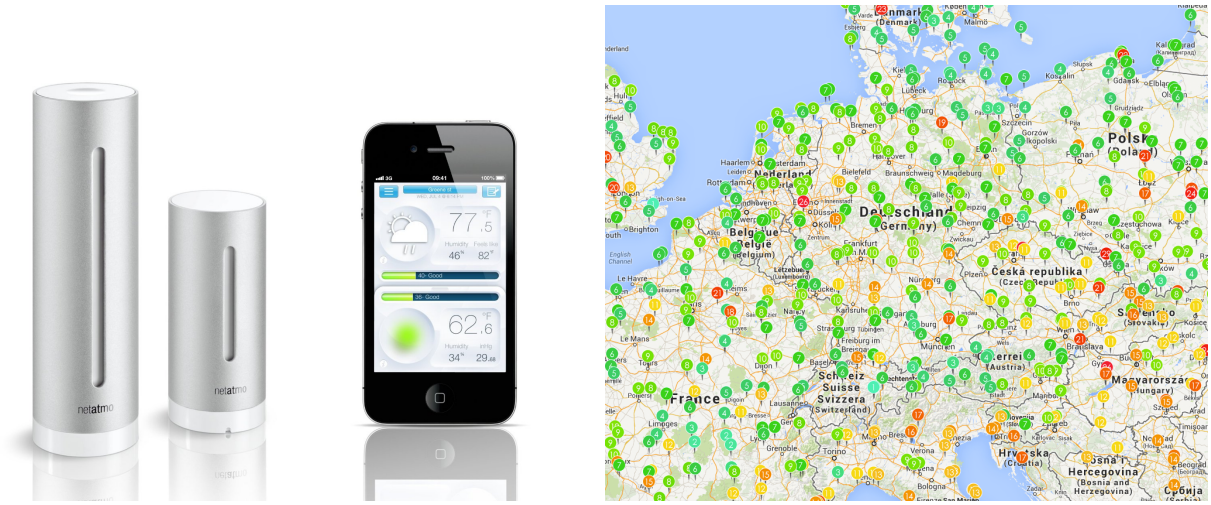
II. Room Climate Sensors

Netatmo is probably the first mainstream connected room climate sensor and outdoor weather station. The system consists of a base station that measures temperature, relative humidity, CO_2, barometric pressure and acoustics in 5min resolution. The base station has both, a wifi and a 868 Mhz, module. The 868 Mhz module allows to connect an outdoor module that measures temperature, relative humidity and barometric pressure, and up to three additional room modules which are similar to the outdoor module but entail an additional CO_2 sensor. The wifi module enables the communication with the Netatmo cloud service where the measurement data is stored and is available through an authenticated RESTful API. The intended main use case is to monitor indoor environmental variables in order to *improve your indoor wellness*. Figure 1a shows the base station, the outdoor module and the main screen of the iPhone app. The map in Fig. 1b indicates the spread of just this particular IOT room climate sensor. Note also that sharing outdoor temperature data is voluntary and meanwhile comparable products have hit the market.

III. Previous Work

There is a vast amount of research focusing on occupancy detection and further concepts of indoor location tracking and identification [3]–[7]. However, the purpose of this work is to illustrate that a widespread commercial room climate sensor can be alienated to infer room-level occupancy. For this reason, the focus of this review is on research that is based on comparable sensors.

Living beings generate heat, exhale moisture and CO_2, and usually produce some noise. Therefore, it is natural to ask the question if room climate sensors can be used to detect occupancy. Thus far, scholars have mainly focused on office buildings. In this scenario, room climate sensors with CO_2-sensing capabilities are already common to



(a) Sensor modules and iPhone app

(b) Weathermap showing outdoor temperatures of participating Netatmo modules

Fig. 1. Netatmo connected room climate sensor/weather station

provide input for HVAC systems. The prevalent approach was to develop a steady-state or dynamic model based on the mass balance of CO_2. However, this approach is only viable if room sizes and air exchange rates are known. The idea of applying machine learning techniques on indoor climate sensor data in order to infer occupancy was first implemented by *Lam and Dong et al.* [5]. They equipped an open plan office space with a CO_2 sensor network (2min sampling rate), and an additional sensing network consisting of luminosity, temperature, relative humidity, motion (PIR) and acoustics sensors (1min sampling rate). Based on a measure called information gain the best set of features to predict occupancy levels were selected. This process led to a feature set consisting of CO_2, acoustics and motion. In a follow up, *Dong et al.* [8] used these features to feed Support Vector Machines, Neural Networks and Hidden Markov Models (HMMs) and concluded that HMMs are better suited because they exhibit less fluctuations due to their inherent temporal structure. Quantitative results in terms of accuracy, however, are not conclusive. *Han et al.* [9] argued that temporal correlations between non-consecutive measurements of environmental parameters may be important. These correlations were taken into account by using an Autoregressive Hidden Markov Model (ARHMM). However, the average accuracy merely improved from 79.63% (HMM) to 80.78% (ARHMM), although the number of model parameters increases significantly. Only recently CO_2-based occupancy detection in residential buildings was investigated in comparison to PIR and device-free localization [7]. A three-bedroom dwelling was equipped with CO_2 sensors, PIR sensors, and an ultra-wideband tracking system in several rooms. Furthermore, the dwelling was equipped with a mechanical ventilation heat recovery (MVHR) system. In this study, occupancy detection on room and dwelling level was not investigated using machine learning techniques but rather by discussing the graphs of the time series data. Concerning CO_2, the authors concluded that air circulation patterns and status of doors and windows strongly effect CO_2 measurements and should therefore taken into account in order to allow reliable occupancy detection.

In summary, besides the recent, qualitative discussion occupancy estimation using room climate sensors was only investigated in office scenarios. Such a setting differs distinctively from the residential setting. While those offices were equipped with ventilation systems, a typical dwelling in central and northern Europe is ventilated manually by opening windows. Furthermore, occupancy patterns and the number of occupants are not comparable.

IV. Data Acquisition

Room climate data was collected in three apartments with 11 rooms in total (see Table I). Binary occupancy data was collected for periods between one and two weeks using switches in two apartments and cameras in one apartment. The switches were installed in the hallway besides every door to a room and a sign was attached to the door in order to remind inhabitants to operate the switch when entering a room as first as well as leaving a room as last. The switches transferred their status to a Raspberry Pi which logged the states and stored it in the database. In the apartment with cameras, two cameras were installed in the hallway which covered every door. The open source computer vision library OpenCV [10] was used to extract sequences with movements in the vicinity of doors. The timestamps of these events were extracted automatically and written to a file. Thereafter events were labeled manually.

TABLE I
List of rooms.

Apartment	Room	Occupancy	Period [days]	Id
1	Kitchen	Switch	9	1
	Livingroom	Switch	9	2
	Bathroom	Switch	9	3
	Bedroom	Switch	9	4
2	Kitchen	Switch	11	5
	Livingroom	Switch	11	6
	Bathroom	Switch	11	7
	Bedroom	Switch	11	8
3	Bedroom	Camera	16	9
	Bedroom	Camera	15	10
	Bathroom	Camera	16	11

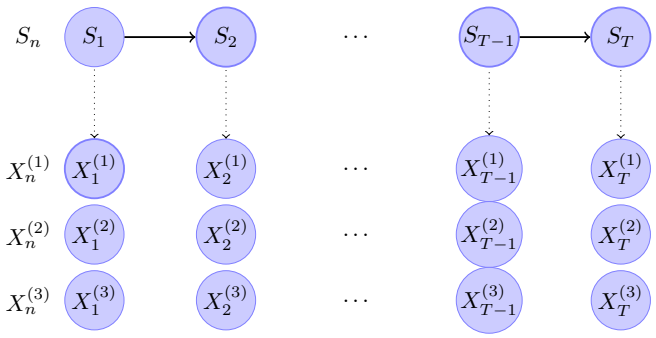


Fig. 2. Illustration of the Hidden Markov Model. S_i denotes the state S at the i-th time step. $X_i^{(j)}$ denotes the j-th observable at the i-th time step.

V. Methodology

A. Hidden Markov Model

A Hidden Markov Model[2] is a statistical model in which the dynamics are described by a first-order Markov process with unobservable (hidden) states. The hidden states are expected to generate distinct observables. In general a HMM is defined by a parameter set which can be written as the 3-tuple $\lambda = (A, b, \pi)$. The matrix A consists of the state transition probabilities

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i) \quad 1 \leq i, j \leq N \tag{1}$$

where $\{S_i\}$ is the set of hidden states with cardinality N. b denotes the observation symbol probability distribution or emission distribution

$$b_i(x_t) = P(X_t = x_t | q_t = S_i) \quad 1 \leq i \leq N \tag{2}$$

where x_t denotes the instantiation of the observables at time t. These may be categorical or continuous. Finally, the initial state probabilities are described by

$$\pi_i = P(q_1 = S_i) \quad 1 \leq i \leq N \tag{3}$$

An illustration of the HMM can be seen at Fig. 2.

In this work, for each room an independent HMM is assumed. The hidden states are identified with the occupancy of a particular room. A room is in the occupied state (O) if at least one person is present and in the vacant state (V) otherwise ($N = 2$). While this model is rather simple it has the advantage of having a small number of parameters which enables an reliable estimation using a limited training set.

B. Feature Identification

In the beginning of Section III the effect of human presence on environmental variables was briefly discussed. Now the question arises which measurement variables qualify as observables for the HMM and how should the respective emission probabilities be modelled. An approach to pursue these questions is to investigate histograms of measurement data. Figure 3 shows histograms of CO_2, its first and second derivative[3] as well as of temperature and relative humidity. Red (blue) color indicates that the room is vacant (occupied). In the histograms of CO_2 and their derivatives occupancy and vacancy can be separated reasonably well. This is not the case for temperature and relative humidity. Other influences cover the effect of human presence on these variables. In the case of temperature for instance, outside temperature, solar radiation and the heating system have a far greater effect on room temperature than a single human being. Hence, CO_2 and its first and second derivative are selected as features, i.e. observables, for the HMM. While the derivatives of CO_2 could be modeled as categorical variables being either positive or negative, for the actual CO_2 values a continuous distribution has to be used. However, in order to have a simple model one multivariate Gaussian distribution is used to model the emission probabilities. A multivariate Gaussian distribution is defined by a vector of the means and the covariance matrix. Here, we assume no correlations between the observables which leads to a diagonal covariance matrix where the elements are given by the variances.

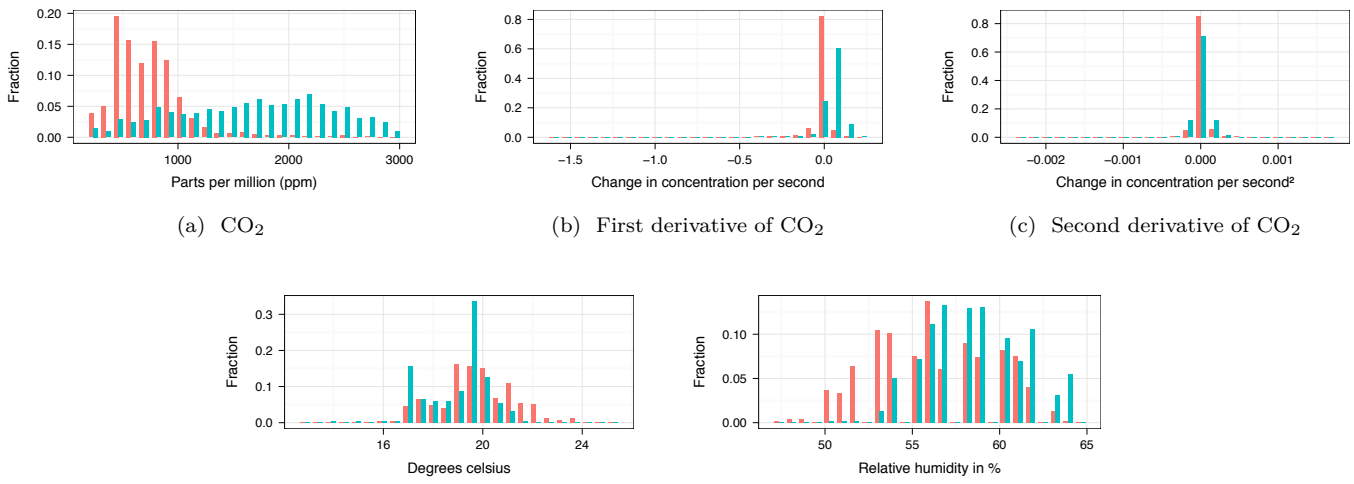
C. Training

The data was divided in a training and validation set. The training set consists of the first 7 days while the remaining days (apartment 1: 2, apartment 2: 4, apartment 3: 9) are used for validation. Through simulations and considering the size of the overall data set, we found that at least one week, i.e. seven days of training information, is needed for the model to adjust its parameters and guarantee an adequate performance.

The state transition probabilities a_{ij} are given by counting the transitions in the training set and computing their relative frequencies. In order to calculate the emission probabilities, the measurement data of the observables in the training set is divided according to the state they belong to. Thereafter, the parameters of the Gaussian

[2]Consult [11] for an excellent introduction.

[3]Derivatives are computed by interpolating the original data with third-order splines and computing derivatives thereof.



(a) CO_2 (b) First derivative of CO_2 (c) Second derivative of CO_2

(d) Temperature (e) Relative humidity

Fig. 3. Histograms for feature identification (room Id 9). The cyan (red) color denotes occupancy (vacancy).

distributions are approximated by

$$\mu_i \approx \bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \tag{4}$$

and

$$\sigma_i^2 \approx s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \bar{x_i})^2 \tag{5}$$

where n_i denotes the cardinality of the training set with state S_i. The initial state probability π_i is set to be the relative frequency of S_i in the training set.

D. Prediction of Hidden State Sequence

The prediction of the hidden state sequence is identified with the problem of finding the *single* best state sequence given the model and observations, i.e. maximizing $P(S|X, \lambda)$. The formal solution to this problem is based on dynamical programming and is called the Viterbi algorithm [12], [13]. Herein, an implementation of the algorithm in R was used [14].

VI. Results

A. Performance Metrics for Binary Occupancy Estimation

In order to evaluate the performance of the binary occupancy estimation the following metrics are considered: accuracy ($\frac{TP+TN}{TP+FP+TN+FN}$), precision ($\frac{TP}{TP+FP}$), sensitivity ($\frac{TP}{TP+FN}$), specificity ($\frac{TN}{FP+TN}$) and F1 score ($\frac{2TP}{2TP+FP+FN}$). Hereby, TP is the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives. Positive (negative) refers to the occupied (vacant) state. In all cases a higher number means that the model is able to make a better prediction.

B. Evaluation

In Figure 4 the performance metrics of the prediction are shown. The accuracy as well as the specificity is above 75% in most cases. At first glance accuracy seems to be a reasonable metric as it is defined as the ratio of correct predictions and all predictions. However, if a room is vacant most of the time, the accuracy might be close to one even though the model failed to predict the short intervals of occupancy. Specificity which is given by the ratio of correct vacancy predictions and all vacant times is even more biased in such a case. Since all residents are working, there is in general a much higher probability to find a room vacant than occupied. Therefore, precision and sensitivity are much more informative. It can be seen that in rooms which are visited frequently for short periods of time like bathrooms and kitchens these measures are particularly low (c.f. Fig. 6). The zeros in these metrics for room 2 are because the living room was not used during the two days which were used for validation.

Figure 5 illustrates two days of occupancy estimation for two different rooms. These are archetypal, since Fig. 5a represents a room with quite continuous periods of occupancy (bedroom) whereas Fig. 5b represents a room with brief visits (bathroom). Besides the delay in the evening, the occupancy estimation of the bedroom resembles the general occupancy pattern. For the bathroom, only the longer visits are detected. Since the room is small the occupancy prediction is hardly delayed. However, the transition to the vacant state is delayed.

VII. Discussion

A. Occupancy Ground Truth

The results for the apartment with camera-based ground truth are considerably better than for the apartments with

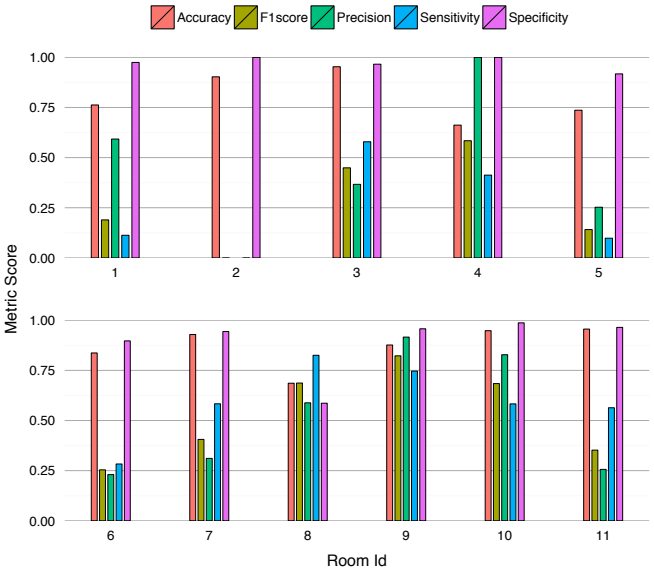


Fig. 4. Performance metrics on validation set.

switch-based ground truth. Looking at the individual time series of each room the reason for this becomes obvious. Although there were signs at each door to remind participants to operate the switches it is uncertain if they did it always correctly. Indeed, there are occurrences in the data which are obviously erroneous. For instance, in one case *ground truth* shows that the kitchen (Id 5) was occupied all over the night but the participants explained that this did not happen. Therefore, the switch-based *ground truth* is dubious. The camera-based approach in contrary is precise but the effort to evaluate the recordings is huge although we used computer vision-based motion detection to streamline the process. Furthermore not every participant agrees to be monitored by cameras. In a first test, we tried to use iBeacons[4] to gain room-level occupancy information. This approach involves that the participants would always carry their smartphones. However, it turned out that location tracking using one iBeacon per room so far is not stable enough to enable trustworthy room-level occupancy information.

B. Limitations

Obviously, the concentration and diffusion of CO_2 depends on the room size, air velocity and infiltration rate and thus the state of windows and doors. Since this study was carried out during winter between the end of January and the beginning of February) windows were typically only opened for short time periods[5]. During these periods CO_2-based observables are inappropriate to estimate occupancy. Hence, in summer, when windows may be open continuously, this approach won't work.

[4]A technology based on Bluetooth Low Energy

[5]Airing can be seen in the data quite well.

Furthermore, as already discussed briefly in Section VI-B, short intervals of occupancy (or vacancy) are hard to predict (see Fig. 5b). This has two main reasons. First, the sampling rate of the sensor is 5 min. Second, depending on the size of the room and the location of the sensor it may take some until significant changes of CO_2 reach the sensor. An additional instant sensing method like acoustics (which is available within the Netatmo base station but again only as 5 min moving averages) or motion could lead to a great improvement in such scenarios.

Finally, in this work a simple form of supervised learning was used to determine the parameters of the HMM. In a real-world setting such training data won't be available. However, considering Figure 3 again, it might be at least possible to define emission probabilities that could work for different rooms. In addition, the room category may give estimates for the transition probabilities.

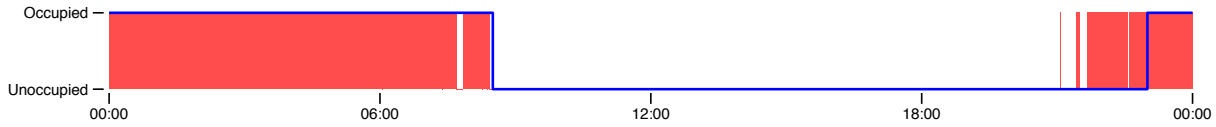
C. Possible Application

Given the aforementioned limitations, a promising application of this type of occupancy estimation is occupancy-based heating. Since the occupancy estimation typically lags behind the ground truth this approach is not suitable for a reactive control system, i.e switching the heating on (off) if occupancy (vacancy) is detected. Note, however, that heating systems, in particular hydronic systems prevailing in Europe, in general exhibit a delayed response. In addition building dynamics are slow. Therefore, it is already too late to switch on the heating when people arrive. For this reason, a predictive control system [15, c.f.] is favourable which would benefit from historical occupancy information inferred from room climate measurements.

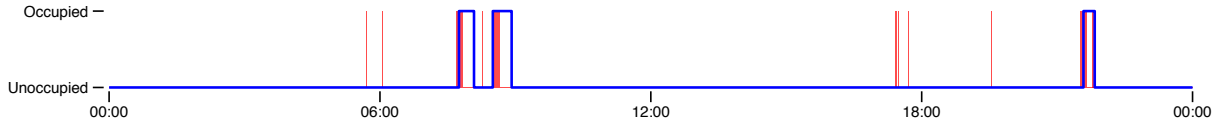
VIII. Conclusion

In this work, we showed how room climate sensor data from a consumer IOT weather station can be used to infer binary occupancy estimation of individual rooms by applying the well-known machine learning technique of Hidden Markov Models. We find that observables based on CO_2 measurements qualify for occupancy estimation, while temperature and humidity are depending more on other environmental conditions. Although accuracies of the occupancy estimations are high ($> 75\%$) in almost all cases, we find a good resemblance of the occupancy profiles only for rooms with continuous periods of occupancy like bedrooms and living rooms. This is supported by looking at finer performance metrics like precision and sensitivity.

As an application we expect that in particular predictive heating control systems could benefit from such a simple and unobtrusive occupancy estimation. The method, however, is not suitable for reactive scenarios like switching light, since it may take some time until the change of CO_2 is sufficiently large. This is in particular true for larger rooms. Additional measurements like luminosity



(a) Ground truth and occupancy prediction for bedroom with Id 9.



(b) Ground truth and occupancy prediction for bathroom with Id 11.

Fig. 5. A single day from the validation period of two rooms to give an example of the ground truth presence data and the corresponding occupancy prediction. The red areas depict the actual occupancy. The blue line demonstrates the prediction our model achieves.

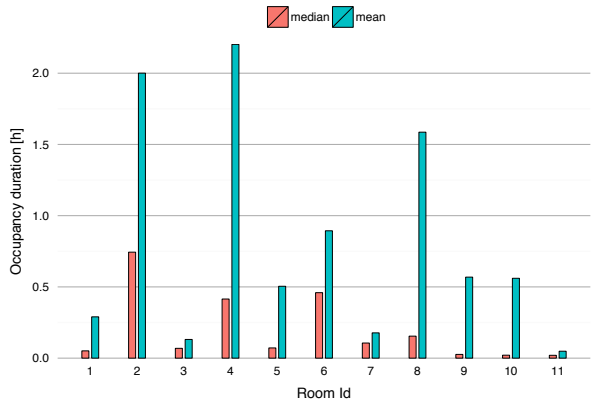


Fig. 6. Median and mean occupancy duration per room.

and acoustics, as e.g. offered by CubeSensors[6] may help to overcome these issues. Further, classical motion-based occupancy detection could be supported by climate sensor-based occupancy estimation because living beings interact with their environment even if they don't move. So far the approach was based on having occupancy ground truth information in order to train the model. Future work will be directed to finding unsupervised approaches. Interestingly, there is more than just occupancy hidden in room climate sensor data. Ventilation behavior, showering, cooking, sleeping and probably many more activities could be extracted such that your smart home knows even more about its inhabitants.

Acknowledgment

This ongoing research is kindly supported by the Bosch IoT Lab at Sankt Gallen University, Switzerland.

References

[1] M. Zorzi, A. Gluhak, S. Lange, and A. Bassi, "From today's intranet of things to a future internet of things: a wireless- and mobility-related view," *Wireless Communications, IEEE*, vol. 17, no. 6, pp. 44–51, 2010.

[2] L. Tan and N. Wang, "Future internet: The internet of things," in *Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on*, vol. 5. IEEE, 2010, pp. V5–376.

[3] R. H. Dodier, G. P. Henze, D. K. Tiller, and X. Guo, "Building occupancy detection through sensor belief networks," *Energy and Buildings*, vol. 38, no. 9, pp. 1033–1043, 2006.

[4] S. Funiak, C. Guestrin, M. Paskin, and R. Sukthankar, "Distributed localization of networked cameras," in *Proceedings of the 5th International Conference on Information Processing in Sensor Networks*, ser. IPSN '06. ACM, 2006, pp. 34–42.

[5] K. P. Lam, M. Höynck, B. Dong, B. Andrews, Y.-S. Chiou, R. Zhang, D. Benitez, J. Choi *et al.*, "Occupancy detection through an extensive environmental sensor network in an open-plan office building," in *Proceedings of the 11th International IBPSA Conference*, 2009, pp. 1452–1459.

[6] M. Seifeldin and M. Youssef, "A deterministic large-scale device-free passive localization system for wireless environments," in *Proceedings of the 3rd International Conference on Pervasive Technologies Related to Assistive Environments*, ser. PETRA '10. New York, NY, USA: ACM, 2010, pp. 51:1–51:8.

[7] E. Naghiyev, M. Gillott, and R. Wilson, "Three unobtrusive domestic occupancy measurement technologies under qualitative review," *Energy and Buildings*, vol. 69, no. 0, pp. 507 – 514, 2014.

[8] B. Dong, B. Andrews, K. P. Lam, M. Höynck, R. Zhang, Y.-S. Chiou, and D. Benitez, "An information technology enabled sustainability test-bed (itest) for occupancy detection through an environmental sensing network," *Energy and Buildings*, vol. 42, no. 7, pp. 1038–1046, 2010.

[9] Z. Han, R. Gao, and Z. Fan, "Occupancy and indoor environment quality sensing for smart buildings," in *Instrumentation and Measurement Technology Conference (I2MTC), 2012 IEEE International*. IEEE, 2012, pp. 882–887.

[10] G. Bradski, *Dr. Dobb's Journal of Software Tools*.

[11] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[12] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *Information Theory, IEEE Transactions on*, vol. 13, no. 2, pp. 260–269, 1967.

[13] G. D. Forney Jr, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[14] J. O'Connell and S. Højsgaard, "Hidden semi markov models for multiple observation sequences: The mhsmm package for R," *Journal of Statistical Software*, vol. 39, no. 4, pp. 1–22, 2011. [Online]. Available: http://www.jstatsoft.org/v39/i04/

[15] F. Oldewurtel, D. Sturzenegger, and M. Morari, "Importance of occupancy information for building climate control," *Applied Energy*, vol. 101, pp. 521–532, 2013.

[6]http://www.cubesensors.com