*INTERSPEECH 2017*
August 20–24, 2017, Stockholm, Sweden

# Occupancy Detection in Commercial and Residential Environments Using Audio Signal

*Shabnam Ghaffarzadegan*[1], *Attila Reiss*[2], *Mirko Ruhs*[2], *Robert Duerichen*[2], *Zhe Feng*[1]

[1]Robert Bosch Research and Technology Center (RTC), Palo Alto, CA
[2]Robert Bosch GmbH, 70465 Stuttgart, Germany

Shabnam.Ghaffarzadegan@us.bosch.com, Attila.Reiss@de.bosch.com, Mirko.Ruhs@de.bosch.com, Robert.Duerichen@de.bosch.com, Zhe.Feng2@us.bosch.com

## Abstract

Occupancy detection, including presence detection and head count, as one of the fast growing areas plays an important role in providing safety, comfort and reducing energy consumption both in residential and commercial setups. The focus of this study is proposing affordable strategies to increase occupancy detection performance in realistic scenarios using only audio signal collected from the environment. We use approximately 100-hour of audio data in residential and commercial environments to analyze and evaluate our setup. In this study, we take advantage of developments in feature selection methods to choose the most relevant audio features for the task. Attribute and error vs. human activity analysis are also performed to gain a better understanding of the environmental sounds and possible solutions to enhance the performance. Experimental results confirm the effectiveness of audio sensor for occupancy detection using a cost effective system with presence detection accuracy of 96% and 99%, and the head count accuracy of 70% and 95% for the residential and commercial setups, respectively.

**Index Terms**: head count, presence detection, human-machine interaction, audio analytics.

## 1. Introduction

Over the past few years, there has been a significant growth towards offering audio analytics solutions for a number of application domains such as smart homes and buildings. These applications are not only important in improving the comfort and safety of the building, but also in reducing the energy consumption. Occupancy detection, including presence detection and head count, is one of the key elements in smart homes and buildings with applications such as: efficient smart heating, ventilation and conditioning system, smart evacuation in emergency situations and discovering the abnormal patterns of occupants' behavior in security systems.

Presence detection and head count have number of challenges in real-life scenarios. One of the key challenges is performing these tasks in different environments, conditions and background noises. For example, presence detection in open office environment or in presence of TV noise in the background is more difficult than in a quiet single office. Moreover, most of the time using only one sensor does not fulfill the task completely, besides using more sensors will have high energy consumption, cost, and computational complexities. Finally, taking people's privacy into account especially in residential areas is another important factors in the system design. For instance, using audio sensors are much preferable than video sensors in smart home applications. Techniques available in the literature attempt to solve the problem using mainly a single modality or fusion of small number of sensors [1–4]. The task of occupancy sensing has been investigated using different sensors in the literature. Vibration and acoustic sensors have been used in [5–7] to detect the presence based on foot-step vibration and the audio. In these studies, simple features such as: mean and variance of windows of the signal or Mel frequency cepstral coefficients (MFCC) are used in the experiments. Furthermore, radar and sonar sensors [8, 9], chemosensors [4, 8], and thermal imaging [10, 11] are among some of the modalities that have been previously explored for the occupancy detection task. However, most of these approaches have limitations such as: limited detection distance, failure in detecting an static person, high computational complexity, etc.

In this paper, our focus is on the design of affordable and cost-effective strategies solely based on audio signal, with the goal to increase the performance of occupancy detection tasks in realistic scenarios. The audio sensor is chosen as the main modality due to: 1) its easy accessibility in different environments and scenarios, 2) its proven accuracy in different applications, 3) its privacy friendly nature, 4) high coverage of the room (vs other sensors such as FIR with limited field of view and critical sensor position), low costs, etc. The main contributions of the paper are threefold. First, we explored the audio feature space and identified the most important features for occupancy detection. Second, we performed a comprehensive study in order to provide an affordable solution considering computational complexity. Third, we conducted an analysis on the output error with respect to human activity to gain a better understanding of the problem and possible solutions.

## 2. Corpus Description

The audio samples utilized in this study are drawn from a multi-sensor occupancy detection corpus collected at Robert Bosch GmbH [12]. This corpus includes recordings of five different commercial and residential environments: open office, single office, meeting room, bedroom, and living room. All three commercial environments in this corpus are collected in real-life scenarios, while the residential environments (bedroom and living room) are collected in a laboratory setting. The audio data are recorded in 44.1 kHz using Robert Bosch GmbH MEMS acoustic sensor model AKU151, which is specifically designed for space-constrained consumer electronic devices. Ground truth (number of people present at a given time) is provided throughout the entire corpus, based on video recordings.

A subset of the corpus is chosen in this study: single office and living room (as representatives of commercial and residential scenarios), with $\sim$ 75 hours and $\sim$ 22 hours of audio data, respectively. Figure 1 represents detailed statistics of the two selected environments. The single office data was acquired in a small, enclosed room including one fully equipped office work-

http://dx.doi.org/10.21437/Interspeech.2017-524

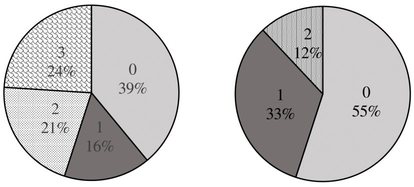

Figure 1: *Head count distributions of the single office (right chart) and living room (left chart) datasets.*

place. The recordings were carried out over seven days during daytime on workdays. The recorded data mostly comprises of regular office work of a single person, including conversations over phone and longer meetings with a second colleague. Additionally, in order to increase data variety and balance the presence ratio, data was recorded on a day off and over one night. The living room data was acquired in a larger lab room, furnished as a simple living room setting. Data was recorded over six sessions, each following a predefined protocol with activities in varying order. The following activities were carried out: watching TV, reading newspaper, talking, and playing a cardgame. The number of people present in the room as well as an approximate length of each activity was defined in the protocol.

## 3. Methodology

In the following, we describe the details of our data processing chain.

**Feature Extraction**: In the first step of the chain, the well-known audio low level descriptive (LLD) features are extracted using a frame-level sliding window with 25 ms length and no overlap. The overlap time was investigated experimentally for the head count task. Table 1 represents the LLD features used in this study and their dimension.

**Feature Segmentation**: Second, features are partitioned into segments with a fixed length in time and a shift of one frame. Different segment lengths are explored in the experimental section to investigate the optimum time window for occupancy detection task.

**Feature Functionals**: Third, functionals are applied to LLDs and their delta and acceleration coefficients in each segment. The functionals are chosen as final features to highlight the most interesting information in a longer time segment. Third column of table 1 shows the functionals used in this study.

**Feature Selection**: In this paper, we exploit the feature space for head count task and investigate the contribution of different audio feature types for the classification accuracy. Given the huge number of possible audio features we typically need to de-correlated and reduce the feature space via feature space transformations or feature selection. In this work, we will not follow feature space transformation strategies, due to the fact that these methods will not answer the question of the contribution of each feature for the task. Instead, we use feature selection methods to pool together the most relevant and uncorrelated features.

Table 1: *Low level descriptive (LLD) features and functionals computed on audio data. MFCC-d: MFCC delta; MFCC-dd: MFCC acceleration; Std: Standard deviation; Abs Integral: Absolute value Integral.*

| Features | Dimension | Functional |
|---|---|---|
| Envelope dynamic range | 1 | Mean<br>Median<br>Std<br>Abs Integral<br>Min<br>Max<br>Dynamic range<br>Dominant-frequency<br>Entropy |
| Zero crossing rate | 1 | |
| Energy | 1 | |
| Brightness | 1 | |
| Spectral variance | 1 | |
| Spectral roll off | 1 | |
| Spectral flux | 1 | |
| MFCC | 20 | |
| MFCC-d | 20 | |
| MFCC-dd | 20 | |

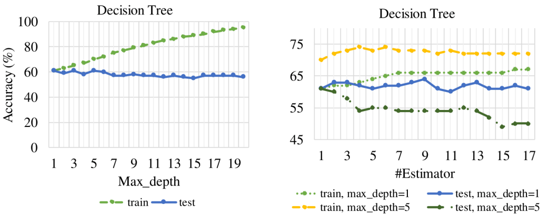

Figure 2: *Decision tree classification accuracy vs maximum tree depth and number of the estimators in living room environment.*

Feature selection as an automatic method to select the most relevant features to the modeling problem has many benefits, such as: improving the performance, providing a faster and simpler model, and allowing better understanding of the data and its underlying process [13]. Different feature selection methods put more emphasis on one aspect than others. In this work, we have used two common algorithms, namely: univariate Chi2 and least absolute shrinkage and selection operator (LASSO) methods due to their simplicity, speed and effectiveness. Univariate Chi2 is an attribute ranking method based on the Chi2 value of attributes and their corresponding target labels [14]. As a result, if the target variable is independent from the feature variable, that feature can be discarded. Otherwise, the feature variable is important. Moreover, LASSO, as another feature selection method used in this study, is an alternative to ridge regression method in linear modeling that yields to a sparse model [15]. In LASSO, $l1$ penalty is used compared to the $l2$ penalty in ridge regression case. Hence, it can be used as feature selection method.

**Classifier**: Finally, the last step of the data processing chain is the classifier. Here, we will not optimize the classifier choice but simply use a decision tree classifier due to its simplicity and interpretability. In the experimental section, decision tree classifier is optimized for Maximal depth of tree and other parameters such as: minimal size for split, minimal leaf size, and decision criterion are assumed to be fixed values. Also, ensemble decision tree has been evaluated to improve generalizability and robustness of the classifier.

## 4. Experimental Results

In all the experiments, functionals are applied on a time window of 5 $s$. Also, leave-one-recording-out cross-validation method is used throughout the experiments, as an evaluation technique. Consequently, 9-fold and 7-fold cross-validation are used in the living room and single office scenarios, respectively. Finally, classification accuracy has been used as the performance measurement.

The classifier performance on the full feature set in living room environment is shown in Fig. 2. Left plot of the figure shows decision tree performance with respect to maximum depth with 1 estimator, 5 minimal size for split, 1 minimal leaf size, and Gini impurity decision criterion. As observed, deeper tree yields to an over-fitted model and increases the computational complexity, dramatically. Right plot of Fig. 2 represents

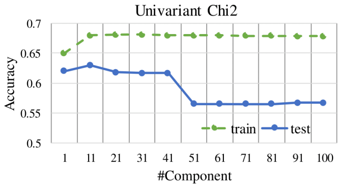

Figure 3: *Decision tree classifier performance after univariate Chi2 feature selection in living room environment; #Component: Number of feature dimension.*

the boosted decision tree classification performance with respect to the number of estimators, other parameters of the model are kept the same. Based on this figure, not much performance has been gained through ensemble method. At the end, a simple decision tree with maximum depth of 5 and 1 estimator is chosen as the baseline system due to the performance and computational cost trade-off. Because of lengthy computations, we avoid repeating the same experiments for single office environment on the full feature set, and will simply use the same decision tree parameters in the rest of the paper.

In the next experiments, we aim to find each feature variable contribution in head count task using Chi2 and LASSO methods in the living room scenario. Figure 3 shows the system accuracy with respect to number of features chosen via Chi2 method. As seen in the figure, choosing only two components out of hundreds of features, which have the highest Chi2 value between the feature variables and labels, can improve the performance up to 7% absolute value and speed up the classifier, greatly. Figure 4 is also representing the classifier accuracy with respect to features chosen by LASSO method in living room scenario. These results are in agreement with the results from Chi2 feature selection. Given the similar accuracy trend between Chi2 and LASSO methods, we continue our experiments based on LASSO feature selection for single office environment. Given Fig. 5, selecting only a few components provides a reasonable classification performance in the single office scenario, as well. Note that the results driven from LASSO and Chi2 methods are in agreement with the decision tree classification results in Fig. 2, in which only one- or two- layer decision tree can lead to descent classification results.

Moreover, we summarized the first 18 relevant attributes, result from LASSO, for the living room and single office environments in Table 2 along with their LASSO coefficients in Fig. 6 and 7. As Table 2 presents the most important feature chosen in both environments is the energy term (first MFCC coefficient is the audio signal energy representation, as well). This result is expected heuristically, too. However, two different functionals namely: Maximum and Mean are picked for living room and single office scenarios, respectively. Also, it is clear from the selected features that the time and frequency audio features are playing more important role in head count task rather than the cepstal (MFCC) features. This results may indicate that MFCC features are not the most suitable features for environmental sounds.

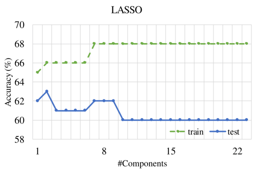

Figure 4: *Living room.*

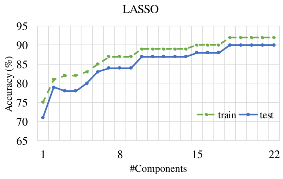

Figure 5: *Single office.*

*Decision tree classifier performance after LASSO feature selection in single office environment, right plot, and living room, left plot; #Component: Number of feature dimension.*

Table 2: *The most important attributes for head count task using LASSO method. d#: MFCC delta; dd#: MFCC acceleration.*

| LASSO-Single Office | LASSO-Living Room |
|---|---|
| 1. Mean-Energy | 1. Max-mfcc1 |
| 2. Mean-Brightness | 2. Dynamic Range-Envelope Dynamic Range |
| 3. Median-Brightness | 3. Dynamic Range-Brightness |
| 4. Std-Spectral Flux | 4. Entropy-d4 |
| 5. Std-Envelope Dynamic Range | 5. Entropy-mfcc19 |
| 6. AbsIntegral-d5 | 6. Entropy-Envelope Dynamic Range |
| 7. Abs Integral-Brightness | 7. Entropy-Zero Crossing Rate |
| 8. Entropy-Brightness | 8. Entropy-Spectral Rolloff |
| 9. Entropy-Spectral Variance | 9. Entropy-Spectral Flux |
| 10.Entropy-d2 | 10.Entropy-mfcc7 |
| 11.Entropy-d4 | 11.Entropy-dd1 |
| 12.Entropy-Energy | 12.Entropy-mfcc5 |
| 13.Entropy-Envelope Dynamic Range | 13.Entropy-Brightness |
| 14.Entropy-d20 | 14.Entropy-mfcc3 |
| 15.Entropy-Spectral Flux | 15.Entropy-d2 |
| 16.Entropy-dd1 | 16.Entropy-d4 |
| 17.Entropy-mfcc5 | 17.Entropy-Energy |
| 18.Entropy-mfcc3 | 18.Entropy-d20 |

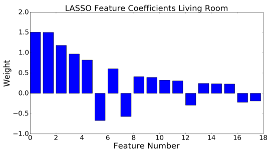

Figure 6: *Living room*

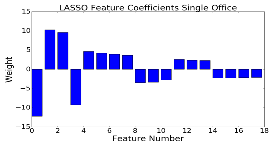

Figure 7: *Single office.*

*Features importance using LASSO feature selection for single office environment, left plot, and living room, right plot (number to feature mapping can be found in Table 2).*

To select the best feature set for both environments, head count performance vs. attributes is studied in Figures 8 and 9 for living room and single office environments, respectively. In these figures, x-axis represent the attributes exploited in each experiment. #-LR and #-SO represent the feature number selected in the living room (LR) and single office (SO) scenarios driven from table 2. Moreover, the + sign at the beginning of the x-axis label shows the attribute accumulation from the previous step. For example, the first point in the plot (1-LR) represents the accuracy when using the first best feature in living room scenario (Max-mfcc1), 1-LR+1-SO shows the performance when using best features of both environments (Max-mfcc1 and Mean-Energy), +2-LR represents the accuracy when using 3 dimensional features 1-LR, 1-SO and 2-LR, and so on. Based on both plots, 11 dimensional features (6-LR point on the x-axis) has been chosen as the final optimum set with 63% and

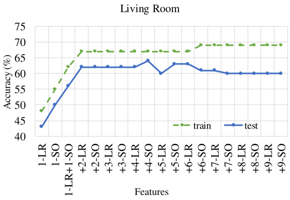

Figure 8: *Living room.*

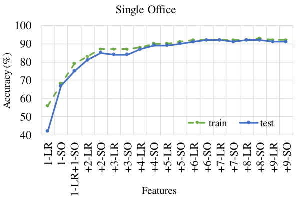

Figure 9: *Single office.*

*Head count accuracy vs. attributes for living room, left plot, and single office, right plot.*

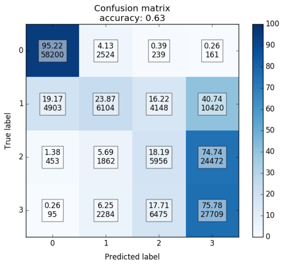

Figure 10: *Living room.*

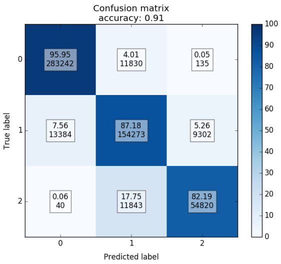

Figure 11: *Single office.*

*Head count confusion matrix of single office environment, right plot, and living room, left plot.*

91% head count accuracy for living room and single office environments, respectively. Moreover, confusion matrices of head count task for the two environments have been calculated using the 11-dimensional final feature set. Based on Fig. 10 and 11, presence detection has a high performance rate using only audio signal with 95% accuracy for living room scenario and 96% accuracy for the single office scenario. As a result, audio sensor can be used confidently as the first step of presence detection and other sensors may be fused in case of head count to increase the accuracy. This way, we can save both computational and energy resources that are important factors in real life applications. Furthermore, the overall head count accuracy using only a subset of features not only speed up the train and test time and gave us a better understanding of the feature variable contributions, but also reached to a satisfactory classification performance, with 63% and 91% classification accuracy for living room and single office environments, respectively. Based on the confusion matrices, performance in single office environment is accurate for 0, 1 and 2 head counts. However, in living room scenario 1 or 2 people presence are usually mistaken with 3 head count. In the next section, we will perform an error analysis to investigate issues related to the head count errors and possible solutions.

Finally, different segment lengths (5, 10, 30, 60, 120, 300, 600 $s$) are extracted to investigate the optimum time window of the functionals. Conducted experiments propose an optimum window length of 30 $s$, with accuracy improvement from 91% to 95% for head count task, and from 96% to 99% for presence detection in single office setup. In living room setting, the performance improved from 63% to 70% for head count and from 95% to 96% for presence detection. These results indicate the longer time-span of head count task compared to other audio analytic tasks, such as: ASR, emotion recognition, etc. As sanity check, we also evaluate the system performance on the other three environments included the corpus using the 11-dimension final features and 30 $s$ segments. The system gains 48%, 61% and 81% accuracy in open office (9-way), bedroom (2-way), and meeting room (7-way) environments. These results show the scalability of the selected features in scenarios outside of the training set.

Table 3: *head count error vs. human activity.*

| True HC | Predicted HC | Living Room | Single Office |
|---|---|---|---|
| 0 | 1 | Background noise | Outside Noise |
| 1 | 0 | Background noise | Background noise |
|  | 2 | Moving objects | Moving objects |
|  | 3 | Moving objects |  |
| 2 | 1 | TV<br>One person talking | One person talking |
|  | 3 | Two people talking |  |
| 3 | 1 | One person talking |  |
|  | 2 | People talking and/or TV |  |

### 4.1. Error Analysis

In final set of experiments, head count error analysis vs. human activity is performed to: 1) identify the most problematic activities in presence detection/head count task, and 2) gain an insight of possible solutions to improve the classification performance. Human activity time boundaries were manually labeled by an expert annotator for 4 hours of data evenly distributed in different conditions of the two environments. Table 3 is summarizing the error analysis results driven from manual annotation and classifier results for both environments. As seen in the table, 0 occupancy might be mistaken by 1 head count mainly due to background or outside noise (ventilation system noise, people walking or talking outside, so on). 1 person occupancy in the room is mistaken by 0 due to background noise or is predicted as 2 or 3 head count due to moving objects (this may include person walking in the room, as well). Moreover, 2 people presence in the room is mistaken by 1 when only one person is talking and the other one is quiet or when there is only TV sound present. 2 people presence is also mistaken by 3 mainly due to two people talking simultaneously and TV in the background. Finally, 3 people presence is mistaken by 1 when only 1 person is talking and the other 2 are silent, and is mistaken by 2 when there are 2 people talking and/or TV sound.

The aforementioned analysis may suggest some possible solutions and future works to enhance the classification accuracy, such as: 1) in scenarios where the person is quiet or not everybody talking, it might be a better idea to fuse other sensors such as motion or temperature. In this case, the audio sensor can be used in the first stage to perform presence detection task and in case of presence, activates other sensors to be included in the classifier. This strategy will save both time and computational resources. 2) Background noise, especially TV sound in the background, is one of the most problematic scenarios in occupancy detection. TV noise might be suppressed using techniques focusing on differences between sound in the room and the audio played from the TV speakers.

## 5. Conclusion

In this study, we explored the audio signal capability for occupancy detection application and analyzed different audio features relevance to the task. Using Chi2 ranking criteria and LASSO regression model, we ranked hundreds of audio variables and extracted a subset of 11-dimensional set. It has been observed that the basic time and frequency domain audio features carry more weights compared to the cepstral features (MFCC) that are more popular in speech related applications. Finally, we analyzed the system error vs. human activities to gain a better understanding of the problematic environmental sounds and possible solutions to improve the system. Based on the analysis, Most errors occur when there is no audio signal (predicting 0 head count) or in presence of background noise such as TV (predicting more head counts). Hence, developing robust features to background noise and using other type of sensors in silence mode might be some possible future directions. The final setup results in 95% head count and 99% presence detection accuracy in single office, and 70% head count and 96% presence detection accuracy for living room setup.

# 6. References

[1] C. Basu, C. Koehler, K. Das, and A. K. Dey, "PerCCS : Person-Count from Carbon dioxide using Sparse Non-negative Matrix Factorization," *Proc. 2015 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. (UbiComp '15)*, pp. 987 – 998, 2015.

[2] Y. Agarwal, B. Balaji, R. Gupta, J. Lyles, M. Wei, and T. Weng, "Occupancy-driven energy management for smart building automation," *Proc. 2nd ACM Work. Embed. Sens. Syst. Energy-Efficiency Build. - BuildSys '10*, p. 1.

[3] K. P. Lam, M. Höynck, R. Zhang, B. Andrews, Y.-S. Chiou, B. Dong, and D. Benitez, "Information-theoretic environmental features selection for occupancy detection in open offices," *Build. Simul. 2009, 11th Int. IBPSA Conf.*, pp. 1460–1467, 2009.

[4] K. P. Lam, M. Höynck, B. Dong, B. Andrews, Y.-s. Chiou, D. Benitez, and J. Choi, "Occupancy detection through an extensive environmental sensor network in an open-plan office building," *IBPSA Conf.*, pp. 1452–1459.

[5] A. Pakhomov, A. Sicignano, M. Sandy, and T. Goldburt, "Seismic Footstep Signal Charicterization," *Proceeding SPIE*, vol. 5071, no. 2003, pp. 297–305, 2003.

[6] S. Uziel, T. Elste, W. Kattanek, E. Sebastianuzielimmsde, D. Hollosi, S. Gerlach, and S. Goetze, "Networked Embedded Acoustic Processing System for Smart Building Applications," *Des. Archit. Signal Image Process. (DASIP), Conf.*, 2013.

[7] A. Khan, J. Nicholson, S. Mellor, D. Jackson, K. Ladha, C. Ladha, J. Hand, J. Clarke, P. Olivier, and T. Plötz, "Occupancy monitoring using environmental; context sensors and a hierarchical analysis framework," in *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*, ser. BuildSys '14, 2014, pp. 90–99.

[8] P. Bahl and V. N. Padmanabhan, "RADAR: An In-building RF-based User Location and Tracking System," *Proc. IEEE INFOCOM 2000. 19th Annu. Conf. Comput. Commun.*, vol. 2, pp. 775–784, 2000.

[9] S. P. Tarzia, R. P. Dick, and P. a. Dinda, "Sonar-based Measurement of User Presence and Attention," *Proc. Int. Conf. Ubiquitous Comput. (UbiComp '09)*, pp. 89–92, 2009.

[10] A. Beltran, V. L. Erickson, and A. E. Cerpa, "ThermoSense: Occupancy Thermal Based Sensing for HVAC Control," *BuildSys'13 Proc. 5th ACM Work. Embed. Syst. Energy-Efficient Build.*, 2013.

[11] P. Hevesi, S. Wille, G. Pirkl, N. Wehn, and P. Lukowicz, "Monitoring household activities and user location with a cheap, unobtrusive thermal sensor array," *Proc. 2014 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. - UbiComp '14 Adjun.*, pp. 141–145.

[12] M. Syty, "Analysis and fusion of ubiquitous sensors for presence detection and people count," Master's thesis, Darmstadt University of Applied Sciences, 2016.

[13] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.

[14] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," in *Tools with artificial intelligence, 1995. proceedings., seventh international conference on.* IEEE, 1995, pp. 388–391.

[15] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.