Non-Intrusive Multi-Modal Estimation of Building Occupancy

Aveek K. Das
University of California, Davis
Computer Science
Davis, CA 95616, USA

Parth H. Pathak
George Mason University
Computer Science
Fairfax, VA 22030, USA

Josiah Jee
University of California, Davis
Computer Science
Davis, CA 95616, USA

Chen-Nee Chuah
University of California, Davis
Electrical and Computer Engineering
Davis, CA 95616, USA

Prasant Mohapatra
University of California, Davis
Computer Science
Davis, CA 95616, USA

ABSTRACT

Estimation of building occupancy has emerged as an important research problem with applications ranging from building energy efficiency, control and automation, safety, communication network resource allocation, etc. In this research work, we propose the estimation of occupancy using non-intrusive information that is already available from existing sensing modes, namely, number of WiFi devices, electrical energy demand and water consumption rate. Using data collected from 76 buildings in a university campus, we study the feasibility of multi-modal fusion between the three data sources for estimating fine-grained occupancy. In order to make the estimation model scalable, we propose three different clustering schemes to identify similarity in building characteristics and training per-cluster occupancy estimation models. The presented multi-modal fusion estimation framework achieves a mean absolute percentage error of 13.22% and we find that leveraging all three modalities provide an improvement of 48% in accuracy as compared to WiFi-only occupancy estimation. Our evaluation also shows that clustering buildings greatly increases the scalability of the proposed approach through significant reduction in training overhead, while providing an accuracy comparable to exhaustive, per-building estimation models.

KEYWORDS

Occupancy counting; Smart buildings; Multi-modal sensing

ACM Reference format:
Aveek K. Das, Parth H. Pathak, Josiah Jee, Chen-Nee Chuah, and Prasant Mohapatra. 2017. Non-Intrusive Multi-Modal Estimation of Building Occupancy. In *Proceedings of SenSys '17, Delft, Netherlands, November 6–8, 2017,* 14 pages.
DOI: 10.1145/3131672.3131680

SenSys '17, Delft, Netherlands

DOI: 10.1145/3131672.3131680

1 INTRODUCTION

Determining the number of people in a building (referred as occupancy counting) has applications spanning from building management to safety and resource allocation. In terms of building management, occupancy estimation is useful in optimizing energy demand and consumption, and Heating Ventilating and Air Conditioning (HVAC) control. It also enables novel building automation applications (for example, lighting control) that improve comfort of the occupants and enable assisted living services. Occupancy estimation plays a crucial role in ensuring building safety through intelligent surveillance and overflow monitoring. In emerging smart cities applications, occupancy is useful in predicting traffic flows between large buildings and mobility dynamics. Furthermore, in businesses and retail store buildings, it improves profitability through physical analytics. Lastly, in terms of computer networks, occupancy estimation provides information essential to adaptively perform network load balancing and network resource switching (activating/deactivating) for energy savings.

Due to its importance, occupancy counting has been investigated by a number of recent research papers [27, 31, 34, 43]. However, there are three major limitations of these approaches: (1) Existing research (such as [5, 29, 46, 52]) rely on deploying special purpose occupancy detection/estimation sensors in buildings. Examples of these sensors include PassiveInfrared (PIR) sensors, CO_2 sensors, acoustic, seismic and motion sensors. In reality, a large fraction of buildings are not equipped with such sensors, and deploying them in existing buildings is cost intensive. (2) Many of the sensing modalities used in occupancy counting are intrusive. For example, camera image based counting [50] can record every movement of occupants. Other approaches like [52] and [44] are short range and detect whether a specific room or office space is occupied or not. This type of monitoring pose serious privacy risks to occupants and can lead to user tracking. (3) Existing approaches largely ignore the issue of scalability and building diversity in occupancy counting. The proposed solutions are often evaluated for controlled settings (like laboratory) and it is not clear how they scale to large scale smart-city type of applications with a variety of buildings having diverse characteristics. Hence, there is a need of low-cost, non-intrusive and scalable solution for building occupancy counting.

In this paper, we present a novel occupancy counting solution that only relies on information already available from existing modes of sensing in buildings including electrical energy demand,

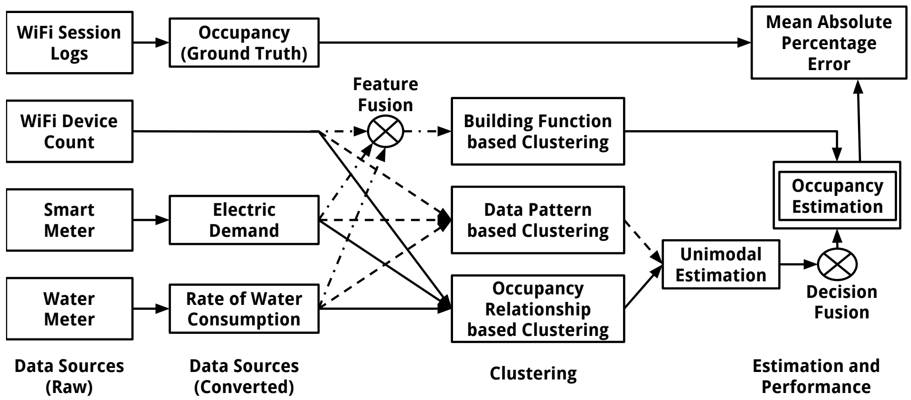


Figure 1: Our approach of estimating occupancy from three data sources (number of WiFi-connected devices, electricity demand and water consumption rate) through clustering and multi-modal fusion. The bold arrow represents the approach for best prediction accuracy.

water consumption and number of wireless (WiFi) network connected devices. Using a large-scale building dataset from a major university campus, we show that these three modalities are strong indicators of number of occupants and they complement each other in deriving accurate occupancy estimates. We build a scalable, multimodal occupancy counting framework that fuses the inferences of occupancy from each modality. The presented approach relies on three sources of information - smart metering for electricity demand, water consumption, and WiFi networks - that are readily available in most commercial and residential buildings. This eliminates the need of deploying dedicated sensor infrastructure for occupancy counting. The information used in our presented solution is only aggregate in nature (total electricity and water demand and total number of WiFi connected devices), protecting occupants' privacy. Lastly, through careful characterization and clustering of over 76 buildings in our dataset, we show that it is possible to build scalable machine learning based occupancy estimation models that can achieve highly accurate occupancy counting with limited training overheads.

Challenges and our approach: There are multiple challenges associated with occupancy counting using the three data sources. We now describe each of the challenges and how our proposed approach addresses them.

Feasibility and the need of multiple modalities: First challenge in developing our occupancy counting framework is that it is not clear how representative the three data sources are of varying levels of occupancy. Our characterization study shows that electrical energy demand and water consumption can indeed be used to estimate the occupancy, however, both the modalities can only provide coarse-grained estimation. Electricity demand is not only dependent on number of occupants but also on other factors like HVAC changes due to outside temperature and weather, building size, HVAC efficiency, etc. On the other hand, typical water meters that monitor domestic water consumption provide very coarse granularity (records one sample every 100 gallons of water in our case), further complicating the occupancy estimation.

The number of WiFi devices connected to a building's wireless network is strongly correlated to its occupancy, but depending on the context and function of a building, users can have multiple WiFi-enabled devices. This is consistent with recent studies [1, 12] which show the rise of multi-device users. More and more users are carrying more than one smart device (smartphone, tablet, laptop, etc.). Also, the number of devices carried by a user is dependent on time and context. For example, [12] shows that average number of devices per user is higher when users are in a residential environment as compared to other location contexts. This means that using number of WiFi devices for occupancy estimation requires careful consideration of building characteristics (its function and typical multi-device user behavior). We characterize the impact of occupancy on each data source using the university dataset and find that utilities data (electricity and water consumption) complement the WiFi network information in accurately estimating occupancy.

Scalability and clustering: In terms of developing and training machine learning based estimation models, it is possible to train one custom estimation model per building. Although such an approach is likely to be more accurate, it scales poorly when considering a large set of buildings in smart city type applications. Additionally, training custom estimator per building requires much longer monitoring and training periods due to sparseness of the training data. To address this challenge, we investigate three clustering techniques (as shown in Fig. 1) that group the buildings based on their characteristics, and train one estimation model per cluster. In the first technique, the buildings are clustered based on their function (e.g., cafeteria, offices, dormitories, etc.), while in the second technique, they are clustered based on the variation in the patterns of the three data sources. Both the clustering approaches are *occupancy-agnostic* which means that they do not require ground truth occupancy values while clustering. The third approach is *occupancy-aware* where clustering is based on similarity of relationships between the data sources and occupancy. It groups buildings in which values of a data source (WiFi, electricity or water) varies similarly with changes in occupancy. We show that clustering substantially increases the scalability with fewer trained models, as compared to per building modeling, while maintaining comparable estimation accuracy.

Multi-modal fusion: The third challenge is to constructively combine the three data sources such that it results in a more accurate occupancy estimation. The sparseness of the feature space associated with each modality further complicates the fusion process. We address this by devising separate multi-modal fusion models based on the characteristics of chosen clustering schemes. Specifically, we use two multi-modal fusion techniques (i) feature-level early fusion and (ii) decision-level late fusion as shown in Fig. 1. In building function based clustering, feature-level early fusion is used to form a large feature space before estimating occupancy. For the data pattern and occupancy relationship based clustering, decision-level late fusion is employed where occupancy estimation from each modality are combined through another regression model. Our results indicate that decision-level fusion techniques achieve higher occupancy estimation accuracy.

The contributions of this work can be summarized as follows:

(1) We show that it is feasible to achieve accurate *fine-grained* occupancy counting using information available from non-intrusive data sources like electrical energy demand, water consumption and number of WiFi-connected devices. Furthermore, it is demonstrated that utilities data plays a crucial role in complementing WiFi data in occupancy estimation through our proposed multi-modal fusion approach.

(2) We present novel building clustering techniques that eliminate expensive per-building training by grouping the buildings based on their characteristics. Specifically, it is shown that when buildings are clustered based on how the three data sources vary with changes in occupancy, the resultant per-cluster occupancy models can achieve highly accurate occupancy estimation with decision-level fusion.

(3) We evaluate the clustering and multi-modal fusion with data collected from 76 buildings over 4 weeks duration with occupancy varying from 0 to 550. It is observed that when buildings are clustered based on similarity in occupancy-data source relationship, the resultant occupancy estimation with three data sources achieve a mean absolute percentage error of 13.22%. We observe that use of multimodal data provides significant improvement in accuracy, almost 48% reduction in error as compared to the model built using only WiFi data.

The rest of the paper is organized as follows. We introduce the dataset in Section 2 and discuss the feasibility of multiple modalities in Section 3. In Section 4 we discuss the different schemes based on which we cluster the buildings in our dataset. Section 5 discusses the multimodal fusion schemes for estimation and the results are discussed in Section 6. After presenting related work in Section 7, we conclude the paper in Section 8.

2 DATASET AND METHODOLOGY

The first major step in building the occupancy estimation model is collecting data from multiple buildings with diverse characteristics (for example, building function and context which affect how WiFi and utilities are used). For this, we collect data from a university campus for a duration of four weeks (28 days). The dataset is described below and is also summarized in Table 1.

2.1 WiFi Network Data

WiFi Device Count: The number of WiFi-connected devices is an important indicator of number of occupants. Our WiFi network data for the university campus includes *aggregate device count* for each building. The aggregate device count is the number of devices (unique MAC addresses) connected to a building's WiFi Access Points (APs). These devices include any WiFi-capable device (e.g., smartphone, tablet, laptop, etc.). The per-building device count is maintained by the network administrator for troubleshooting and management purposes. We query this database once every 10 minutes for each building.

Data Type	No. of Buildings	Sampling Rate
WiFi Device Count	76	6 samples/hr
Electricity Demand	56	360 samples/hr
Domestic Water Consumption Rate	19	180 samples/hr
Occupancy (Ground Truth)	76	6 samples/hr

Table 1: Primary data sources and sampling rates for a university campus dataset used in our occupancy estimation

Ground truth occupancy: The WiFi network data also includes *WiFi session logs* which record the start and end times of each device's connection to an AP. The other relevant information in the logs are: <IP address, MAC address, connected AP, Username>. The AP name, when combined with the WiFi AP deployment map (discussed in Section 2.3) can identify the building where each session is created. As highlighted earlier, most users carry more than one WiFi enabled device. The username field in the session logs can be used to identify users carrying multiple devices. The number of unique users (or usernames) at a specific building represents the actual number of occupants of the building. This occupancy value, in all instances, is lower or equal to the number of WiFi devices (as collected above). In this work, this value is used as the *ground truth occupancy* in all instances. Based on the session logs, the occupancy count per building is calculated for 10 minute windows. Session logs are collected from a total of 76 different buildings on campus. We use the 10 minute interval as the granularity for all the occupancy estimation that we report in this paper.

Devices that are not carried by users when leaving the building create sessions of very long duration. In addition, since the WiFi network is not limited within the physical confines of the building, some transient passerby users can create sessions of very short duration. In order to not count such instances in our ground-truth, we remove sessions that are longer than twenty-four hours or shorter than five minutes. The ground truth occupancy does not include a small fraction of the occupants in a building who do not connect to the wireless network or who connect their devices only to the wired network. This fraction of users, which we miss in our ground truth estimation and can only be estimated using dedicated sensors is a limitation of our approach. Also, our method of ground truth occupancy estimation depends on username based login information which might not be available in buildings that employ password based login. However, the username based login provides us an accurate estimate of occupancy for developing our multimodal estimation models.

The wireless data collection was performed in collaboration with Information Technology Department of a university campus under Institutional Review Board (IRB) approval. We collaborated with the department to anonymize the collected network logs to remove any personally identifiable information before using it in our study. Specifically, we anonymize the IP addresses, MAC addresses and usernames. We employ prefix-preserving anonymization as proposed in [17]. The anonymization methods and parameters are kept consistent over all logs.

2.2 Utilities Data

General purpose electricity and water meters which are commonly used in commercial buildings are deployed in university buildings. In collaboration with the campus utilities department, we collect electrical energy demand and water consumption data from these meters. This information is then maintained by the utilities department with the aim of better energy conservation. All the data is accessed through a central university server using a RESTful API.

(1) Smart Meter Electricity Data: The electrical energy consumed by a specific building is referred to as its electricity demand. This includes electrical energy consumption from infrastructure-based setups, like corridor lights and HVAC systems, energy consumed by user plug-load devices, like personal computers, phones, and televisions and energy expenditure of shared common space devices like refrigerators and microwave ovens. These electric consumption values from various buildings are collected using smart meters and stored in a central university database. This dataset collects energy demand (in KiloWatts) at each meter at a 10 second interval. In some instances, more than one meter is deployed per building, each accounting for specific sections or floors. The data from multiple meters is aggregated together to calculate the total demand of the building. There are also some instances, in which one meter is used to record information from multiple buildings. Such buildings are ignored in this study due to unavailability of accurate per building smart meter data. The smart meter data is collected at the aforementioned 10 second interval, from 56 buildings on the campus.

(2) Domestic Water Consumption Data: Domestic water is the water directly used by occupants in a building (like restroom usage, kitchen use, etc.). Other usage of water, like water used for sprinkler functioning outside a building, are not considered as domestic water. Data collection of domestic water is different from the aforementioned electric meter. Instead of reporting a continuous stream of water consumption rate, the meter records instances whenever 100 gallons of water is used in the building. We use the data to calculate the mean rate of water consumption (gallons per minute). Since the buildings with water meters on campus - connected to the central server - are fewer in number than smart meters, we collect data from 19 buildings on campus.

2.3 Auxiliary Data

In addition to this primary data, we also collect necessary auxiliary data including building type and size. We also collect information regarding the outside weather as it governs HVAC load. The auxiliary information collected as a part of this work are:

Building Function	Count
Classrooms	15
Laboratories	21
Offices	15
Dormitories	15
Cafeteria/Dining	5
Health Center/Hospital	3
Special Use	2

Table 2: Different building functions and their count for the 76 buildings in our campus dataset

Building function (or type): A university campus closely resembles a smart-city type environment where buildings have diverse functions and characteristics. As we demonstrate later, the function of a building has significant impact on its occupancy pattern. Since the function information is readily available, it can be exploited for clustering and training occupancy estimation models. As a result, building function data is an important part of our study. The types of buildings on campus include among others, laboratories, dormitories, office buildings, cafeterias and classrooms. Table 2 lists the different building types in our dataset of 76 buildings and their count. The special use buildings include the campus gymnasium and the performance theater.

Building size: Larger buildings have a higher baseline energy consumption (energy consumed in absence of any occupants). This is due to higher HVAC load and other factors such as corridor lighting and other electrical appliances. To compare the occupancy patterns of buildings with varied sizes, the building size information is used in normalizing based on area. As a result, we collect information of total area of a building in square foot. The area information includes all floors of the building.

Weather data: The weather dictates the inside building temperature, which in turn drives the HVAC load and its electrical energy consumption. As a result, relevant weather data is collected using the OnPoint API service provided by Weather Source [2]. We calculate the air temperature (in °F) and relative humidity for the campus location at the rate of 60 samples per hour.

AP - Building Map: In order to determine the building where an AP is located, we collect information that maps each AP on campus to a specific building. As explained above, this information is used in conjunction with other WiFi network data to calculate the ground truth occupancy.

3 FEASIBILITY AND THE NEED OF MULTIPLE MODALITIES

In order to establish feasibility of occupancy estimation based on the three data sources, it is first necessary to evaluate that each of these sources individually is correlated to occupancy. In order to do this, we first calculate the ground truth occupancy at 10 minute intervals over the duration of our dataset. For the same 10 minute intervals, the WiFi device count and the mean of electrical energy demand and water consumption rates are calculated. We calculate the building-wise correlation coefficient between the occupancy and each of the data streams. The calculated values are shown in Table 3.

Data Source	Mean Correlation	Std. dev. Correlation
WiFi Device Count	0.878	0.097
Mean Electricity Demand	0.696	0.241
Mean Water Consumption Rate	0.455	0.281

Table 3: Mean and standard deviation of correlation between each of the three data streams and occupancy

We observe that the WiFi device count and electrical energy demand are highly correlated with the occupancy, whereas the water meter reading has loose, yet positive, correlation. This implies that the change of each data stream and occupancy follows a similar trend. However, since correlation does not imply causality and the values of three data sources can be dependent on other factors, we model their relationship with occupancy in Section 4. Also, since WiFi device count is highly correlated with occupancy, we now turn our focus on determining its sufficiency in occupancy estimation.

3.1 Why is single modality not sufficient?

It might appear at first that WiFi device count can be used to accurately estimate the occupancy, however, there are a number of factors that can lead to incorrect estimation using this approach.

Multi-Device Users: Using only the number of wireless devices at each location at a specific time to estimate the occupancy at that location will lead to over-estimation due to many users having more than one device. Nevertheless, if the average number of devices per user is constant at multiple locations over different time points, we can use a *proportionality factor* to estimate the occupancy by simple multiplication. This, we observe, is indeed not the situation.

We define *device-user ratio* as the ratio of WiFi device count and number of occupants. We calculate the average value of device-user ratio for each building and plot the distribution (minima, first quartile, median, third quartile and maxima) of the average device-user ratio values for the different buildings belonging to the same function. Four different building functions are shown via a boxplot in Fig. 2. The difference in the distributions of the average device-user ratio among the different building types is a direct effect of users' behavior at each building type. It can be observed that classrooms have the least average number of wireless devices carried per user. Since most of the occupancy in classrooms are governed by students attending classes, there are very few instances of user accessing multiple devices at the same time. On the other hand in dormitories where users are in a residential setting, the number of devices per user increases significantly due to the use of devices that users do not always carry around. For labs and office spaces, this number is higher than classrooms but lower than dormitories. This could be a result of the fact that many of users' laptops or computers are connected through wired networks. The different building types vary from the point of view of their *context*. Due to this contextual difference between buildings using the proportionality factor cannot correctly estimate occupancy. Even the buildings belonging to the same building functions have significant difference in the mean device-user ratio, thereby signifying that a per-building type proportionality factor would not be enough for estimation.

In addition, we plot in Fig. 3 the device-user ratio variation over five working days for four different types of buildings on campus.

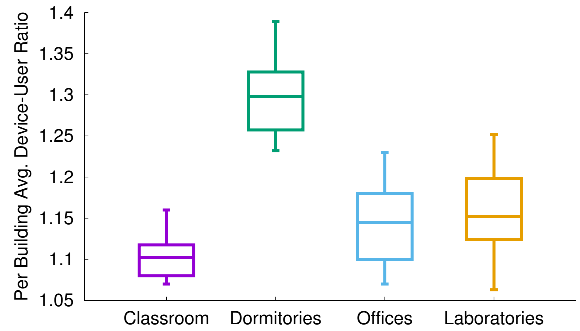


Figure 2: Distribution of per-building mean device-user ratio represented in the form of <min, first quartile, median, third quartile, max> for the four major types of buildings on campus.

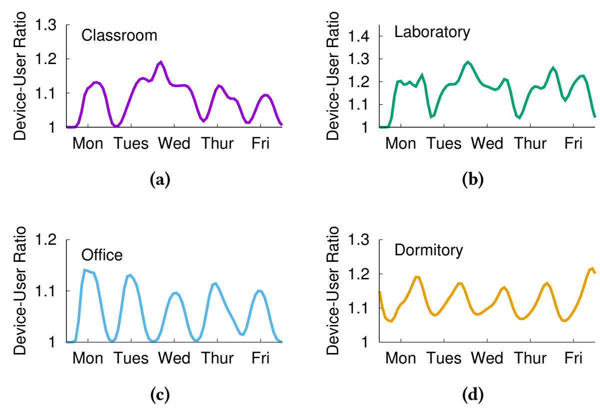
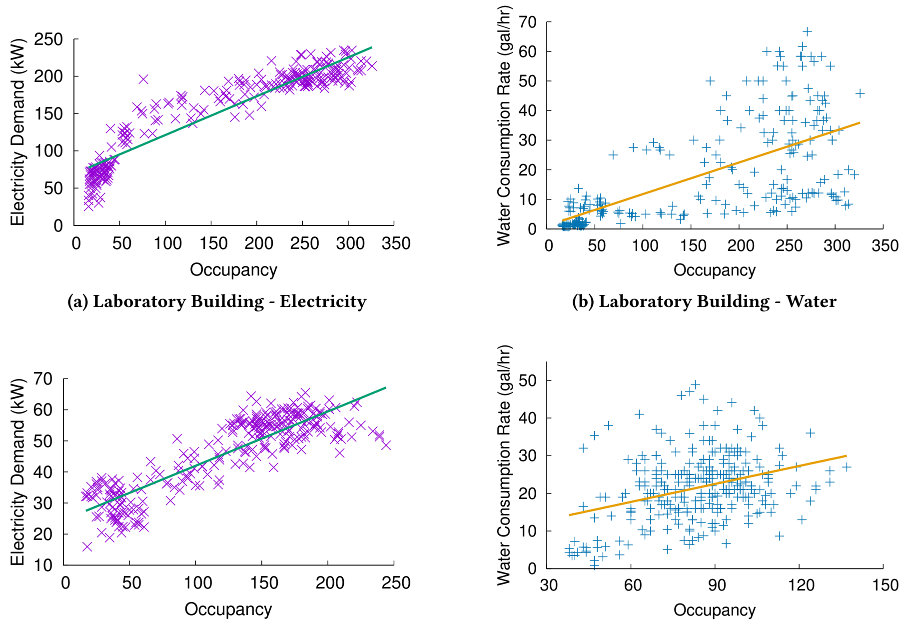


Figure 3: Device-user ratio pattern for four different types of campus buildings in a week; the ratio varies based on contextual factors (building function and time)

It can be observed that the device-user ratio values are continually different based on the time of the day. The pattern of the device-user ratio variation throughout the day is representative of the type of building in question. Some locations have higher values of device-user ratio during working hours, while the values drop down to 1 (one device per user) during the night time. Dormitory buildings, due to the presence of overall larger number of wireless devices, always have a device-user ratio greater than 1 and the change in the value is more gradual. We can therefore validate that the device-user ratio is not only dependent on the building function, but also on the time of the day. Based on these observations, it is difficult to derive a fixed *proportionality factor* to directly estimate occupancy from the count of wireless devices. Any sort of factor, if used, has to be trained based on the building type, the time of the day and possibly other contextual factors.

Another additional information that hinders occupancy estimation only from the point of view of WiFi device count is the fact that some users might not carry wireless devices, or not connect to the WiFi network. As a result, the access point would have no record of such devices, and in return, not count the users towards estimating



(a) Laboratory Building - Electricity

(b) Laboratory Building - Water

(c) Classroom - Electricity

(d) Cafeteria - Water

Figure 4: Variations in utilities data for occupancy instances that have the same WiFi device count, showing that the utilities data can complement WiFi data in estimating occupancy

occupancy. This is a limitation of our proposed approach, as we cannot account for users without any connected wireless device.

3.2 Complementing WiFi with utilities data

With the knowledge that the electricity demand and water consumption data are both correlated to the occupancy, we further investigate if the utilities data can be useful in estimating occupancy when the WiFi device count is not sufficient. To evaluate this, we investigate the variation in utilities data with changing occupancy for instances where device-user count is constant. This is demonstrated in Fig. 4.

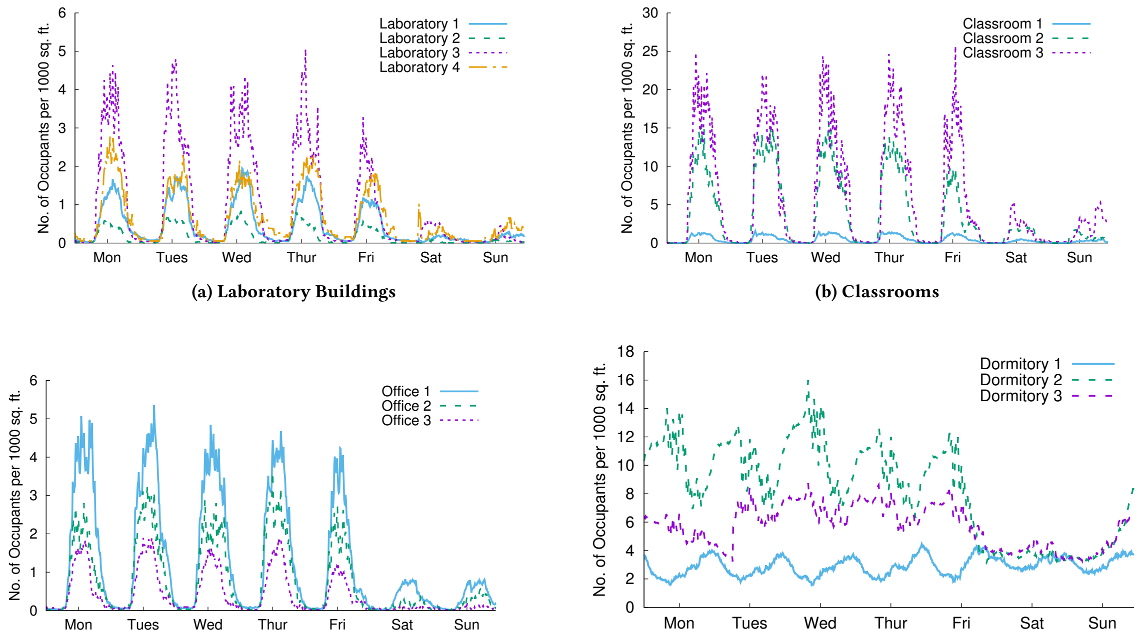
The pattern of variation of electricity demand in Figs. 4a and 4c show that for two different types of buildings on campus, electrical energy demand increases with occupancy (the relationship is approximated using a linear fit). This indicates that even with the same device-user ratio values, the presence of smart meter data can provide useful information towards the estimation of occupancy. The pattern of variation of water consumption rate with occupancy, observed in Figs. 4b and 4d, also indicates a gradual increase in consumption rate with occupancy. The slope for the fitted regression line is smaller as compared to the electricity data instances, which is in line with lower correlation values.

Based on these observations, we can conclude that when WiFi device count is not enough for estimating occupancy, the utilities data can complement it. This motivates us to develop a multi-modal fusion based occupancy estimation framework which we present in Section 5. Our evaluation in Section 6 further confirms that use of multiple modality can indeed result in more accurate occupancy estimation.

4 CLUSTERING FOR SCALABILITY

Since our objective is to estimate building occupancy based on the three data sources, one possible solution is to develop an estimation model (based on machine learning algorithms) for each building. This approach is likely to result in better estimation accuracy. However, it scales poorly in scenarios like smart cities where there can be hundreds of buildings. This scalability issue motivates us to cluster buildings based on similarity in their characteristics, and develop per-cluster occupancy estimation models. Depending on the chosen characteristics and accuracy of clustering, such per-cluster models are more scalable and likely to provide reasonable estimation of occupancy. We divide the clustering methods into **occupancy-agnostic clustering** and **occupancy-aware clustering**.

The motivation behind this categorization is that a building, for which corresponding occupancy estimation model is not trained, can be assigned to a cluster first (based on similarity criteria described below) and then corresponding cluster-specific model can be employed for estimation. It is possible that ground truth occupancy is not readily available for the untrained buildings as measuring occupancy requires costly deployment of sensors at the entry/exit points or motion sensors to detect presence of an individual in different rooms. To cluster such buildings, it is desirable that characteristics other than occupancy are considered in clustering, specifically, building's primary function and daily or weekly variation pattern of the three data sources. We refer to such a clustering technique as occupancy-agnostic clustering. If the occupancy information is available for the untrained buildings, the knowledge of building's true occupancy and its relationship with the three observed data sources is used for occupancy-aware



(a) Laboratory Buildings

(b) Classrooms

(c) Office Buildings

(d) Dormitories

Figure 5: Weekly occupancy patterns of the four building types in the campus environment shows that clustering based on building type/function clusters buildings with similar occupancy pattern but not similar absolute normalized occupancy

clustering. We use k-means clustering algorithm and optimize the number of clusters using the change in the sum of squared errors. We iteratively increase the number of clusters (while limiting the maximum cluster count to less than 10) until we do not further observe a significant decrease in the error value.

4.1 Occupancy-agnostic clustering

Occupancy-agnostic clustering makes use of any information other than occupancy for clustering. We explore two types of occupancy-agnostic clustering as described next.

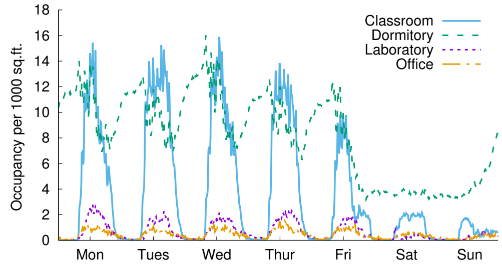
(1) Clustering based on building function: As shown in Table 2, the buildings on campus have different functions (e.g., classroom buildings, office buildings, dormitories, etc.). The intuition behind using a building's function as the criterion for clustering is that buildings with similar function are likely to have similar behavioral pattern in terms of occupancy, and hence the three data sources (WiFi, electrical energy and water consumption). For example, buildings serving as cafeteria or dining halls are likely to be more occupied during breakfast, lunch and dinner times. Since the building function is already known, such a clustering does not require any additional data analysis.

Fig. 5 shows occupancy patterns for a week for buildings of four different functional categories (laboratory, classroom, office and dormitories). In order to remove the impact of building size (area and number of floors), normalized occupancy (no. of occupants per thousand square feet) is shown. It can be observed that (as expected) each building has a repetitive pattern over the different days of the week - with occupancy being lower during the last two days (weekends). In the classrooms, laboratories and offices, there is a very clearly distinguishable higher occupancy period during "work-hours" and lower occupancy outside work hours. Dormitories do not have such a distinct difference and has peaks during nighttime. However, even after normalization based on the size, the same type of buildings do not have similar normalized occupancy. Also, buildings falling in different clusters can have similar absolute normalized occupancy values(for example, Laboratory-1 and Office-3).

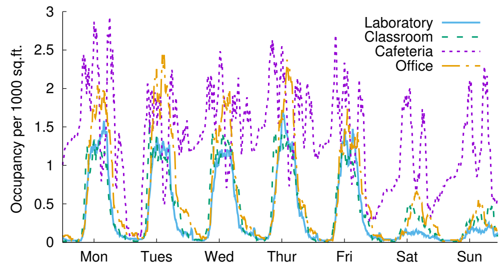
This way, the function based clustering is effective in clustering buildings with similar variation in *occupancy pattern*. However, it cannot accurately cluster buildings that have similar values of *absolute normalized occupancy*.

(2) Clustering based on data pattern: We investigate another occupancy-agnostic clustering where buildings are clustered based on the similarity in patterns of the three data sources. The intuition here is that since occupancy is correlated with the data sources (WiFi device count, electrical energy demand and water consumption rate), this type of clustering will group the buildings that have similar occupancy. The benefit of this approach is that, similar to function-based clustering, no information about actual occupancy is required for categorizing the buildings. The clustering can be performed through monitoring the three data sources over a predefined period of time.

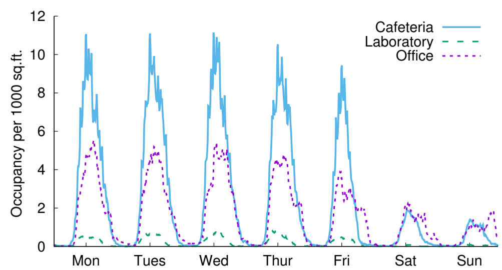
To evaluate the clustering, we calculate the mean data reading, normalized for size, every 10 minutes for a total of four weeks for each building in our dataset. Based on this data for each building, we perform k-means clustering. The clustering is performed separately



(a) A cluster of buildings based on electric demand pattern



(b) A cluster of buildings based on WiFi device count variation



(c) A cluster of buildings based on water consumption

Figure 6: Buildings when clustered based on the weekly patterns of the three data sources, do not necessarily have similar occupancy patterns even when they belong to the same cluster.

based on each data source, that is, one particular building is a part of 3 different clusters, one based on each data type. Fig. 6 show the occupancy of buildings of the same cluster for each of the three data sources. We observe that although buildings can have similarity in their patterns of number of WiFi devices, electrical energy demand and/or water consumption, their true occupancy can only be loosely similar.

4.2 Occupancy-aware clustering

In cases when occupancy of buildings is available for clustering, relationships between the data sources and occupancy can be modeled, and buildings with similar relationships can be clustered. After training separate models of each cluster, the occupancy can be estimated using the three data sources for each building. Different from the occupancy-agnostic approach, the occupancy values have to be monitored through additional infrastructure (for example, by deploying sensors) for certain amount of time. On the other hand,

Relationship with Occupancy	Mean R^2 Score	Std. Dev. of R^2 Score
WiFi Device Count	0.845	0.118
Electricity Demand	0.579	0.207
Water Consumption Rate	0.382	0.149

Table 4: Mean and standard deviation of R^2 score for linear regression model of different buildings; linear model approximates the relationship between a data source and occupancy

the availability of true occupancy helps in development of more accurate clustering.

In this type of clustering, we first model the relationship between individual data sources and occupancy for each building and thereafter, cluster the buildings that have similar relationships.

Modeling relationship between individual data sources and occupancy: We apply linear regression separately on number of WiFi devices, electrical energy demand and water consumption data for each building using its occupancy. We note that number of WiFi devices and (domestic) water consumption are directly dependent on number of occupants. Higher occupancy likely results in more devices and more water consumption. However, in addition to occupancy, there is no other building-specific attribute that directly affects values of water consumption and WiFi device count. The linear regression model based on these two data sources is based on only one input factor - occupancy.

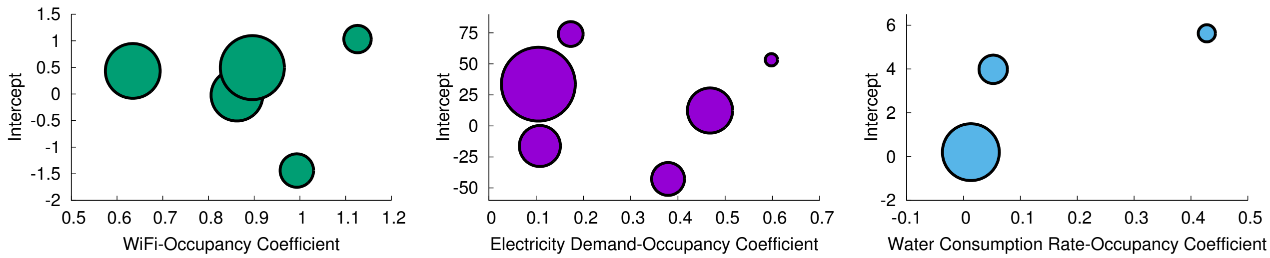
On the other hand, electrical energy demand is also a function of other factors including outside weather conditions. Depending on the temperature, the building HVAC can consume more or less to maintain the desired indoor temperature and ventilation. To account for this, we also include our weather auxiliary data (Section 2) in linear regression of energy demand in the form of outdoor temperature. The building size is invariant and modeled in the form of intercept in the linear regression. The linear regression model based on electrical energy demand for each building is thus based on two input factors - occupancy and outside air temperature.

We model the relationship of a data source with occupancy using linear regression. For WiFi and water, this takes a form of $ax + c = z$ where a is the occupancy coefficient, x is the occupancy, c is the intercept (building dependent) and z is the value of data source (i.e. WiFi device count or water consumption rate). For electrical energy, the linear model take a form of $ax + by + c = z$ where y is the outside air temperature, b is the temperature coefficient and z is the electrical energy demand. We do not use these models for actual occupancy estimation as they do not involve multimodal information and are based off just one of the three data sources.

To estimate the *goodness of fit*, the coefficient of determination or R^2 score for each building's regression model is calculated as

$$R^2 = 1 - \frac{\sum_i (r_i - \hat{r}_i)^2}{\sum_i (r_i - \bar{r})^2} \quad (1)$$

where r_i are the original data values, $\hat{r}_i$ are the predicted data values, and $\bar{r}$ is the mean of the original data values. Values of R^2 score closer to 1 indicates a model which represents a good fit. Corresponding values for our regression models are shown in Table 4.



(a) WiFi Device Count (b) Electricity Demand (c) Water Consumption Rate

Figure 7: Clusters formed based on the linear regression relationship between occupancy and each of the three data sources

Data Type	Cluster No. (Size)	Classroom	Laboratories	Offices	Dormitores	Cafeteria	Hospital	Special Use
WiFi Device Count	1 (17 buildings)	3	2	6	1	1	2	2
	2 (18 buildings)	5	8	2	-	2	1	-
	3 (11 buildings)	-	5	3	3	-	-	-
	4 (21 buildings)	7	6	4	2	2	-	-
	5 (9 buildings)	-	-	-	9	-	-	-
Electricity Demand	1 (8 buildings)	1	5	2	-	-	-	-
	2 (10 buildings)	2	2	2	3	-	-	1
	3 (11 buildings)	1	6	2	-	1	1	-
	4 (18 buildings)	5	2	3	7	1	-	-
	5 (6 buildings)	1	3	-	-	1	-	1
	6 (3 buildings)	-	2	-	-	-	1	-
Water Consumption Rate	1 (3 buildings)	1	-	1	-	1	-	-
	2 (9 buildings)	1	3	-	-	2	2	-
	3 (6 buildings)	1	3	2		-	-	-

Table 5: Clustering based on similarity of relationship between data sources and occupancy results in clusters with diverse type of buildings

Clustering based on occupancy relationship: The occupancy coefficient (a) and intercept (c) calculated from the three regression models of each building can now be used for clustering. Note that we do not use the temperature coefficient (b) from electrical demand regression in clustering. Fig. 7 shows the three sets of clusters (one set for each data source) after applying the K-means clustering. The size of the circle indicates the number of buildings in the cluster. Table 5 shows how each cluster obtained using this method contains a mix of different types of buildings (e.g., offices, dorms, etc.). It shows that clustering buildings based on similarity of relationship between data sources and occupancy results in drastically different clusters compared to occupancy-agnostic function or pattern based clustering.

We will show in Section 6 that the relationship based clustering models are more accurate in occupancy estimation compared to occupancy-agnostic models. However, they require occupancy values in order to model the relationships.

5 MULTIMODAL FUSION FOR OCCUPANCY ESTIMATION

Based on the clusters calculated using the techniques discussed above, it is now possible to train an occupancy prediction model for each cluster. One important design challenge before this is to decide how the data from three sources can be fused for training. Multimodal fusion has been studied extensively in machine learning and data mining literature for analyzing and combining multiple data sources together. In recent times, multimodal fusion has been applied to a variety of data including multimedia [6, 49]. In general, multimodal data can be fused using two following ways (refer to Fig. 8):

(1) Feature-level (early) fusion: This type of fusion involves combining the features extracted from the various data sources before learning from the data [20, 35, 36]. The advantage of early fusion is that it requires only one training phase that is performed after the combination of the features. However, a disadvantage is that if the data sources are significantly different, it is possible that the features extracted from each cannot be represented in a similar, fusible format.

(2) Decision-level (late) fusion: The data from each modality is used to create a unimodal learning model and the decisions obtained from the unimodal models are merged in the late fusion approach. This involves creating multiple learning models - one for each modality and one for the final decision fusion [3, 53]. This type of fusion works even for significantly different input feature space as the decisions from each dataset still have similar representation.

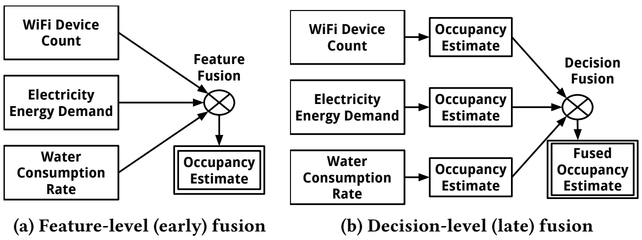


Figure 8: Two types of multimodal fusion; feature-level fusion is used with building function-based clustering, and decision-fusion is used with data pattern-based and occupancy-relationship based clustering

Common methods for decision fusion involve rule-based methods [24], like weighted-sum, max-min or voting. Classification algorithms like Support Vector Machines, Logistic Regression, Bayesian Networks, etc. have also been used.

5.1 Feature Space

In our system, the occupancy estimation is performed every 10 minutes. As a result, the following feature space for estimation is built based on consecutive 10 minutes intervals.

- *Number of WiFi devices:* As the interval of calculating the number of wireless devices at each location is equal to the interval of estimation, we use the *wifi device count* as the feature corresponding to the network data.
- *Electric energy demand*, which is collected every 10 seconds, has 60 samples in the 10 minute interval. We calculate *<min, max, mean, standard deviation, range, sum of absolute differences>* based on these 60 samples.
- *Water consumption rate*, calculated every 30 seconds, has 20 samples in our time interval. We calculate the same 6 features as the electric demand data.

5.2 Occupancy Estimation Models

Using the feature space and multimodal fusion techniques described above, we now train three occupancy estimation models for the three type of clustering described in Section 4.

[M1] Building function based clustering and early fusion: Since building function based clustering results in only one set of clusters (not three sets as with other types of clustering), we use early fusion where the features from each modality are combined to form one feature vector. All the buildings in each cluster are then combined to train a per-cluster linear regression model for occupancy estimation.

[M2] Data pattern based clustering and late fusion: The clusters are created based on the patterns of each individual data source (WiFi, electrical energy and water consumption). As a result, one building can be part of three different clusters (one for each data source). This makes feature-level fusion unsuitable in this scenario. For each data source, we train a cluster-specific regression model (based on all the buildings in the cluster) and arrive at a data-source based decision. The individual decisions from the models of all the three data sources are further combined using linear regression to estimate occupancy. We note that in the decision-level fusion stage, learning is based on all the buildings as there is no clustering involved in fusion.

Type of Clustering	Type of Fusion	1 Source (WiFi)	2 sources (WiFi & Electric)	3 sources (WiFi, Water & Electric)
Building Function [M1]	Feature Level	27.3%	25.82%	24.4%
Data Pattern [M2]	Decision Level	30.13%	23.7%	19.49%
Occupancy Aware [M3]	**Decision Level**	**24.14%**	**14.59%**	**13.22%**
No Clustering (Per Building Model)	Feature Level	22.73%	13.91%	13.69%

Table 6: Mean Absolute Percentage Error values for occupancy estimation. The occupancy-aware clustering with decision level fusion yields lowest estimation error among the three clustering based models (M1, M2, M3)

[M3] Occupancy-aware clustering and late fusion: Similar to the data pattern based clustering, since clusters are based on relationship between occupancy and each individual data source, we cannot perform feature-level early fusion. The decisions obtained from the cluster-specific regression models for each data source is linearly combined (for all buildings) based on regression coefficients for occupancy estimation. This approach is identical to the previous approach with the only difference being the mode of clustering. Decision-level fusion allows for easier accommodation of cases when one or two of the data sources are not available (for example, buildings not equipped with water meters).

6 EVALUATION

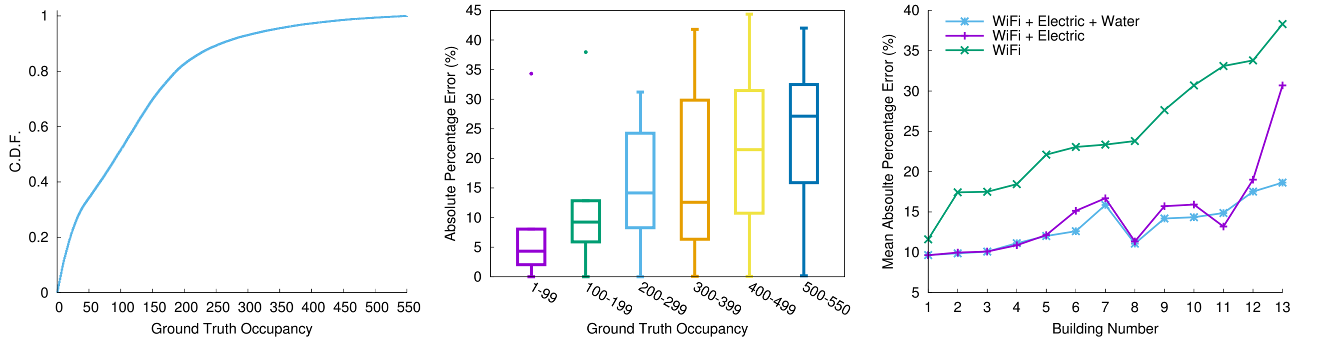
We now evaluate the clustering and fusion based occupancy estimation models developed above using our dataset.

Performance Metric: To measure the accuracy of our estimation models, we use *Mean Absolute Percentage Error* (MAPE) per building, which is calculated as follows:

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{A_i - E_i}{A_i} \right| \tag{2}$$

where, A_i is the ground truth occupancy, E_i is the estimated occupancy and n is the number of predictions per building. We use this metric instead of using the Mean Absolute Error, as the latter does not capture the degree of the error compared to the true occupancy. Fig. 9a shows that the range of ground truth occupancy for all the buildings under consideration varies from 0 to 550.

After training the regression models for each cluster of buildings we test the models on a per building basis using 10-fold cross validation. Overall our dataset is four weeks (28 days) long and has 144 samples of occupancy per building per day. We calculate the MAPE value for all the estimation instances of a specific building



(a) Distribution of ground truth occupancy for all buildings in our dataset

(b) Distribution of absolute percentage error value represented as <min, first quartile, median, third quartile, max> for different occupancy instances in occupancy aware clustering and late fusion (M3)

(c) Decrease in estimation error with inclusion of electrical energy and water consumption demonstrated for 13 representative buildings

Figure 9: The proposed occupancy estimation framework with clustering and multimodal fusion results in low estimation error especially in presence of all three data sources. It is observed that more than 90% of tested instances have occupancy estimation error of less than 20%

and use the average to represent the performance of our system for that building. The accuracy is calculated based on three different sets of buildings according to the availability of the data sources.

- A total of 19 buildings in our dataset have all the 3 sources of data (WiFi, electricity and water). For these buildings, all the **3 sources** of data are used to estimate occupancy.
- 56 buildings on campus with at least **2 sources** of data, specifically, electricity data and WiFi data. The results reported for these 56 buildings are based on models built on the basis of two data sources.
- We build estimation models based on **1 source** of data, the wireless device count, for all the 76 buildings on campus.

Occupancy estimation accuracy: Occupancy estimation results for the aforementioned cases are shown in Table 6. The results show that using multi-modal data to estimate the occupancy is actually beneficial as increase in the number of sources directly result in a reduction of error. It can be observed that clustering the buildings based on the relationship of data source to the occupancy (M3) produces the least error. This confirms that occupancy estimation models should be developed for buildings that have similar data source and occupancy relationships. The percentage error values for the occupancy instances, divided into bins of 100, for all the buildings in this model are shown in Fig. 9b. We can observe that the maximum value of MAPE for any instance, when occupancy-aware clustering is employed, is less than 45%. Overall, we observe that the absolute percentage error increases for higher occupancy instances. In the occupancy-agnostic methods, data pattern based clustering yields better estimation accuracy compared to building function based clustering. This reflects that even though it is occupancy agnostic, data pattern based clustering in fact performs better in grouping buildings with similar occupancy patterns compared to building function based clustering.

For a new, untrained building, we can deploy dedicated sensors for a short duration of time (for example, two weeks) and collect ground truth data. Based on this ground truth information, occupancy aware clustering can be performed. Depending on the cluster this new building belongs to, we can perform late fusion (M3) using a pre-trained model for occupancy estimation for a longer duration (beyond the two week period). On the other hand, if it is not feasible to get ground truth occupancy values for a new building, we can use the WiFi and utilities data pattern from that building to find which cluster the building belongs to. Thereafter, a pre-trained model corresponding to the cluster can be chosen and decision level fusion (M2) can be employed for occupancy estimation. In most practical scenarios, where occupancy data cannot be retrieved easily, this model (M2) could be used to calculate occupancy with a minor compromise in accuracy.

Data Sources in Model	Data Sources in Compared Model	% Improvement
3 Sources	2 sources	**9.4%**
3 Sources	1 source	**48.3 %**
2 Sources	1 source	**43 %**

Table 7: Improvement in the estimation accuracy with addition of data sources to occupancy estimation model

Impact of multiple modalities: The improvement in accuracy as we move from 1-source based model to the 3-sources based model for the occupancy-aware clustering is shown in Table 7. This variation in the per building MAPE values for a few representative buildings are shown in Fig 9c. It can be observed that addition of electric demand information to the WiFi data produces significant improvement in the performance (improvement of 43%). It also confirms that the presented multi-modal fusion techniques are effective in supplementing WiFi with electrical energy demand. The addition of the third source (water consumption) results in decrease of estimation error (improvement of 9.4%), albeit not significantly. As a result, in buildings without water metering available, the presence of smart-meter in addition to WiFi can provide reasonable accuracy for occupancy estimation.

Length of Data	3 sources in model	2 sources in model	One source in model
1 week	16.7 %	17.79 %	24.08 %
2 weeks	**13.89** %	**15.41** %	**24.47** %
4 weeks	13.22 %	14.59 %	24.14 %

Table 8: Increasing the duration of data available for monitoring and training increases reduces the estimation error but the improvements are limited beyond two weeks

Impact of clustering: One of the advantages of our occupancy estimation approach is the reduction in number of prediction models that are required to be trained. In the instance of building function based clustering, the number of trained models are equal to the number of clusters (five in our dataset). For clustering with decision level fusion, the unimodal decision making phase involves training models equal to the total number of clusters in all three modalities. In addition, another model needs to be trained for the decision fusion. In our model, there are five clusters for WiFi data, six for electric data and three for water data - resulting in 15 models including the decision fusion. On the other hand, if estimation models were built for each building, our dataset would end up having 76 different models.

Table 6 shows that our clustering-based models in fact achieve similar accuracy to that of per-building models albeit with significantly fewer trained models. Also, the use of cluster-specific models is particularly beneficial to untrained buildings in reducing the training overhead.

Impact of training duration: We now evaluate the impact of duration of data available/monitored while clustering and training on estimation accuracy. In case of occupancy aware clustering and decision level fusion (M3), it is required that data from the three sources and the ground truth occupancy are monitored for a significant period of time. The results presented in Table 6 are based on data collected from the sources for a duration of four weeks. However, in practice, it is beneficial if the monitoring/data collection duration can be reduced while ensuring reasonable accuracy. To this end, we re-cluster and retrain our models with data collected for varying durations. Table 8 show the error values as the training data duration varies from one week to four weeks. We observe that increasing the training period from one week to two week results in noticeable increase in accuracy which is likely due to improved clustering. However, further increasing the training duration from two weeks to four weeks does not provide comparable improvements. This suggests that biweekly pattern is sufficient for training our estimation models. However, our models do not consider seasonal patterns or changes due to holidays or breaks in the university. Further investigation using data collected over a longer duration (over a year) is necessary to observe how our models perform during a different season or during the summer break. We leave this exploration to future work.

7 RELATED WORK

Occupancy Detection and Estimation: Estimating building occupancy has benefits from the point of view of energy-efficiency, crowd control, building security, etc. The major challenge in efficient occupancy detection is finding a method that is inexpensive (does not require costly sensor deployment) and reliable (can be implemented across buildings with diverse characteristics).

Utility Data: One traditional approach of non-intrusive occupancy detection relies on observing the energy usage of a building. This data usually comes in the form of monitoring electricity usage from smart meters [11, 26, 27]. These approaches monitor the change in the smart meter readings to detect if a building is in an occupied state. Real-time electricity data can be used for effective real-time occupancy detection which can be used for building automation systems. The effect of occupancy on electricity data variation is explored in [33] and [15] and these works show that occupancy is just one of the various factors that effect the smart meter readings. Utilization of smart meter is popular because of its ubiquity and accessibility in commercial and residential buildings. Although detecting whether a building (or building zone) is occupied can be achieved accurately using smart meter data, it is not viable to determine the number of occupants. To counter this problem, [51] has combined motion-based sensors with electricity data to detect number of occupants in a house. Similarly, [22] utilizes an individual power monitoring system in conjunction with ultrasonic sensors, motion sensors and WiFi access points within a commercial building setting to detect user presence.

Wireless Networks: A number of research efforts have utilized information available from the WiFi network to count occupants in a building. This is an attractive method because most commercial and residential buildings already have WiFi infrastructure in place and therefore do not require additional installation or cost. In [18, 19], WiFi has been used to detect pedestrian flow within a large area by tracking the broadcasted MAC address of mobile devices of pedestrians. Analyzing pedestrian flow can provide useful insights during times of disaster and also assist in urban planning. The presence of multiple users affect the WiFi signal strength and the Channel State Information values. Authors in [13, 37] exploit this change in measurements and train prediction models to calculate occupancy in smaller areas or buildings. Scanning the received signal strength (RSS) values from multiple APs in users' phones or laptops [21, 25, 30] have been used to locate users in a specific zone of a building, and consequently, to find number of occupants. WiFi based approaches have also been utilized within city buses in [23] to detect occupancy and to passively track patrons by using broadcasted MAC addresses of user devices.

Other deployed sensors: Deploying specific sensors for detecting presence of a user can provide high-accuracy and room-level occupancy estimates. But installing these additional hardware is potentially expensive and time consuming. One commonly used approach involves use of PassiveInfrared (PIR) sensors inside rooms [5, 14] to detect the presence of an individual. Since detecting idle occupants in this method can be difficult, [45] proposed deploying of such sensors in the walkways or entry/exit points. Use of acoustic sensors [9, 43, 44] can aid in occupant detection as the human body reflects sound waves in ways different to other furniture in a room. A number of research works have used ultrasonic signals and observed the multipath characteristics or changes in received signal strength as compared to the non occupied state to detect occupancy. Image processing techniques applied on images taken by deployed cameras [47, 50], or monitoring the thermal hotspots in camera based thermal sensors [8, 42] can also help in counting occupants. Deploying ambient motion sensors [29, 32, 39] or pressure sensors [38, 46] have also been used to detect occuapancy. [52] has combined CO_2 sensors and ambient light sensors with the afore-mentioned sensors to detect the occupancy.

Many of the sensor based approaches (thermal imaging, motion sensors) are efficient in detecting low occupant counts and fail in estimating higher values of occupancy. In addition to cost incurred from setting up sensors, another challenge in deploying sensors in specific rooms is the preservation of user privacy. Usage of camera based techniques have serious flaws related to user privacy. This approach of occupancy detection has mostly been used to count occupants within specific zones or areas of buildings and not estimate the entire occupancy as a whole.

Other Related Work: Many of the occupancy detection techniques are agnostic to varying contextual information, specific of the space where the occupancy is being detected. The behavior of the sensors (intrusive or non-intrusive) can vary significantly depending on the building context. Use of smart meter data to decipher contextual household characteristics, like age-range, employment-status, floor area, number of occupancts, number of bedrooms, etc. has been studied in [7]. Other research works [40, 41] have discussed how varying context, like presence of holidays, or low-income levels, or extent of urbanization effect the overall energy demand. The electric demand change for these works are coarse-grained, whereas, in this work, the focus in on more fine-grained variation related to context. Location context and its effect on usage of WiFi network and devices have been discussed in [48] and [16]. Contextual information has also been extensively used for efficient recommendation systems [4], better sentiment analysis and opinion mining [10] and mobility prediction [28]. Another important related work is about users having multiple WiFi enabled devices. Recent works [1, 12] have shown that more than 50% users possess multiple devices and that different locations and device types govern the usage of the WiFi network in various ways. Our proposed work is in line with the research on multi-device users and exploits their existence for occupancy estimation.

8 CONCLUSION

In this paper, we investigated the feasibility of estimating occupancy using three readily available, non-intrusive building data sources - WiFi device count, electrical energy demand and water consumption rate. We find that although the utilities (electricity and water) data by themselves can only provide coarse-grained building occupancy estimation, they can supplement the WiFi data in achieving fine-grained, accurate occupancy estimation. We use WiFi session logs with username information to estimate our ground truth - which can be limited to a certain extent as it does not count the percentage of users with wired devices or users without devices. Through occupancy-agnostic and occupancy-aware clustering, we propose methods to reduce the per-building training overhead by developing occupancy estimation models for building clusters. We evaluate our clustering and multi-modal fusion techniques using a large university dataset collected over 28 days from 76 buildings. Our evaluation shows that clustering based on occupancy-data source relationship with decision-level fusion achieves an accuracy of 13.22%. We use the occupancy-agnostic clustering scheme to facilitate the occupancy estimation of new untrained buildings where occupancy ground truth is unavailable to perform the initial clustering. We also observe that addition of water consumption data does not provide significant improvement over WiFi and smart meter based estimation. In our approach, we use linear regression for building our estimation models. Change in the estimation performance by using non-linear approaches are left as a scope for future work.

REFERENCES

[1] 2015. More than 40% of online adults are multi-device users. (2015). https://econsultancy.com/blog/64464-more-than-40-of-online-adults-are-multi-device-users-stats/

[2] 2016. WeatherSource OnPoint API. (2016). https://developer.weathersource.com/

[3] W. H. Adams, Giridharan Iyengar, Ching-Yung Lin, Milind Ramesh Naphade, Chalapathy Neti, Harriet J. Nock, and John R. Smith. 2003. Semantic Indexing of Multimedia Content Using Visual, Audio, and Text Cues. *EURASIP Journal on Advances in Signal Processing* 2003, 2 (2003), 987184. DOI:http://dx.doi.org/10.1155/S1110865703211173

[4] Gediminas Adomavicius and Alexander Tuzhilin. 2015. *Context-Aware Recommender Systems.* Springer US, Boston, MA, 191–226. DOI:http://dx.doi.org/10.1007/978-1-4899-7637-6_6

[5] Yuvraj Agarwal, Bharathan Balaji, Rajesh Gupta, Jacob Lyles, Michael Wei, and Thomas Weng. 2010. Occupancy-driven Energy Management for Smart Building Automation. In *Proceedings of the 2Nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building (BuildSys '10).* ACM, New York, NY, USA, 1–6. DOI:http://dx.doi.org/10.1145/1878431.1878433

[6] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. 2010. Multimodal Fusion for Multimedia Analysis: A Survey. *Multimedia Syst.* 16, 6 (Nov. 2010), 345–379. DOI:http://dx.doi.org/10.1007/s00530-010-0182-0

[7] Christian Beckel, Leyna Sadamori, Thorsten Staake, and Silvia Santini. 2014. Revealing household characteristics from smart meter data. *Energy* 78 (2014), 397–410.

[8] Alex Beltran, Varick L. Erickson, and Alberto E. Cerpa. 2013. ThermoSense: Occupancy Thermal Based Sensing for HVAC Control. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings (BuildSys'13).* ACM, New York, NY, USA, Article 11, 8 pages. DOI:http://dx.doi.org/10.1145/2528282.2528301

[9] D. Caicedo and A. Pandharipande. 2012. Ultrasonic array sensor for indoor presence detection. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO).* 175–179.

[10] E. Cambria, B. Schuller, Y. Xia, and C. Havasi. 2013. New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems* 28, 2 (March 2013), 15–21. DOI:http://dx.doi.org/10.1109/MIS.2013.30

[11] Dong Chen, Sean Barker, Adarsh Subbaswamy, David Irwin, and Prashant Shenoy. 2013. Non-Intrusive Occupancy Monitoring Using Smart Meters. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings (BuildSys'13).* ACM, New York, NY, USA, Article 9, 8 pages. DOI:http://dx.doi.org/10.1145/2528282.2528294

[12] A. K. Das, P. H. Pathak, C. N. Chuah, and P. Mohapatra. 2015. Characterization of wireless multi-device users. In *2015 12th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON).* 327–335. DOI:http://dx.doi.org/10.1109/SAHCN.2015.7338332

[13] S. Depatla, A. Muralidharan, and Y. Mostofi. 2015. Occupancy Estimation Using Only WiFi Power Measurements. *IEEE Journal on Selected Areas in Communications* 33 (2015), 1381–1393. DOI:http://dx.doi.org/10.1109/JSAC.2015.2430272

[14] Robert H. Dodier, Gregor P. Henze, Dale K. Tiller, and Xin Guo. 2006. Building occupancy detection through sensor belief networks. *Energy and Buildings* 38, 9 (2006), 1033 – 1043. DOI:http://dx.doi.org/10.1016/j.enbuild.2005.12.001

[15] Bing Dong, Zhaoxuan Li, and Gaelen Mcfadden. 2015. An investigation on energy-related occupancy behavior for low-income residential buildings. *Science and Technology for the Built Environment* 21, 6 (2015), 892–901. DOI:http://dx.doi.org/10.1080/23744731.2015.1040321 arXiv:http://dx.doi.org/10.1080/23744731.2015.1040321

[16] Hossein Falaki, Ratul Mahajan, Srikanth Kandula, Dimitrios Lymberopoulos, Ramesh Govindan, and Deborah Estrin. 2010. Diversity in Smartphone Usage. In *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services (MobiSys '10).* ACM, New York, NY, USA, 179–194. DOI:http://dx.doi.org/10.1145/1814433.1814453

[17] Jinliang Fan, Jun Xu, Mostafa H. Ammar, and Sue B. Moon. 2004. Prefix-preserving IP Address Anonymization: Measurement-based Security Evaluation and a New Cryptography-based Scheme. *Comput. Netw.* 46, 2 (Oct. 2004), 253–272. DOI:http://dx.doi.org/10.1016/j.comnet.2004.03.033

[18] Yuki Fukuzaki, Masahiro Mochizuki, Kazuya Murao, and Nobuhiko Nishio. 2014. A Pedestrian Flow Analysis System Using Wi-Fi Packet Sensors to a Real Environment. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct)*. ACM, New York, NY, USA, 721–730. DOI: http://dx.doi.org/10.1145/2638728.2641312

[19] Yuuki Fukuzaki, Masahiro Mochizuki, Kazuya Murao, and Nobuhiko Nishio. 2015. Statistical Analysis of Actual Number of Pedestrians for Wi-Fi Packet-based Pedestrian Flow Sensing. In *2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers (UbiComp/ISWC'15 Adjunct)*. ACM, New York, NY, USA, 1519–1526. DOI: http://dx.doi.org/10.1145/2800835.2801623

[20] P. Gehler and S. Nowozin. 2009. On feature combination for multiclass object classification. In *2009 IEEE 12th International Conference on Computer Vision*. 221–228. DOI: http://dx.doi.org/10.1109/ICCV.2009.5459169

[21] S. K. Ghai, L. V. Thanayankizil, D. P. Seetharam, and D. Chakraborty. 2012. Occupancy detection in commercial buildings using opportunistic context sources. In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. 463–466. DOI: http://dx.doi.org/10.1109/PerComW.2012.6197536

[22] Ming Jin, Ruoxi Jia, Zhaoyi Kang, Ioannis C. Konstantakopoulos, and Costas J. Spanos. 2014. PresenceSense: Zero-training Algorithm for Individual Presence Detection Based on Power Monitoring. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings (BuildSys '14)*. ACM, New York, NY, USA, 1–10. DOI: http://dx.doi.org/10.1145/2674061.2674073

[23] Lei Kang, Bozhao Qi, and Suman Banerjee. 2016. A Wireless-Based Approach for Transit Analytics. In *Proceedings of the 17th International Workshop on Mobile Computing Systems and Applications (HotMobile '16)*. ACM, New York, NY, USA, 75–80. DOI: http://dx.doi.org/10.1145/2873587.2873589

[24] Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. 1998. On combining classifiers. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* 20 (1998), 226–239.

[25] Wilhelm Kleiminger, Christian Beckel, Anind Dey, and Silvia Santini. 2013. Using Unlabeled Wi-Fi Scan Data to Discover Occupancy Patterns of Private Households. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems (SenSys '13)*. ACM, New York, NY, USA, Article 47, 2 pages. DOI: http://dx.doi.org/10.1145/2517351.2517421

[26] Wilhelm Kleiminger, Christian Beckel, and Silvia Santini. 2015. Household Occupancy Monitoring Using Electricity Meters. In *2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 975–986. DOI: http://dx.doi.org/10.1145/2750858.2807538

[27] Wilhelm Kleiminger, Christian Beckel, Thorsten Staake, and Silvia Santini. 2013. Occupancy Detection from Electricity Consumption Data. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings (BuildSys'13)*. ACM, New York, NY, USA, Article 10, 8 pages. DOI: http://dx.doi.org/10.1145/2528282.2528295

[28] Julian F.P. Kooij, Nicolas Schneider, Fabian Flohr, and Dariu M. Gavrila. 2014. *Context-Based Pedestrian Path Prediction*. Springer International Publishing, Cham, 618–633. DOI: http://dx.doi.org/10.1007/978-3-319-10599-4_40

[29] Timilehin Labeodan, Wim Zeiler, Gert Boxem, and Yang Zhao. 2015. Occupancy measurement in commercial office buildings for demand-driven control applications - A survey and detection system evaluation. *Energy and Buildings* 93 (2015), 303 – 314. DOI: http://dx.doi.org/10.1016/j.enbuild.2015.02.028

[30] Du Li, Bharathan Balaji, Yifei Jiang, and Kshitiz Singh. 2012. A Wi-fi Based Occupancy Sensing Approach to Smart Energy in Commercial Office Buildings. In *Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys '12)*. ACM, New York, NY, USA, 197–198. DOI: http://dx.doi.org/10.1145/2422531.2422568

[31] Nan Li, Gulben Calis, and Burcin Becerik-Gerber. 2012. Measuring and monitoring occupancy with an {RFID} based system for demand-driven {HVAC} operations. *Automation in Construction* 24 (2012), 89 – 99. DOI: http://dx.doi.org/10.1016/j.autcon.2012.02.013

[32] Salvatore Longo and Bin Cheng. 2015. Privacy Preserving Crowd Estimation for Safer Cities. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers (UbiComp/ISWC'15 Adjunct)*. ACM, New York, NY, USA, 1543–1550. DOI: http://dx.doi.org/10.1145/2800835.2801631

[33] Claudio Martani, David Lee, Prudence Robinson, Rex Britter, and Carlo Ratti. 2012. ENERNET: Studying the dynamic relationship between building occupancy and energy consumption. *Energy and Buildings* 47 (2012), 584–591.

[34] S. Meyn, A. Surana, Y. Lin, S. M. Oggianu, S. Narayanan, and T. A. Frewen. 2009. A sensor-utility-network method for estimation of occupancy in buildings. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*. 1494–1500. DOI: http://dx.doi.org/10.1109/CDC.2009.5400442

[35] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad, and P. Natarajan. 2012. Multimodal feature fusion for robust event detection in web videos. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 1298–1305. DOI: http://dx.doi.org/10.1109/CVPR.2012.6247814

[36] Ara V. Nefian, Luhong Liang, Xiaobo Pi, Xiaoxing Liu, and Kevin Murphy. 2002. Dynamic Bayesian Networks for Audio-Visual Speech Recognition. *EURASIP Journal on Advances in Signal Processing* 2002, 11 (2002), 783042. DOI: http://dx.doi.org/10.1155/S1110865702206083

[37] Sameera Palipana, Piyush Agrawal, and Dirk Pesch. 2016. Channel State Information Based Human Presence Detection Using Non-linear Techniques. In *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments (BuildSys '16)*. ACM, New York, NY, USA, 177–186. DOI: http://dx.doi.org/10.1145/2993422.2993579

[38] Shijia Pan, Mostafa Mirshekari, Pei Zhang, and Hae Young Noh. 2016. Occupant traffic estimation through structural vibration sensing. (2016). DOI: http://dx.doi.org/10.1117/12.2222024

[39] Juhi Ranjan, Erin Griffiths, and Kamin Whitehouse. 2014. Discerning Electrical and Water Usage by Individuals in Homes. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings (BuildSys '14)*. ACM, New York, NY, USA, 20–29. DOI: http://dx.doi.org/10.1145/2674061.2674066

[40] Patricia Romero-Lankao, Kevin R. Gurney, Karen C. Seto, Mikhail Chester, Riley M. Duren, Sara Hughes, Lucy R. Hutyra, Peter Marcotullio, Lawrence Baker, Nancy B. Grimm, Christopher Kennedy, Elisabeth Larson, Stephanie Pincetl, Dan Runfola, Landy Sanchez, Gyami Shrestha, Johannes Feddema, Andrea Sarzynski, Joshua Sperling, and Eleanor Stokes. 2014. A critical knowledge pathway to low-carbon, sustainable futures: Integrated understanding of urbanization, urban areas, and carbon. *Earth's Future* 2, 10 (2014), 515–532. DOI: http://dx.doi.org/10.1002/2014EF000258 2014EF000258.

[41] Miguel O. Romn and Eleanor C. Stokes. 2015. Holidays in lights: Tracking cultural patterns in demand for energy services. *Earth's Future* 3, 6 (2015), 182–205. DOI: http://dx.doi.org/10.1002/2014EF000285 2014EF000285.

[42] Fisayo Caleb Sangoboye and Mikkel Baun Kjærgaard. 2016. PLCount: A Probabilistic Fusion Algorithm for Accurately Estimating Occupancy from 3D Camera Counts. In *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments (BuildSys '16)*. ACM, New York, NY, USA, 147–156. DOI: http://dx.doi.org/10.1145/2993422.2993575

[43] Oliver Shih, Patrick Lazik, and Anthony Rowe. 2016. AURES: A Wide-Band Ultrasonic Occupancy Sensing Platform. In *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments (BuildSys '16)*. ACM, New York, NY, USA, 157–166. DOI: http://dx.doi.org/10.1145/2993422.2993580

[44] Oliver Shih and Anthony Rowe. 2015. Occupancy Estimation Using Ultrasonic Chirps. In *Proceedings of the ACM/IEEE Sixth International Conference on Cyber-Physical Systems (ICCPS '15)*. ACM, New York, NY, USA, 149–158. DOI: http://dx.doi.org/10.1145/2735960.2735969

[45] Elahe Soltanaghaei and Kamin Whitehouse. 2016. WalkSense: Classifying Home Occupancy States Using Walkway Sensing. In *3rd ACM International Conference on Systems for Energy-Efficient Built Environments (BuildSys '16)*. ACM, New York, NY, USA, 167–176. DOI: http://dx.doi.org/10.1145/2993422.2993576

[46] Arun Subramanian, Kishan G. Mehrotra, Chilukuri K. Mohan, Pramod K. Varshney, and Thyagaraju Damarla. 2010. *Feature Selection and Occupancy Classification Using Seismic Sensors*. Springer Berlin Heidelberg, Berlin, Heidelberg, 605–614. DOI: http://dx.doi.org/10.1007/978-3-642-13025-0_62

[47] T. Teixeira and A. Savvides. 2008. Lightweight People Counting and Localizing for Easily Deployable Indoors WSNs. *IEEE Journal of Selected Topics in Signal Processing* 2, 4 (Aug 2008), 493–502. DOI: http://dx.doi.org/10.1109/JSTSP.2008.2001426

[48] Ionut Trestian, Supranamaya Ranjan, Aleksandar Kuzmanovic, and Antonio Nucci. 2009. Measuring Serendipity: Connecting People, Locations and Interests in a Mobile 3G Network. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement (IMC '09)*. ACM, New York, NY, USA, 267–279. DOI: http://dx.doi.org/10.1145/1644893.1644926

[49] Yi Wu, Edward Y. Chang, Kevin Chen-Chuan Chang, and John R. Smith. 2004. Optimal Multimodal Fusion for Multimedia Data Analysis. In *12th Annual ACM International Conference on Multimedia (MULTIMEDIA '04)*. ACM, New York, NY, USA, 572–579. DOI: http://dx.doi.org/10.1145/1027527.1027665

[50] D. B. Yang, H. H. Gonzalez-Banos, and L. J. Guibas. 2003. Counting people in crowds with a real-time network of simple image sensors. In *Proceedings Ninth IEEE International Conference on Computer Vision*. 122–129 vol.1. DOI: http://dx.doi.org/10.1109/ICCV.2003.1238325

[51] Longqi Yang, K. Ting, and M. B. Srivastava. 2014. Inferring occupancy from opportunistically available sensor data. In *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 60–68. DOI: http://dx.doi.org/10.1109/PerCom.2014.6813945

[52] Zheng Yang and Burcin Becerik-Gerber. 2015. Cross-Space Building Occupancy Modeling by Contextual Information Based Learning. In *Proceedings of the 2Nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments (BuildSys '15)*. ACM, New York, NY, USA, 177–186. DOI: http://dx.doi.org/10.1145/2821650.2821668

[53] G. Ye, D. Liu, I. H. Jhuo, and S. F. Chang. 2012. Robust late fusion with rank minimization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 3021–3028. DOI: http://dx.doi.org/10.1109/CVPR.2012.6248032