Energy and Buildings 133 (2016) 790–798

Contents lists available at ScienceDirect

Energy and Buildings

journal homepage: www.elsevier.com/locate/enbuild

A fusion framework for occupancy estimation in office buildings based on environmental sensor data

Zhenghua Chen, Mustafa K. Masood, Yeng Chai Soh*

School of Electrical and Electronics Engineering, Nanyang Technological University, 50 Nanyang Ave, Singapore

ARTICLE INFO

Article history:
Received 11 March 2016
Received in revised form
13 September 2016
Accepted 19 October 2016
Available online 20 October 2016

Keywords:
Occupancy estimation
Data-driven models
Occupancy models
ELM-based wrapper

ABSTRACT

Occupancy information that can be used to determine heating, ventilation and lighting requirements is one of the important parameters for the control of energy efficient buildings. In this paper, we propose a fusion framework for building occupancy estimation with environmental parameters. Based on the environmental sensor data, a coarse estimation of building occupancy can be achieved using data-driven models that include extreme learning machine (ELM), support vector machine (SVM), artificial neural network (ANN), K-nearest neighbors (KNN), linear discriminant analysis (LDA) and classification and regression tree (CART). Due to the extremely fast learning speed of the ELM algorithm, we apply an ELM-based wrapper method to select the best feature set of environmental parameters. To further improve the estimation accuracy of building occupancy, taking occupancy dynamics into consideration, we fuse the results of data-driven models with well developed occupancy models by using a particle filter algorithm. Real experiments have shown that our proposed fusion framework can achieve improvements of around 5–14% and 3–12% for the estimation accuracy and the detection accuracy (presence/absence) respectively among the different methodologies.

1. Introduction

According to the United Nations Environment Programme (UNEP), buildings contribute to about one-third of total global greenhouse emissions [1]. In Singapore alone, the greenhouse gas emissions from buildings are expected to be around 5 million tons, or 13.8% of the city's total [2]. There is a need for more energy-efficient technologies for buildings. In this regard, occupancy information can be a valuable asset. Occupancy-driven control can save up to 15% of the total energy usage in buildings [3]. Another application of occupancy information is adaptive lighting control which has been shown to be effective and can reduce 35–75% of energy usage for lighting in buildings [4]. Moreover, building security and emergency evacuation also require an accurate monitoring of building occupancy. But obtaining reliable occupancy estimates is a challenging problem that requires more attention.

Occupancy estimation can be accomplished with different sensors. Some works focus on binary detection, i.e. presence and absence, usually with PIR sensors [5]. But recently, greater focus is being given to the estimation of actual occupancy numbers. Methods with various sensors, such as RFID tags [6], cameras [7] and wearable sensors [8], have been presented. However, these approaches are intrusive to occupants. Other works take an activity-dependent approach, such as the use of chair sensors [9] and monitoring appliance power consumption [10]. These approaches cannot detect occupants who are disengaged from the activities being monitored.

A recent development has been the use of environmental sensors for occupancy estimation, which is non-intrusive for users and of low-cost. The environmental sensors include CO_2, temperature, humidity and air pressure which are influenced by the change of occupancy. The inherent relationship between occupancy and environmental parameters can be utilized for occupancy estimation [11–13]. Specifically, from environmental sensor data, useful features that capture changes in occupancy can be extracted. Out of these extracted features, the most relevant ones can be used with data-driven models to generate occupancy estimates.

Various feature selection techniques can be applied to find the most relevant features. In the literature on occupancy estimation, feature selection using correlation measures like Information Gain and Symmetric Uncertainty has been reported [14,12,15]. These are filter-based approaches, which select features independently of classifiers. While these approaches are fast, they compromise on the estimation accuracy. We instead use a wrapper-based

* Corresponding author.
E-mail addresses: chen0832@e.ntu.edu.sg (Z. Chen), must0006@e.ntu.edu.sg (M.K. Masood), EYCSOH@ntu.edu.sg (Y.C. Soh).

http://dx.doi.org/10.1016/j.enbuild.2016.10.030
0378-7788/

approach, which is generally more accurate than filter methods [16]. However, it has a problem of high computational load with normal classifiers. Owing to the extremely fast learning speed of the extreme learning machine (ELM) algorithm, we perform an ELM-based wrapper method which is feasible for real implementation and has higher accuracy. Building occupancy has been shown to contain specific patterns [17–19]. However, traditional data-driven approaches did not consider the information of occupancy patterns. By taking occupancy patterns into consideration, we combine a well developed occupancy model with the outputs of data-driven models via a particle filter algorithm to further improve the occupancy estimation accuracy. The data-driven models that are employed in this work include ELM, support vector machine (SVM), artificial neural network (ANN), K-nearest neighbors (KNN), linear discriminant analysis (LDA) and classification and regression tree (CART). Experiments have been conducted in a multi-occupant office environment to evaluate the performance of the proposed framework.

The main contributions of this paper are as follows:

- We propose a fusion framework which combines data-driven models with occupancy models for building occupancy estimation based on environmental sensor data.
- We apply a new approach of ELM with regularization term for occupancy estimation, and perform an ELM-based wrapper method for feather selection.
- We evaluate the proposed approach using real experiment data, and compare it with the state-of-the-art data-driven approaches.

2. Related works

2.1. Occupancy models

Occupancy models which reveal the patterns of occupancy have been well developed for decades. Wang et al. proposed two exponential distributions to model occupied and vacant intervals in a single person office [20]. Page et al. presented an inhomogeneous Markov chain with two states of presence and absence to model occupancy states in single person offices [17]. A parameter of mobility which denotes the changing rate of the two states is defined to calculate the transition probability matrix of the inhomogeneous Markov chain model.

An extension of the model in [17] into a more complicated multi-occupant situation can be found in [21]. Richardson et al. presented an inhomogeneous Markov chain where the state is the number of occupants to simulate occupants in residential buildings [21]. Another work in [22] proposed an event-driven approach for modeling building occupancy. An inhomogeneous Markov chain is applied for each event which contains walking around, going to office, getting off work and lunch break. Erickson et al. applied a multivariate Gaussian and an agent-based models to capture occupancy patterns in buildings [23]. The multivariate Gaussian model attempts to fit a Gaussian distribution for building occupancy at each time step. And the agent-based model tries to model each occupant's behaviour individually.

Liao et al. proposed a different agent-based model with four modules to simulate building occupancy [18]. Based on the occupancy property, a damping process which claims that occupants tend to stay at their working place for a long time and an acceleration process which claims that occupants tend to leave hallways or restrooms quickly were established. A more recent work can be found in [19], where the authors proposed two novel inhomogeneous Markov chain models for regular occupancy modeling in commercial buildings. Instead of using the number of occupants as the states, they leverage on the increment information of occupancy. In this way, they can dramatically simplify the calculation of Markov transition probability matrices.

2.2. Occupancy estimation and detection

A feature selection approach to occupancy estimation was introduced by Dong et al. [11]. Their method involved four main steps: (1) collect environmental sensor data, (2) extract features from the data, (3) select the most relevant features, and (4) input the selected features to machine learning algorithms. They recorded various environmental parameters in an open plan office using a wireless sensor network. Feature selection was performed using a filter method based on information gain. They concluded that the most relevant parameters were CO_2 and acoustic levels. The estimation accuracy with over a week of data was limited to 70%.

Yang et al. [24] used a combination of non-intrusive sensors that measured indoor temperature, CO_2, humidity, light, sound and motion. They collected data in two multi-occupancy rooms and trained a Radial Basis Function (RBF) neural network with a number of features to estimate the occupancy. They achieved an accuracy of up to 88.74% with an error tolerance of one occupant. However, they were not selective in their use of features, and the entire dataset is used in the estimation. Yang et al. addressed this problem in [12], in which they explored different combinations of sensors, i.e. motion, sound, door, temperature, humidity, CO_2, light, and passive infrared, to detect occupancy in both single-occupancy and multi-occupancy offices by using six data-driven models. The contributions of each sensor were evaluated based on information gain theory. They found the CO_2, door status and light levels to be the three most informative variables. Among the six methodologies, the decision tree algorithm performs the best with detection accuracies of 96.0–98.2% and the RMSEs (Root Mean Square Errors) of 0.109–0.156 for the four different rooms.

Khan et al. [25] measured temperature, light, humidity, PIR and audio levels in an office space. For classification, they proposed the use of three hierarchies with varying resolutions of occupancy and employed the posterior probabilities of lower levels to improve the feature vectors of higher levels. This improved the estimation accuracy of higher levels. The classifiers that they used are KNN and SVM. They also incorporated meeting schedule and computer usage information into the features. However, they found that the meeting schedules could be inaccurate, such as when a scheduled meeting would not take place. This limited the positive contribution of the scheduled information to the overall estimation accuracy.

Recently, an occupancy detection system that leverages on CO_2, temperature, humidity and light levels was presented [26]. The authors applied statistical models for classification. They found that with combinations of just two features, good detection accuracy could be achieved, especially with linear discriminant analysis (LDA). However, the Random Forest model yielded poor accuracies. They also included the time of day and week status in their model, which improved the accuracy. This work was limited to detection of presence or absence of occupants, rather than the actual occupancy levels.

The state-of-the-art approaches reviewed above employed environmental sensors and occupancy related sensors, e.g. PIR, motion sensor and light sensor, for occupancy estimation and detection using various data-driven approaches. In this work, we only employ environmental sensors, i.e. temperature, humidity, CO_2 and air pressure. No specific occupancy sensors, e.g. PIR, motion or light sensors, are required in the environment. We noted that feature selection has been shown to be effective in increasing estimation accuracy [11,12,26]. We perform an ELM-based wrapper method for feature selection. The data-driven models we used include ELM, SVM, ANN, KNN, LDA and CART. To further enhance the performance of the estimation system, we combine them with

existing well developed occupancy models which can extract occupancy patterns for building occupancy estimation.

3. Methodology

In this section, we will first introduce the data-driven models of ELM, SVM, ANN, KNN, LDA and CART. Then, feature engineering which consists of feature extraction and feature selection will be presented. After that, we will select a proper occupancy model for our case. Finally, fusion of the occupancy model and the output of data-driven models by using a particle filter will be presented.

3.1. Data-driven models

The data-driven models of ELM, SVM, ANN, KNN, LDA and CART will be introduced in this section.

3.1.1. Extreme learning machine

Extreme learning machine (ELM) was developed for single-hidden layer feedforward neural networks (SLFNs). It randomly chooses the parameters of hidden layer neurons and analytically determines the weights of output neurons [27]. Given training samples $(\mathbf{x}_k,\ \mathbf{y}_k)$, where $\mathbf{x}_k = [x_{k1}, x_{k2}, \ldots, x_{km}]^T \in \mathbb{R}^m$ and $\mathbf{y}_k = [y_{k1}, y_{k2}, \ldots, y_{kn}]^T \in \mathbb{R}^n$, the activation function is $g(\cdot)$ and the number of hidden nodes is L. Then, the output of SLFNs $\mathbf{t}_k$ is given by

$$\sum_{j=1}^{L} \mathbf{w}_j g(\boldsymbol{\alpha}_j, \mathbf{x}_k, \beta_j) = \mathbf{t}_k \tag{1}$$

where $k = 1, \ldots, N$, N is the total number of samples, $\boldsymbol{\alpha}_j$ is the weight of input nodes to the hidden node j, β_j is the bias of hidden node j, and $\mathbf{w}_j \in \mathbb{R}^n$ is the weight of the hidden node j to output nodes. Assume that the activation function $g(\cdot)$ can approximate these N samples with no error, which means $\sum_{k=1}^{N} \|\mathbf{t}_k - \mathbf{y}_k\| = 0$, i.e., there exist $\mathbf{w}_j$, $\boldsymbol{\alpha}_j$ and β_j such that

$$\sum_{j=1}^{L} \mathbf{w}_j g(\boldsymbol{\alpha}_j, \mathbf{x}_k, \beta_j) = \mathbf{y}_k \tag{2}$$

We can rewrite the above equation into matrix form as

$$\mathbf{H}\mathbf{w} = \mathbf{Y} \tag{3}$$

where

$$\mathbf{H} = \begin{bmatrix} g(\boldsymbol{\alpha}_1, \mathbf{x}_1, \beta_1) & \cdots & g(\boldsymbol{\alpha}_L, \mathbf{x}_1, \beta_L) \\ \vdots & \cdots & \vdots \\ g(\boldsymbol{\alpha}_1, \mathbf{x}_N, \beta_1) & \cdots & g(\boldsymbol{\alpha}_L, \mathbf{x}_N, \beta_L) \end{bmatrix}_{N\times L}, \tag{4}$$

$$\mathbf{w} = \begin{bmatrix} \mathrm{w}_1^T \\ \vdots \\ \mathrm{w}_L^T \end{bmatrix}_{\mathrm{L}\times\mathrm{n}} \quad \textbf{and} \quad \mathbf{Y} = \begin{bmatrix} \mathrm{y}_1^T \\ \vdots \\ \mathrm{y}_N^T \end{bmatrix}_{\mathrm{N}\times\mathrm{n}} \tag{5}$$

As named by Huang et al. [27], $\mathbf{H}$ is called hidden layer output matrix; the kth column of $\mathbf{H}$ is the kth hidden node output with respect to all the inputs. According to [27], the smallest least-squares solution of Eq. (3) can be expressed as

$$\hat{\mathbf{w}} = \mathbf{H}^{\dagger}\mathbf{Y} \tag{6}$$

where $\mathbf{H}^{\dagger}$ is the *Moore Penrose generalized inverse* of matrix $\mathbf{H}$. Based on [28], an orthogonal projection method can be employed to calculate $\mathbf{H}^{\dagger}$ in two cases: when $\mathbf{H}^T\mathbf{H}$ is nonsingular, $\mathbf{H}^{\dagger} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$, or when $\mathbf{H}\mathbf{H}^T$ is nonsingular, $\mathbf{H}^{\dagger} = \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}$. To avoid the singularity of $\mathbf{H}^T\mathbf{H}$ or $\mathbf{H}\mathbf{H}^T$, the authors in [28] suggest to add a positive value to the diagonal of $\mathbf{H}^T\mathbf{H}$ or $\mathbf{H}\mathbf{H}^T$, which can be expressed as

$$\hat{\mathbf{w}} = \left(\frac{\mathbf{C}}{\mathbf{I}} + \mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{Y} \quad \textbf{or} \tag{7}$$

$$\hat{\mathbf{w}} = \mathbf{H}^T\left(\frac{\mathbf{C}}{\mathbf{I}} + \mathbf{H}\mathbf{H}^T\right)^{-1}\mathbf{Y} \tag{8}$$

where $\mathbf{I}$ is the identity matrix, and C is the regularization term. In this way, the solution will be more stable and tends to have better generalization performance [28].

In summary, the implementation of ELM can be divided into three steps:

(1) Randomly assign hidden nodes parameters, i.e. $\boldsymbol{\alpha}_j$ and β_j where $j = 1, 2, \ldots, L$.
(2) Calculate the hidden layer output matrix $\mathbf{H}$.
(3) Calculate the output weight $\mathbf{w}$.

3.1.2. Support vector machine

Support vector machine (SVM) was developed under the principles of structural risk minimization which not only minimizes the training error, but also reduces the system complexity [29]. By applying the kernel trick, it can map the inputs into a high dimensional feature space, where different classes can be separated easily. The widely used radial basis function (RBF) kernel [25] is employed in this paper.

3.1.3. Artificial neural network

Another widely used data-driven approach is artificial neural network (ANN). The neurons in ANN are computational models inspired by natural neurons. The ANN model applies these neurons to process information. In this paper, a single hidden layer feed-forward neural network is used for building occupancy estimation. The Sigmoid functions are used as the activation functions for all hidden neurons. A back-propagation method is employed for the learning of the parameters. The details of ANN can be found in [30].

3.1.4. K-nearest neighbors

K-nearest neighbors (KNN) algorithm is quite popular for classification because of its simple structure and easy interpretation. Given a training data set $(\mathbf{x}_k, y_k)$, where $k = 1, 2, \ldots, N$, and a testing sample $\mathbf{t}$, the distance, d_k, between $\mathbf{t}$ and each $\mathbf{x}_k$ is calculated as

$$d_k = \|\mathbf{t} - \mathbf{x}_k\|_{dis} \tag{9}$$

where $\|\cdot\|_{dis}$ is the distance calculation algorithm. One of the most widely used distance calculation algorithms is Euclidean distance. After obtaining the distances d_k, $k = 1, 2, \ldots, N$, to each training sample, the labels of K training samples with the smallest distance will be selected. Then, a majority voting will be performed to determine the label of the testing sample. Note that any ties can be handled by using a random selection.

3.1.5. Linear discriminant analysis

Linear discriminant analysis (LDA), also known as Fisher linear discriminant [31], is a popular data-driven approach in pattern recognition area. It aims at finding a projection hyperplane for features that minimizes the variance within each class and maximizes the means of classes after projection, which means the best separability in the new feature space. This objective can be resolved by using eigenvalue analysis. The corresponding eigenvector can be leveraged to determine the projection hyperplane. The detailed mathematical expressions can be found in [32].

Table 1
List of features extracted from sensor data.

Feature	Description
First order difference	$fd(i) = raw(i) - raw(i-1)$
Second order difference	$fd2(i) = fd(i) - fd(i-1)$
First order shifted difference	$fds(i) = raw(i) - raw(i-2)$
5-min moving average	$mavg = (\sum_{j=i-4}^{i} raw(j))/5$
10-min shifted difference	$fds10(i) = raw(i) - raw(i-10)$

3.1.6. Classification and regression tree

classification and regression tree (CART) is established upon the principle of recursively partitioning feature space and fitting a simple threshold model within each partition. It can be represented graphically as a decision tree. The partition in CART is a binary split which relies on the criterion of Gini index. The CART splits a node by exhaustively searching all the possible routes that minimize the total Gini index of its two child nodes. The detailed description of CART can be found in [33].

3.2. Feature engineering

The raw environmental sensor data may be noisy and is not likely to give accurate classification results if used directly. Thus, we need to extract more representative information, known as features that can indicate occupancy information. Among all the extracted features, some of them may be redundant, which will lead to poorer performance and incur unnecessary computational time. Therefore, feature selection is necessary. In this work, feature engineering contains two parts, i.e. feature extraction and feature selection.

3.2.1. Feature extraction

The features extracted from the raw environmental data are shown in Table 1. The first order difference (*fd*), second order difference (*fd*2) and first order shifted difference (*fds*) are selected to capture the temporal variations in the data. The moving average (*mavg*) and 10-min shifted difference (*fds*10) features take into account the time delay in the build-up and decay of the environmental parameters. The choice of 5 min in the moving average is based on its effectiveness in [15]. Note that, if the data is not complete, we treat the unavailable data the same as the nearest one for difference related features and perform the average to the available data for moving average related features. For instance, if only one sample is available for the first order difference, since we treat the unavailable one the same as the nearest one, the feature of that is zero. If only two samples (one sample per min) are available for 5-min moving average, we perform the average of these two as the result of 5 min moving average.

3.2.2. Feature selection using an ELM-based wrapper method

In general, feature selection can be accomplished by leveraging on two approaches, i.e. the filter method and the wrapper method. In the filter method, the merit of the features can be assessed using criteria, e.g. Information Gain [14,12] and Symmetric Uncertainty [15], that are independent of classifiers. This makes the feature selection quite fast, but compromises the estimation accuracy. In the wrapper method, classifiers are employed to assess the merit of features. The feature selection is thus optimized for classifiers. This generally yields better performance when compared to the filter method. One issue of the wrapper method is that the computational burden is quite high, because multiple classifier models are created in the feature selection process. Most previous works that estimated occupancy using environmental parameters used filter methods. We instead apply a wrapper method [34]. To address the issue of computational time, we perform an ELM-based wrapper method which is quite fast owing to the extremely fast learning speed of the ELM algorithm [13].

1 **for** $N = 2$ *to* n
2 **for** $c = 1$ *to* ${}_{N}C_{n}$
3 $S^{c} \Leftarrow feature\_combo^{c}; size(S^{c}) = N$
4 $A^{c} \Leftarrow ELM(S^{c}, O)$
5 **end for**
6 $S^{c^*} \Leftarrow argmax_{f_s} A$
7 **end for**
8 $Rank(f_i) \Leftarrow Count(f_i \in S^{c^*})$

Fig. 1. The ELM-based wrapper algorithm.

Table 2
Results of individual domain analysis for the ELM-based wrapper method.

Domain	Best feature set
CO_2	CO_2_fds; CO_2_fd; $CO_2(raw)$
Relative humidity	RH_*fds*10; RH_*fd*2; RH_*mavg*
Temperature	Temp_*fds*10; Temp(*raw*); Temp_*fds*
Pressure	Press_*fds*; Press_*mavg*; Press_*fd*

We conduct the feature selection in two stages. First, we select the best features for each sensing domain. Secondly, we combine these best features to form a multi-domain feature set, which we search exhaustively for the final feature set. These stages are explained in the following paragraphs.

Stage I: Analysis of individual sensing domains. The objective of the individual domain analysis is to select the best features for each environmental parameter. This reduces the feature space to a more manageable number. To begin with, we have six features (that is, including the raw data) for each of the four environmental parameters. We thus have 24 features. We attempt to select the three best features from each sensing domain, thus yielding 12 features. To determine the best features, we apply a wrapper-based ranking. That is, we use the estimation accuracy of a classifier, in our case the ELM, as the criterion to rank features.

The algorithm of the individual domain analysis is outlined in Fig. 1. Note that the algorithm shown applies for a single domain. Let $\{f_i\}$ be the set of features of an environmental parameter. Let N be the size of the combination of features and n the total number of features. In our case, $n = 6$. For each $N = 2, 3, \ldots, n$, all combinations of features are evaluated for their estimation accuracy. In each iteration, the set of features S^c represents a combination of N features. S^c is taken as the input and the occupancy O is made the target output of an ELM. A^c is the accuracy obtained with S^c. We note the combination S^{c^*} that yields the highest accuracy. The rank of each feature is the number of times it appears in the highest accuracy combination.

The algorithm described above was implemented on the extracted features. The algorithm was implemented in MATLAB R2015a on a 1.80 GHz CPU. The approximate running time for the analysis in MATLAB was 7 min and 24 s. The results of the analysis are shown in Table 2.

Stage II: Multi-domain analysis. With the feature space reduced to half by the individual domain analysis, the best features from Stage I are combined into a multi-domain feature set. All possible combinations of the elements of this set were evaluated for their estimation accuracy with an ELM. The best features were CO_2_fd, Temp_*fds*10, Temp(*raw*) and Press_*fds*. It is notable that the pressure data, which has not generally been considered for occupancy estimation, is deemed a relevant feature by the algorithm. Also, the

relative humidity has no clear contributions to the best feature set in this work.

3.3. Selected occupancy models

Occupancy models can reveal occupancy patterns which can be utilized to further improve the estimation accuracy of occupancy. To obtain the exact number of occupants in a multi-occupant area is quite difficult, especially when the occupancy is relatively high. In most situations, for example, discriminating 10 occupants from 11 occupants in one thermal zone is not very meaningful. Also, this discrimination would require high cost devices such as cameras. Normally, the information of occupancy range, i.e. zero, low, medium and high, will be enough for the control systems in buildings. In this work, we consider the estimation of these four states which can be defined based on different deployment scenarios in one thermal zone. The model developed by Richardson et al. [21] is a good candidate for our case. They presented an inhomogeneous Markov chain where the state is the number of active occupants in one house (one thermal zone). The deployment scenario is similar to ours. We only need to change the definition of the Markov state from the number of active occupants to the four occupancy ranges. Recent occupancy models developed by Liao et al. [18] and Chen et al. [19] mainly deal with the simplification of the inhomogeneous Markov chain with a huge dimension of transition probability matrix. They are not suitable for our case. When we consider different deployment scenarios, proper occupancy models need to be chosen. Due to the slow response of the building control system, 15 min resolution will be adequate [35,17,18]. Finally, an inhomogeneous Markov chain model with 96 transition probability matrices of dimension 4×4, where the 4 states are zero, low, medium and high, is constructed for building occupancy modeling.

3.4. Fusion algorithm

Previous works only applied data-driven approaches to estimation building occupancy based on environmental sensor data. They did not consider occupancy dynamics which have been well studied in occupancy modeling [17–19]. In this work, we attempt to incorporate occupancy dynamics with data-driven approaches to enhance the performance of the occupancy estimation system. Thus, we propose to fuse the output of data-driven models with occupancy models using a particle filter algorithm. The particle filter intends to represent the posterior distribution by using a set of particles [36]. In our case, the state of ith particle at time step k is defined as x_k^i which represents the occupancy of a zone. The implementation of the particle filter in this work consists of three steps shown below:

Propagation: The occupancy model we used is a first-order inhomogeneous Markov chain model developed by Richardson et al. [21]. Given ith particle state at time step $k-1$, x_{k-1}^i, the particle can reach the state, $\hat{x}_k^i$, with the transition probability $P(\hat{x}_k^i|x_{k-1}^i)$, $i = 1, 2, \ldots, M$.

Weight calculation: The observation, o_k, can be derived from data-driven models with current environmental sensor data. Given the occupancy model output, $\hat{x}_k^i$, the ith particle weight at time step k can be calculated as

$$\hat{w}_k^i = w_{k-1}^i P\left(o_k|\hat{x}_k^i\right) \tag{10}$$

where $P(o_k|\hat{x}_k^i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-((o_k-\hat{x}_k^i)^2)/2\sigma^2}$. Then, we normalize the weight for each particle as

$$w_k^i = \frac{\hat{w}_k^i}{\sum_{i=1}^M \hat{w}_k^i} \tag{11}$$

Table 3
The resolution and accuracy of the environmental sensors.

Sensor	Measured parameter	Resolution	Accuracy
Rotronic CL11	CO_2	1 ppm	±5% of the measured value
	Temperature	0.05 °C	±0.3 K
	Relative humidity	0.1% RH	<2.5% RH
Lutron MHB-382SD	Pressure	0.1 hPa	±2 hPa

Resampling: We draw M new particles from the current particle set, $\hat{X}_k = \{\hat{x}_k^1, \hat{x}_k^2, \ldots, \hat{x}_k^M\}$, proportional to the weight of each particle and set the weights of all particles to $1/M$. Then, this new particle set, $X_k = \{x_k^1, x_k^2, \ldots, x_k^M\}$, is the desired one to determine the final occupancy of the zone at time step k, y_k, which can be expressed as

$$y_k = \frac{1}{M}\sum_{i=1}^M x_k^i \tag{12}$$

4. Experimental results

In this section, we first introduce the data acquisition process. Then, variables and criteria for performance evaluation are defined. Finally, we present the experimental results and discussions.

4.1. Data acquisition

In this work, we recorded the CO_2, humidity, temperature and pressure levels in the Process Instrumentation Lab at Nanyang Technological University (NTU), Singapore. The lab contains an office space with 24 cubicles and 11 open seats. The room seats 9 PhD students and 11 research staffs. About 6 to 10 of them are regularly present during working hours. Additionally, there are 6 PCs for final year undergraduate students and 5 PCs open to all students. These are less frequently used. In this work, we set the occupancy range of low (1–5 occupants), medium (6–10 occupants), and high (more than 10 occupants). The room is conditioned using both Active Chilled Beam (ACB) and the conventional Variable Air Volume (VAV) systems, and is mechanically ventilated using Air Handling Unit (AHU) which delivers a constant supply of fresh air.

The measurements of CO_2, relative humidity and temperature were done using the CL11 sensor from Rotronic. Pressure levels were measured using the MHB-382SD sensor from Lutron. The sampling time was 1 min. The resolution and accuracy of the sensors are shown in Table 3. The sensors are attached on the wall at a height of 1.1 m from the ground. We applied one sensor per environmental parameter. To record ground truth occupancy, three Internet Protocol (IP) cameras were installed. The layout of the zone is shown in Fig. 2. The main door (referring to the location of camera 1) opens to an outer office space for administrative staffs. Another door (referring to the location of camera 2) opens to a lab area, while the third door remains closed. All the windows are closed. The size of the zone is 20 m long, 9.3 m wide and 2.6 m high. Note that the locations of the sensors are randomly chosen. The optimal placement of the sensors is out of the scope of this paper and will be one of our future works.

The data collected from the sensors was transferred to a laptop using a USB cable. Preprocessing of the data was done in MATLAB. This involved removing missing values, synchronizing the time stamps of the sensors and synchronizing the sensor data with occupancy values. Since the occupancy dynamics on weekdays and weekends are different, we only consider occupancy estimation on weekdays in this work. Note that the occupancy estimation method for weekends is the same. Finally, we collected 32 days of data in

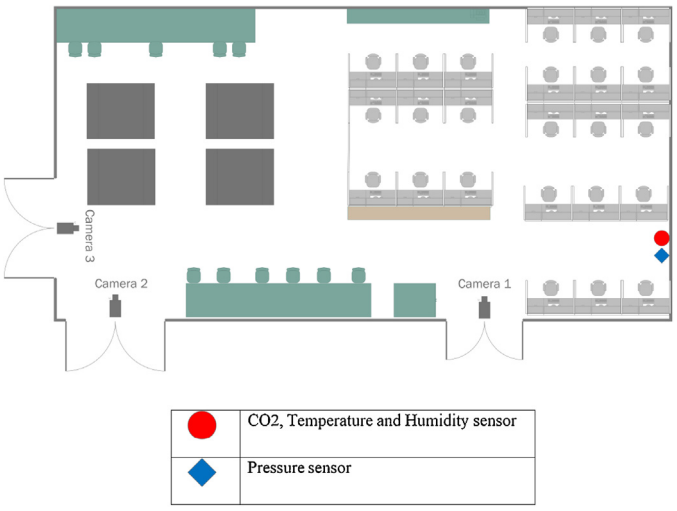


Fig. 2. Layout of the zone.

weekdays for performance evaluation. Among them, we utilize 25 days of data for training and the last 7 days of data for testing.

4.2. Variables and criteria

In this paper, the estimation of building occupancy is treated as a classification problem. Thus, one basic criterion is classification accuracy which is defined as the correctly estimated results against the total number of the examined cases. Another evaluation metric that we used is Normalized Root Mean Square Error (NRMSE) which indicates the magnitude of the estimation error [24]. Time of first arrival (TOFA) and time of last departure (TOLD) are the two most important parameters for building control systems. They will determine the operation period of the control systems. Therefore, the estimation accuracy of these two parameters is vital for an occupancy estimation system. We follow the definitions of these two parameters in [18,19] as follows:

- TOFA: The time of first arrival is the first time when the zone becomes occupied. Precisely, for zone i, if $X_k^i > 0$ and $X_t^i = 0$ for all $t < k$, where X_k^i is the occupancy of zone i at time step k, then, k is the time of first arrival.
- TOLD: The time of last departure is the time from which the zone becomes unoccupied. For zone i, if $X_k^i > 0$ and $X_t^i = 0$ for all $t > k$, then, $k+1$ is the time of last departure.

The normalized mean errors of TOFA and TOLD will be evaluated to demonstrate the performance of the proposed framework.

4.3. Results and discussion

The overall results can be found in Table 4 which presents the classification accuracy, the NRMSE, the normalized mean errors of TOFA and TOLD for the six methodologies with and without fusion. The values of the parameters of some data-driven models, i.e. ELM, SVM, ANN and KNN, are presented in Table 5. All these parameters are adjusted based on grid-search with cross validation of the training data. Specifically, for example, the number of hidden neurons of the ELM algorithm needs to be tuned. We selected the number of hidden neurons from 1 to 100. Fivefold cross validation was employed on the training data with different number of hidden neurons. Then, we chose the optimal number of hidden neurons based on the mean testing accuracy of the fivefold cross validation.

Table 4
The classification accuracy, the NRMSE, the normalized mean errors of time of first arrival and time of last departure for the six approaches with and without fusion. Here, W denotes the approach with fusion and WO denotes the approach without fusion.

Model		Accuracy	NRMSE	TOFA (%)	TOLD (%)
ELM	WO	0.6885	0.2490	7.6340	4.1369
	W	0.7418	0.2099	0.8184	1.2351
SVM	WO	0.6592	0.2660	1.0417	7.4405
	W	0.7167	0.2125	0.8035	1.6369
ANN	WO	0.6427	0.2692	6.6965	3.2441
	W	0.7037	0.2295	1.8899	1.2351
KNN	WO	0.6100	0.2687	14.3452	7.8125
	W	0.7074	0.2126	0.8482	1.5774
LDA	WO	0.6964	0.2450	0.8928	2.2322
	W	0.7390	0.2133	0.7441	1.0417
CART	WO	0.6250	0.2722	14.4345	7.7381
	W	0.7188	0.2176	0.7441	1.4583

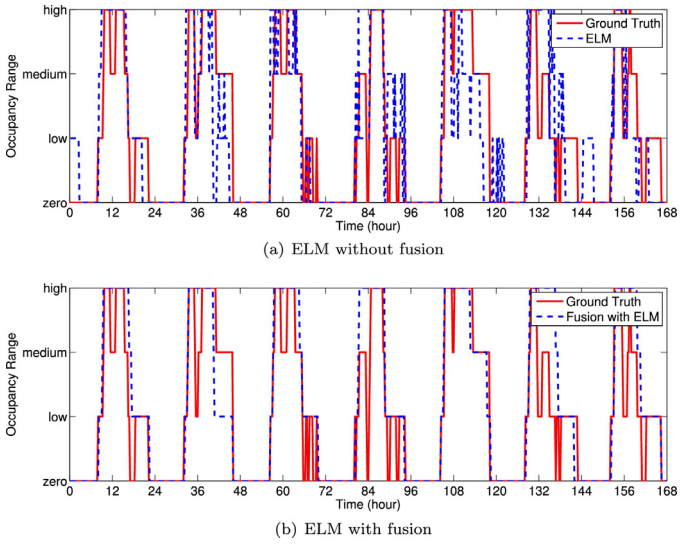


Fig. 3. The classification results of the ELM approach with and without fusion.

The same strategy was applied for determining the optimal parameters for the other data-driven approaches.

Since we takes the occupancy dynamics into consideration, all the approaches with fusion show significant improvements on the estimation accuracy and reductions on the NRMSE. In particular, the fusion algorithm demonstrates an impressive improvement of some 5–14% on estimation accuracy among the different methodologies. Among all the approaches, the ELM with fusion outperforms the others under the two criteria of the estimation accuracy and the NRMSE, which demonstrates the effectiveness of applying this algorithm for building occupancy estimation. For the criteria of the normalized mean errors of TOFA and TOLD, all the approaches with fusion show great reductions. In addition, the LDA and CART with fusion have the lowest estimation error on the normalized mean error of TOFA, and the LDA with fusion has the lowest estimation error on the normalized mean error of TOLD.

Fig. 3 shows an example of the classification results of the 7 testing days for the ELM approach with and without fusion. The results after fusion are smoother. Thus, they are more suitable for building control systems. Meanwhile, the proposed fusion framework eliminates the errors in midnight which may be caused by the slow spread of the environmental parameters, i.e. CO_2, temperature, humidity, after becoming unoccupied. Due to these wrong detections at midnight in the methodologies of ELM, ANN, KNN and CART

Table 5
The values of the parameters of some data-driven models, i.e. ELM, SVM, ANN and KNN.

ELM		SVM			ANN		KNN	
No. of hidden neurons	Activation function	C	ε	Kernel function	No. of hidden neurons	Activation function	Distance metric	No. of neighbors
50	Sigmoid	1.7	0.01	RBF	50	Sigmoid	Euclidean	20

Table 6
The classification accuracy of presence/absence for the six methodologies with and without fusion.

	ELM		SVM		ANN		KNN		LDA		CART	
	WO	W	WO	W	WO	W	WO	W	WO	W	WO	W
Accuracy	0.8729	0.9318	0.8586	0.9281	0.8728	0.9345	0.8152	0.9278	0.9033	0.9345	0.8125	0.9304

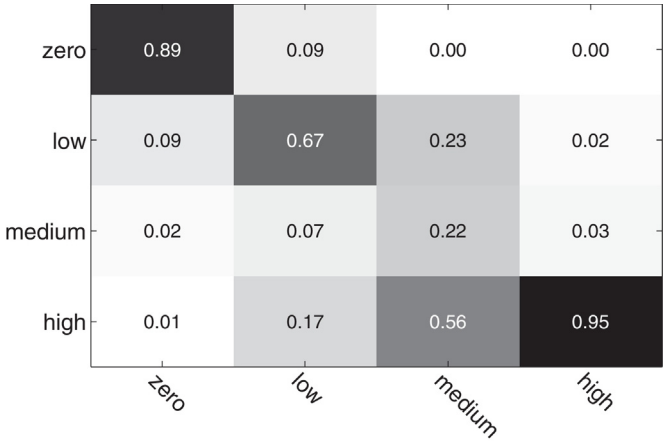


Fig. 4. Confusion matrix of the estimation result of the ELM with fusion.

without fusion, the normalized mean errors of TOFA for the four approaches are quite large. After fusion, the four approaches show significant improvements of the normalized mean errors of TOFA. The SVM and LDA do not contain this wrong detection at midnight. But our fusion approach still shows encouraging improvements. One example of error cumulative distribution functions (CDF) of TOFA for the ELM approach with and without fusion is shown in Fig. 5(a). Note that each step refers to 15 min. The large error is caused by the wrong estimation of the occupancy range at midnight.

Fig. 4 presents the confusion matrix of the estimation result of the ELM with fusion. We can find that the detections of zero and high occupancy are easier than the detections of low and medium occupancy. The possible reason can be that, since the environmental parameters are smoothly changing, the low and high thresholds can be easily identified, but the low and medium ones can be confusing. Moreover, the confusion matrix also indicates a high detection accuracy of presence and absence.

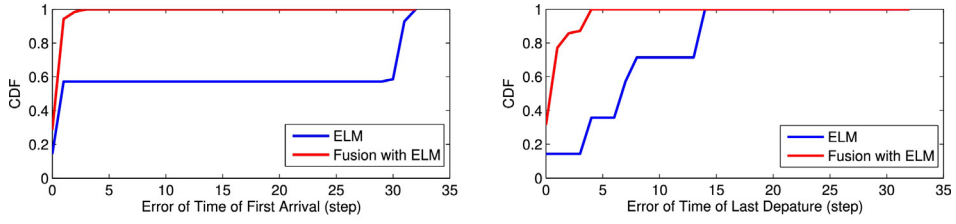
Another observation is that the normalized mean error of TOLD is larger than that of TOFA in most cases after fusion. To explain this phenomenon, we take the environmental parameter of CO_2 as an example. When the zone becomes first occupied in the morning, the CO_2 level will increase to indicate occupancy change. After that, the zone will become occupied for almost the entire day, thus, the CO_2 level accumulates to a relatively high level in the evening. Then, when after all occupants have left, the CO_2 level will take some time to decrease to a low level which indicates vacancy. That is the main drawback of the environmental parameter based occupancy estimation method. Nevertheless, according to Fig. 5(b), the mean error of TOLD has been greatly reduced by our proposed fusion algorithm, as compared to that without fusion.

Presence and absence of occupant information is important for applications such as lighting control. The classification accuracies of presence/absence for the six methodologies with and without fusion are also tested and shown in Table 6. The fusion algorithm demonstrates an impressive improvement of around 3–12% among the different approaches. Among all the approaches, ANN and LDA with fusion perform the best. The average detection accuracy for all the approaches after fusion is as high as 93% with only the environmental sensors, rather than any extra occupancy related sensors, e.g. motion sensor [11,12], acoustic sensor [11,12,25], PIR sensor [12,25] and light sensor [25,26].

The state-of-the-art systems for occupancy estimation with environmental sensor data are applying various data-driven approaches. The approaches of SVM, ANN, KNN, LDA and CART are widely used and have been shown to be effective for occupancy estimation. The estimation results obtained from these approaches can be treated as the state-of-the-art results. By taking occupancy dynamics into consideration, we improve the estimation performance of these data-driven approaches by using our proposed fusion framework. Moreover, we implemented a new approach of ELM with regularization term, which achieved competitive performance after fusion for building occupancy estimation.

4.4. Limitations

- Environmental sensors: Due to the slow response of the environmental parameters to the change of occupancy, the intermediate transition between occupied and unoccupied states can be



(a) TOFA for ELM with and without fusion (b) TOLD for ELM with and without fusion

Fig. 5. The CDFs of TOFA and TOLD for the ELM approach with and without fusion.

difficult to detect. This is one of the limitations of employing environmental sensors for occupancy estimation. However, this can be easily resolved by using some low cost motion sensors, which will be further investigated in our future works. We also note that the slow response of the environmental sensors may not be able to track the frequent change of occupancy, which can be observed from Fig. 3. Note that this frequent change of occupancy is unfavorable for the slow response control systems.

- Ground truth: The occupancy estimation using data-driven approaches involves a model training process which requires the ground truth occupancy, which in this paper is obtained by using cameras. However, in real situations, the cameras may not be available. One possible way to obtain that can be based on the passive localization systems in [37,38]. Another possible solution is to rely on crowdsourcing which requires the occupants to report their own profiles.

5. Conclusion and future works

In this paper, we propose a fusion framework for building occupancy estimation with environmental parameters. Data-driven models that include ELM, SVM, ANN, KNN, LDA and CART are employed to achieve an initial estimation of building occupancy. To select the best feature set, we perform an ELM-based wrapper method owing to the extremely fast learning speed of the ELM algorithm. Then, we fuse the results of data-driven models with well-developed occupancy models which can extract occupancy patterns to further improve the estimation accuracy. Experiments have been conducted in an office environment at a university campus. The results show impressive improvements of estimation accuracy after using our proposed fusion framework. Moreover, two important parameters of time of first arrival (TOFA) and time of last departure (TOLD) have been defined for performance evaluation. The proposed fusion framework significantly reduced the estimation errors of these two parameters. We also tested the detection accuracy of the presence and absence of the zone with our proposed fusion approach. The improvements of the detection accuracy among the different methodologies are in the range of 3–12%. After implementing our fusion framework, the detection accuracy is around 93% with just the environmental parameters.

In future works, we will attempt to further improve the performance of the occupancy estimation system by incorporating schedule information, e.g. meetings, classes, etc. Moreover, we intend to develop an occupancy-driven control system based on our occupancy estimation results and quantify the potential energy saving by using energy simulation tools, such as EnergyPlus and DOE2.

Acknowledgements

This work is jointly supported by the Building Efficiency and Sustainability in the Tropics (SinBerBEST) program which is funded by Singapore's National Research Foundation and led by the University of California, Berkeley (UCB) in collaboration with Singapore Universities, and Singapore's National Research Foundation under NRF2011NRF-CRP001-090.

References

[1] U. Sbci, Buildings and Climate Change: A Summary for Decision-makers, United Nations Environmental Programme, Sustainable Buildings and Climate Initiative, Paris, 2009, pp. 1–62.
[2] A.P. Boranian, B. Zakirova, J.N. Sarvaiya, N.Y. Jadhav, P. Pawar, Z. Zhe, Building Energy Efficiency R&D Roadmap, Building and Construction Authority (BCA), Singapore, 2013, pp. 1–44.
[3] Y. Agarwal, B. Balaji, R. Gupta, J. Lyles, M. Wei, T. Weng, Occupancy-driven energy management for smart building automation, in: Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building, ACM, 2010, pp. 1–6.
[4] T. Leephakpreeda, Adaptive occupancy-based lighting control via grey prediction, Build. Environ. 40 (7) (2005) 881–886.
[5] R.H. Dodier, G.P. Henze, D.K. Tiller, X. Guo, Building occupancy detection through sensor belief networks, Energy Build. 38 (9) (2006) 1033–1043.
[6] N. Li, G. Calis, B. Becerik-Gerber, Measuring and monitoring occupancy with an RFID based system for demand-driven HVAC operations, Autom. Construct. 24 (2012) 89–99.
[7] V.L. Erickson, S. Achleitner, A.E. Cerpa, POEM: power-efficient occupancy-based energy management system, in: 2013 ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), IEEE, 2013, pp. 203–216.
[8] B. Balaji, J. Xu, A. Nwokafor, R. Gupta, Y. Agarwal, Sentinel: occupancy based HVAC actuation using existing WiFi infrastructure within commercial buildings, in: Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems, ACM, 2013, p. 17.
[9] T. Labeodan, W. Zeiler, G. Boxem, Y. Zhao, Occupancy measurement in commercial office buildings for demand-driven control applications a survey and detection system evaluation, Energy Build. 93 (2015) 303–314.
[10] J. Zhao, B. Lasternas, K.P. Lam, R. Yun, V. Loftness, Occupant behavior and schedule modeling for building energy simulation through office appliance power consumption data mining, Energy Build. 82 (2014) 341–355.
[11] B. Dong, B. Andrews, K.P. Lam, M. Höynck, R. Zhang, Y.-S. Chiou, D. Benitez, An information technology enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network, Energy Build. 42 (7) (2010) 1038–1046.
[12] Z. Yang, N. Li, B. Becerik-Gerber, M. Orosz, A systematic approach to occupancy modeling in ambient sensor-rich buildings, Simulation 90 (8) (2014) 960–977.
[13] M.K. Masood, Y.C. Soh, V.W.-C. Chang, Real-time occupancy estimation using environmental parameters, in: 2015 International Joint Conference on Neural Networks (IJCNN), IEEE, 2015, pp. 1–8.
[14] K.P. Lam, M. Höynck, R. Zhang, B. Andrews, Y.-S. Chiou, B. Dong, D. Benitez, Information-theoretic environmental features selection for occupancy detection in open offices, in: P.A. Strachan, N.J. Kelly, M. Kummert (Eds.), Eleventh International IBPSA Conference, Citeseer, 2009, pp. 1460–1467.
[15] T. Ekwevugbe, N. Brown, V. Pakka, Real-time building occupancy sensing for supporting demand driven HVAC operations, in: 13th International Conference for Enhanced Building Operations, Montreal, Quebec, 2013.
[16] M.A. Hall, Correlation-based feature selection for machine learning (Ph.D. thesis), The University of Waikato, 1999.
[17] J. Page, D. Robinson, N. Morel, J.-L. Scartezzini, A generalised stochastic model for the simulation of occupant presence, Energy Build. 40 (2) (2008) 83–98.
[18] C. Liao, Y. Lin, P. Barooah, Agent-based and graphical modelling of building occupancy, J. Build. Perform. Simul. 5 (1) (2012) 5–25.
[19] Z. Chen, J. Xu, Y.C. Soh, Modeling regular occupancy in commercial buildings using stochastic models, Energy Build. 103 (2015) 216–223.
[20] D. Wang, C.C. Federspiel, F. Rubinstein, Modeling occupancy in single person offices, Energy Build. 37 (2) (2005) 121–126.
[21] I. Richardson, M. Thomson, D. Infield, A high-resolution domestic building occupancy model for energy demand simulations, Energy Build. 40 (8) (2008) 1560–1566.
[22] C. Wang, D. Yan, Y. Jiang, A novel approach for building occupancy simulation Building Simulation, vol. 4, Springer, 2011, pp. 149–167.
[23] V.L. Erickson, Y. Lin, A. Kamthe, R. Brahme, A. Surana, A.E. Cerpa, M.D. Sohn, S. Narayanan, Energy efficient building environment control strategies using real-time occupancy measurements, in: Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings, ACM, 2009, pp. 19–24.
[24] Z. Yang, N. Li, B. Becerik-Gerber, M. Orosz, A multi-sensor based occupancy estimation model for supporting demand driven HVAC operations, in: Proceedings of the 2012 Symposium on Simulation for Architecture and Urban Design, Society for Computer Simulation International, 2012, p. 2.
[25] A. Khan, J. Nicholson, S. Mellor, D. Jackson, K. Ladha, C. Ladha, J. Hand, J. Clarke, P. Olivier, T. Plötz, Occupancy monitoring using environmental & context sensors and a hierarchical analysis framework, Embed. Syst. Energy Effic. Build. (BuildSys) (2014) 90–99.
[26] L.M. Candanedo, V. Feldheim, Accurate occupancy detection of an office room from light, temperature, humidity and CO_2 measurements using statistical learning models, Energy Build. 112 (2016) 28–39.
[27] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, Extreme learning machine: theory and applications, Neurocomputing 70 (1) (2006) 489–501.
[28] G.-B. Huang, H. Zhou, X. Ding, R. Zhang, Extreme learning machine for regression and multiclass classification, IEEE Trans. Syst. Man Cybern. B: Cybern. 42 (2) (2012) 513–529.
[29] V. Vapnik, The Nature of Statistical Learning Theory, Springer Science & Business Media, 2013.
[30] S.-C. Wang, Artificial neural network, in: Interdisciplinary Computing in Java Programming, Springer, 2003, pp. 81–100.
[31] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (2) (1936) 179–188.
[32] P. Xanthopoulos, P.M. Pardalos, T.B. Trafalis, Linear discriminant analysis, in: Robust Data Mining, Springer, 2013, pp. 27–33.
[33] W.-Y. Loh, Classification and regression trees, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 1 (1) (2011) 14–23.

[34] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1) (1997) 273–324.
[35] N. Nassif, S. Kajl, R. Sabourin, Optimization of HVAC control system strategy using two-objective genetic algorithm, HVAC&R Res. 11 (3) (2005) 459–486.
[36] M.S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, IEEE Trans. Signal Process. 50 (2) (2002) 174–188.
[37] F. Colone, P. Falcone, C. Bongioanni, P. Lombardo, WiFi-based passive bistatic radar: data processing schemes and experimental results, IEEE Trans. Aerosp. Electr. Syst. 48 (2) (2012) 1061–1079.
[38] P. Falcone, F. Colone, A. Macera, P. Lombardo, Localization and tracking of moving targets with WiFi-based passive radar, in: 2012 IEEE Radar Conference, IEEE, 2012, pp. 0705–0709.