DA-HOC: Semi-Supervised Domain Adaptation for Room Occupancy Prediction using CO_2 Sensor Data

Irvan B. Arief-Ang
School of Science
RMIT University
Melbourne, Australia 3000
irvan.ariefang@rmit.edu.au

Flora D. Salim
School of Science
RMIT University
Melbourne, Australia 3000
flora.salim@rmit.edu.au

Margaret Hamilton
School of Science
RMIT University
Melbourne, Australia 3000
margaret.hamilton@rmit.edu.au

ABSTRACT

Human occupancy counting is crucial for both space utilisation and building energy optimisation. In the current article, we present a semi-supervised domain adaptation method for carbon dioxide - Human Occupancy Counter (DA-HOC), a robust way to estimate the number of people within in one room by using data from a carbon dioxide sensor. In our previous work, the proposed Seasonal Decomposition for Human Occupancy Counting (SD-HOC) model can accurately predict the number of individuals when the training and labelled data are adequately available. DA-HOC is able to predict the number of occupancy with minimal training data, as little as one-day data. DA-HOC accurately predicts indoor human occupancy for a large room using a model trained from a small room and adapted to the larger room. We evaluate DA-HOC with two baseline methods - support vector regression technique and SD-HOC model. The results demonstrate that DA-HOC's performance is better by 12.29% in comparison to SVR and 10.14% in comparison to SD-HOC.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing design and evaluation methods**;

KEYWORDS

Transfer learning; domain adaptation; ambient sensing; building occupancy; presence detection; number estimation; cross-space modeling; contextual information

1 INTRODUCTION

Predicting human occupancy in a building is crucial for maximising building utilisation and improving the building management and operations. However, to predict the number of occupants accurately, there needs to be sufficient labelled data for training the predictors and validating the model.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
BuildSys '17, November 8–9, 2017, Delft, Netherlands
© 2017 Copyright held by the owner/author(s). Publication rights licensed to Association for Computing Machinery.
ACM ISBN 978-1-4503-5544-5/17/11…$15.00
https://doi.org/10.1145/3137133.3137146

Many time-series prediction problems suffer from missing labelled data and are unable to achieve acceptable accuracy. Unfortunately, there is an abundance of unlabelled time series data which exists for either training or test data for machine learning model input. To utilise an unlabelled dataset, domain adaptation techniques need to be implemented. There are three types of domain adaptation based on the availability of the labels in the test dataset:

(1) Unsupervised domain adaptation;
(2) Semi-supervised domain adaptation; and
(3) Supervised domain adaptation.

To address the unlabelled problem, the unsupervised domain adaptation is the best option because supervised and semi-supervised domain adaptation techniques will not work in the absence of labelled data. For semi-supervised domain adaptation, the learning model needs to have a set of labelled source samples, a set of unlabelled source samples and an unlabelled set of target samples.

In this paper, we focus on implementing a domain adaptation technique for indoor human occupancy prediction. Data obtained from the U.S. Department of Energy indicates that 35% - 45% of total maintenance costs within a building are spent on heating, ventilation, and air conditioning (HVAC) [10]. Reducing HVAC usage will massively lessen overall energy consumption. A Building Management System (BMS) can intelligently adjust the HVAC based on the occupancy pattern.

Many practical real life advantages can be obtained by knowing the number of occupants residing in one building or one room at any given time. The benefits include reducing energy consumption, human indoor comfort and security area. Knowing the number of people in advance can be used to adjust the HVAC to reduce power cost if there is nobody inside a room for a given period. Comfort can be improved by increasing or decreasing the temperature based on the crowdedness of a room. For security, if the owner of a building knows that there should be no one inside the building at a particular time, detecting a person could indicate a security breach.

Understanding the occupancy pattern is also necessary for space utilisation. If the usage pattern of two similar rooms is below 50% for both rooms, it may be more beneficial to move the audience of one room to another and utilise the newly vacant room for other more useful purposes. Maximising space and room utilisation creates greater space efficiency and could increase both individual and group productivity.

In this paper, we propose a domain adaptation model for carbon dioxide - human occupancy counter (DA-HOC). Domain adaptation concept is shown in Figure 1. The human occupancy counting baseline model accurately predicts the number of people when the training and labelled data are available. We have designed a novel, semi-supervised domain adaptation for a human occupancy counting model so it can be implemented in any room without adequate labelled data. We compare our results with two baseline methods: Support Vector Regression (SVR) and Seasonal Decomposition for Human Occupancy Counting (SD-HOC) [3].

Domain adaptation is useful as most of the time we do not have (or only have a very minimum amount of) historical data. With domain adaptation, we leverage the previous working model and adapt the model to the new environment. We picked the latest baseline model for indoor human occupancy and integrated our domain adaptation method on the top of baseline model to make new DA-HOC.

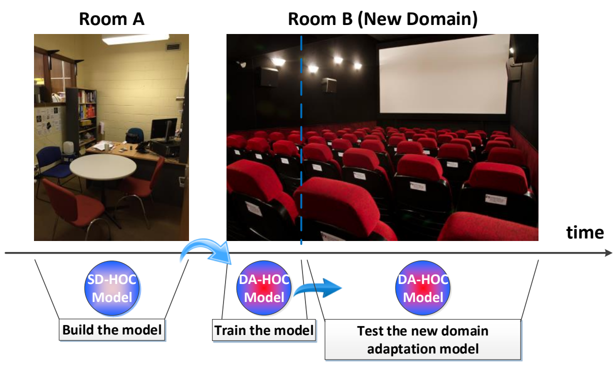

Figure 1: Illustration for the Main Algorithm DA-HOC.

1.1 Research Contribution

The main contributions of this paper are two-fold. First, we propose a new semi-supervised domain adaptation data preprocessing method for human occupancy counting. Second, by making use of this preprocessing method, we present a novel framework to address semi-supervised domain adaptation problems. The performance is evaluated using a real world dataset and is compared with state-of-the-art techniques to measure the robustness of our new framework.

The remainder of the paper is organised as follows. Section 2 presents the background of domain adaptation, specifically semi-supervised domain adaptation and carbon dioxide - human occupancy counting model. Section 3 covers the problem definition. Section 4 introduces the methodology for DA-HOC model. Section 5 describes the experiments, results and comparisons with state-of-the-art algorithms as a baseline. Section 6 concludes the paper with directions for future work.

2 BACKGROUND & RELATED WORK

2.1 Related Work

Data mining and machine learning technologies have made tremendous achievements in many knowledge engineering areas covering classification, clustering and regression [23, 26, 28]. Unfortunately, most of machine learning techniques are designed to work with the best results under one condition, the training and test data are extracted from the same distribution and the same feature space. Every time the distribution changes, the majority of the statistical models need to be re-created with new gathered training data. Due to this reason, domain adaptation research is emerging in the recent years [20, 24] as a way of leverage to make a prediction on the dataset that contains minimum labelled data.

There is a variety of knowledge engineering methods that can fully utilise the benefit of the domain adaptation technique, for example in the field of image recognition [17] and web-document classification [1, 7, 14, 22]. A more prominent example where domain adaptation can shine is to classify product categories and their product reviews automatically. In [5], a domain adaptation technique is used to save a significant amount of labelling effort and be more cost effective. A classification model is adapted and trained on some products to help classification models adapt to the other products to reduce the effort for annotating and labelling reviews for various products.

Domain adaptation can also be implemented with deep learning [19]. The author performed the domain adaptation technique for fault analysis. Domain adaptation is suitable to analyse regression series [6]. Other research in semi-supervised domain adaptation including [8], [18] and [27]. A pipeline for unsupervised domain adaptation is proposed by [12]. An active domain adaptation framework is proposed by [16] for the purpose of adding the label for unlabelled target data and to generate effective label queries during active learning. Both supervised and unsupervised domain adaptation experiments are applied for activity recognition from simple in-home sensors by [15].

The other case for domain adaptation is when data can be easily outdated. Data that was gathered during a given period might have a different distribution when it is collected in the later period. This problem usually appears on indoor WiFi localisation, where a system may be required to locate user position based on WiFi data. The WiFi signal value varies throughout time and the recalibration process can be expensive. Domain adaptation technique can adopt the localisation model from the previous period and adjust it for later period [21].

2.2 Semi-supervised Domain Adaptation

In this paper, we focus on covering background for semi-supervised domain adaptation.

Within pattern classification, methods for semi-supervised domain adaptation are designed to handle cases in which one cannot assume that training and test sets are sampled from the same distribution. The reason is because they are collected from different domains. However, some unlabelled samples that belong to the same domain as the test set are available, enabling the learner to adapt their parameters.

By definition, semi-supervised domain adaptation may be explained as is: Given a source domain D_S and a corresponding learning task L_S, a target domain D_T and a corresponding learning task L_T, semi-supervised domain adaptation aims to improve the learning of the target predictive function $f_T(\cdot)$ in D_T using the knowledge in D_S and L_S, where $D_S \neq D_T$ and $L_S = L_T$. In addition, some unlabelled target-domain data must be available at training time.

In the semi-supervised domain adaptation setting, the source and target tasks are the same, while the source and target domains are different. In this situation, a small amount of labelled data in the target domain is available while labelled data from the source domain exist. Besides, according to different situations between the source and target domains, we can further categorise the semi-supervised domain adaptation setting into two cases.

(1) The feature spaces (χ) between the source and target domains are different, $\chi_S \neq \chi_T$.
(2) The feature spaces between domains are the same, $\chi_S = \chi_T$, but the marginal probability distributions of the input data are different, $P(\chi_S) \neq P(\chi_T)$.

In this paper, we focus on the second case where the feature space between source and domain is similar, but there is a difference in their marginal probability distribution.

2.3 Carbon Dioxide - Human Occupancy Counter Model

Many techniques can be utilised to predict human occupancy prediction. Dutta P.K. et. al. [11] used a radar to predict human. Other research used multiple ambient sensors including temperature, humidity, illuminate, sound and CO_2 [2]. Wi-Fi power measurement can also be used to estimate the occupancy [9].

Human occupancy prediction with CO_2 is gaining in popularity as a model and PerCCS is a model with a non-negative matrix factorization method to count people [4] using only one predictor in CO_2. In predicting vacant occupancy, they achieved up to 91% but suffer from 15% accuracy in predicting the number of occupants. PerCCS used SVR as their baseline method.

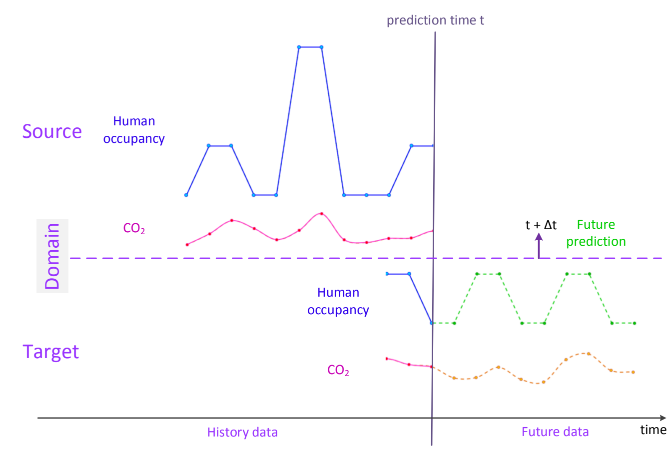

Figure 2: Domain adaptation prediction scenario for continuous t in target domain showing the amount of CO_2 fluctuations. The fundamental task is to predict the number of occupants at time t+Δt.

Seasonal Decomposition for Human Occupancy Counting (SD-HOC) model is used to solve the non-linear correlation issue between CO_2 and indoor human occupancy by decomposing both CO_2 and occupancy data [3]. In seasonal trend decomposition (STD), there are three main components. The trend feature (TF_t) reflects the long-term progression of the time series during its secular variation. The seasonal feature (SF_t) is a systematic and regularly repeated event during a short period of time. The irregular feature (IF_t also known as error or residual) is a short-term fluctuation from the time series and is the remains after the trend and season features have been removed.

This general SD-HOC formula will be applied to both time series for CO_2 dataset and human occupancy datasets. The main formula is shown as below:

$$O(t) = TF_O(t) + SF_O(t) + IF_O(t) + ZF_O(t) \quad (1)$$

$O(t)$ Indoor human occupancy
$TF_O(t)$ Trend feature for occupancy
$SF_O(t)$ Seasonal feature for occupancy
$IF_O(t)$ Irregular feature for occupancy
$ZF_O(t)$ Zero pattern adjustment for occupancy

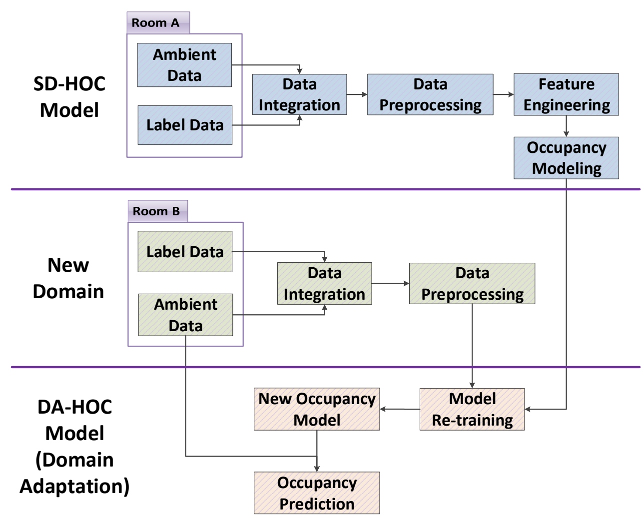

Figure 3: Data Collection and Analysis Framework.

3 PROBLEM DEFINITION

We proposed a solution to build a transferable model to be used for indoor human occupancy prediction by just using carbon dioxide (CO_2) in other domain. From Figure 2, there is a full labelled dataset from the source domain. As occupancy data is dependent on CO_2, a model can be built to predict indoor human occupancy from both source and target domain CO_2 sensors. We then develop a semi-supervised domain adaptation method from the SD-HOC model (the source model) and integrate this with the limited labelled data from the target domain to predict the future human occupancy.

3.1 Scenario

Assume χ represents the length of a time series and is expressed as $\chi = \{\chi_1, \chi_2, ..., \chi_q\}$, where q means the number of sample points.

There are two types of time series, source domain time series (χ_S) and target domain time series (χ_T). For each time series datasets, each has two aspects, CO_2 concentration C and indoor human occupancy O. To summarise, in total we have four aspects:

- CO_2 concentration from source domain C_S, defined as $C_S = \{C_{S1}, C_{S2}, ..., C_{Sq}\}$
- CO_2 concentration from target domain C_T, defined as $C_T = \{C_{T1}, C_{T2}, ..., C_{Tq}\}$
- Indoor human occupancy from source domain O_S, defined as $O_S = \{O_{S1}, O_{S2}, ..., O_{Sq}\}$
- Indoor human occupancy from target domain O_T, defined as $O_T = \{O_{T1}, O_{T2}, ..., O_{Tq}\}$

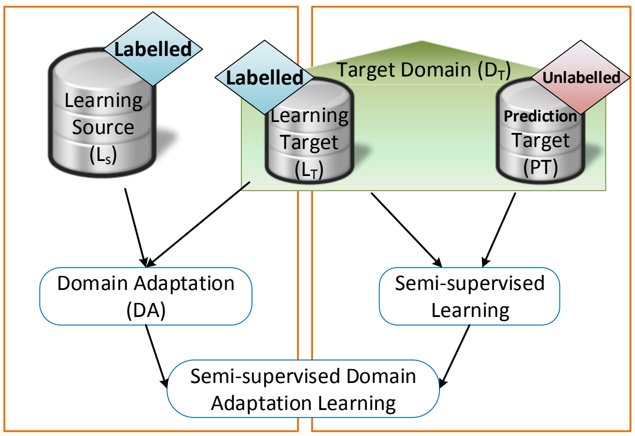

Figure 4: Overview of the Semi-supervised Domain Adaptation Learning Method.

3.2 Domain Adaptation

In the non-domain adaptation problem, χ_S and χ_T are both assumed to have been drawn from the same distribution, χ. In the domain adaptation setting, however, we would like to apply our trained classifier to examples drawn from a distribution different from the one upon which it was trained. We, therefore, assume there are two separate distributions, χ_S and χ_T, from which data may be drawn.

It is important to note that one can have a semi-supervised algorithm that does or does not make the domain adaptation assumption, and vice versa. Much of the work in this paper, however, was inspired by the belief that, although distinct, these problems are nevertheless intimately related. More specifically, when trying to solve a transfer problem between two domains, it seems intuitive that looking at the data of the target domain during training will improve performance over ignoring this source of information. Similarly, even if one believes he is not solving a transfer problem, it may still be beneficial to model one's training and test data as if they were not identically distributed.

For the domain adaptation prediction model, two problems need to be focused on:

- Data preprocessing method for DA-HOC to ensure source domain and target domain have the same granularity.
- Develop semi-supervised domain adaptation for each SD-HOC component (TF_t, SF_t, IF_t and ZF_t).

For the domain adaptation method, we build a model using one dataset from a single location as source domain (χ_S) and then construct a semi-supervised domain adaptation framework so the previous model can be utilised to predict in another environment as target domain (O_T).

4 METHODOLOGY

There are three phases of methodology steps that are implemented for the data collection and analysis framework as shown in Figure 3. The first step is to build the SD-HOC model to the source model that have full labelled data. The second step is to train the model with new labelled data from the target domain. The final step is to implement the DA-HOC model to predict the human occupancy. Semi-supervised domain adaptation concept is shown in Figure 4.

DA-HOC model consists of three main phases: the preprocessing CO_2 data, the main algorithm and the post adjustment model.

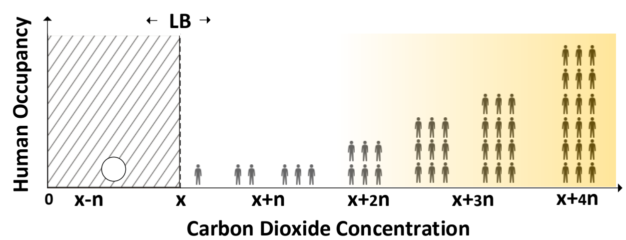

Figure 5: Correlation between human occupancy number and carbon dioxide concentration.

4.1 Preprocessing DA-HOC

During the preprocessing phase, both data from source and target are merged. Autocorrelation and the line of best fit and time lag analysis [3] are an integral part of the preprocessing phase. The purpose of this step is to adjust the time for the duration of CO_2 gas to permeate and populate the whole room. The time lag value differs for each case and the larger the room is, the greater the time lag value.

$$LB_{CO_2} = \frac{\sum_{n=1}^{N_{max}-1} C_{min}}{N_{max} - 1} \tag{2}$$

LB_{CO_2}	Lower bound for CO_2 value
n	Counter number
N_{max}	Total number of local maximum points
C_{min}	local minimum CO_2 concentration value located after local maximum n

In Equation 2, we calculate the lower bound value (LB) to adjust the time series and find a lower threshold for the CO_2 concentration when the room is supposed to be vacant. This value reduces the calculation complexity for the main algorithm DA-HOC. Figure

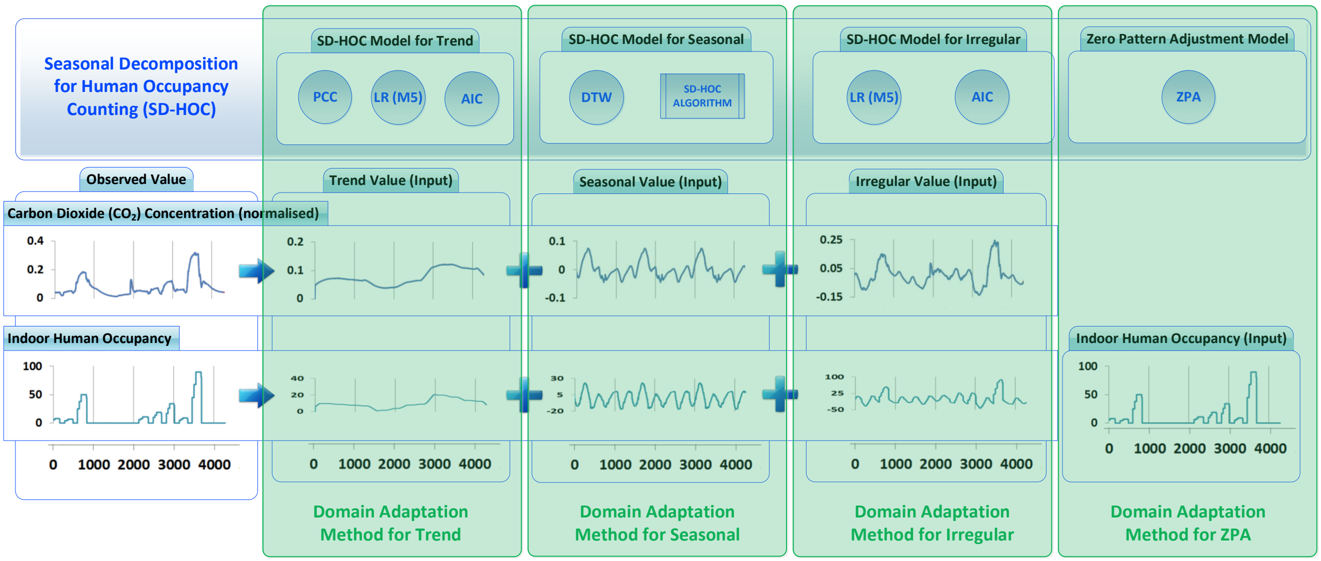

Figure 6: Semi-supervised domain adaptation method for Seasonal Decomposition for Human Occupancy Counter (DA-HOC).

5 shows that the *LB* value fluctuated based on the rooms' characteristics, conditions and the demographics of humans inside it. Complexity reduction algorithm is shown at Algorithm 1.

Algorithm 1 Infused LB_{CO_2} with the dataset for complexity reduction calculation

procedure VACANT_LOWER_BOUND(LB_{CO_2})
  $len \leftarrow 0$ ▷ len: Length for C_q
  $temp \leftarrow 0$
  for each node $c \in C_q$ **do**
    len++
    if $c \leq LB_{CO_2}$ **then**
      $temp \leftarrow len$
      $C_q[temp] \leftarrow LB_{CO_2}$
    end if
  end for
end procedure

The value of LB_{CO_2} is used for the input parameter in DA-HOC to modify any concentration value which is lower than the threshold value. Binary prediction accuracy is boosted with this carbon dioxide concentration uniformity of threshold value for the vacant room.

Temporal frequencies and time intervals from both source and target domains with each time lag value datasets are then correlated and adjusted based on the differences in temporal frequencies. Preprocessing values can be extracted using the formula in Equation 3.

With the preprocessing value, the dataset with the higher temporal frequency can be adjusted by reducing the period to equalise the lower temporal frequency dataset. By reducing the period, excess data issues can be mitigated and the higher temporal frequency

$$A = \left|\frac{\alpha \times F_S^{Temp} - (1-\alpha)F_T^{Temp}}{T\bar{i}}\right| \tag{3}$$

A	Preprocessing value
α	Weight value for temporal frequency for source
F_S^{Temp}	Temporal frequency for source value
F_T^{Temp}	Temporal frequency for target value
$T\bar{i}$	Time interval value

dataset undergoes a data reduction process. Once both datasets have similar temporal frequencies, the preprocessing phase is complete.

4.2 Main algorithm DA-HOC

Main algorithm DA-HOC divides the datasets into three portions. The source dataset (D_S), the target dataset (D_T) and the learning target dataset (L_T). The source dataset contains all the labelled datasets from the source domain, the target dataset contains all the unlabelled datasets from the target domain and the learning target dataset contains all the labelled datasets from the target domain. The learning target dataset quantity is exiguous compared to both source and target datasets. A proportional comparison between learning target and target datasets is 0.1 or less. Figure 2 gives an illustration for this division.

The second data partition is divided between the vacant room prediction and occupied room prediction. The original SD-HOC algorithm does not differentiate between these two because by retraining the model for every domain, the new model can capture the condition and the essence from each domain. Due to the majority of target domain dataset is unlabelled, however, differentiating the vacant and occupied rooms can be designed by the model in a more proportional way.

As DA-HOC is the adapted from SD-HOC, the data set is factorised to four different features: trend feature (TF_t^{DA}), seasonal feature (SF_t^{DA}), irregular feature (IF_t^{DA}) and zero pattern adjustment feature (ZF_t^{DA}) as shown in Figure 6. Below is the general formula for DA-HOC shown in Equation 4.

$$O_t^{DA} = TF_t^{DA} + SF_t^{DA} + IF_t^{DA} + ZF_t^{DA} \tag{4}$$

4.2.1 Domain Adaptation Method for the Trend Feature (TF_t^{DA}). The trend feature behaves similarly between CO_2 concentration and human occupancy due to its linear correlation. The first algorithm for trend feature model is the maximum mean discrepancy (MMD) [13]. The purpose of MMD is to bring the average of the two distributions closer to each other while it projects the data into its principal components directions of the full data including the source and target domain. We want to find a function that assumes different expectations on two separate distributions.

Equation 5 definition: Let $\mathcal{F}$ be a class of functions f:$X \to \mathbb{R}$. Let p and q be Borel probability distributions and let $X = (x_1, x_2, ..., x_m)$ and $Y = (y_1, y_2, ..., y_m)$ be the samples composed of independent and identically distributed observations drawn from p and q, respectively. We define the maximum mean discrepancy (MMD) and its empirical estimate as

$$MMD[\mathcal{F}, p, q] := \sup_{f \in F} \left(\mathbf{E}_p \left[f(x) \right] - \mathbf{E}_q \left[f(y) \right] \right) \tag{5}$$

Putting trend feature source domain (TF_S^{DA}) as p and trend feature target domain (TF_S^{DA}) as q , we can elaborate the MMD formula as shown in Equation 6.

$$MMD[\mathcal{F}, TF_S^{DA}, TF_T^{DA}] := \sup_{f \in F} \left(\frac{1}{m} \sum_{i=1}^{m} f(TF_{Si}^{DA}) - \frac{1}{n} \sum_{i=1}^{n} f(TF_{Ti}^{DA}) \right) \tag{6}$$

With DA-HOC, understanding the trend feature distribution's distance for each domain is necessary due to lacking training data from the related environment dataset, trend component pseudo model needs to be created so that it mimics the model from other domain and compare the boundary between the top and the bottom value for human occupancy. The CO_2 concentration behaves similarly due to normalisation process during preprocessing data. Finding the mapping between CO_2 level and the number of people inside is the key contribution from the trend feature model. The general function with Taylor's expansion for trend feature analysis is shown in Equation 7.

$$TF_t^{DA} = \alpha_0 + \alpha_1(TF_t^{DA}) + \alpha_2(TF_t^{DA})^2 + ... + \alpha_n(TF_t^{DA})^n + \epsilon \tag{7}$$

4.2.2 Domain Adaptation Method for the Seasonal Feature (SF_t^{DA}). The seasonal feature is the repeating part of data during a short period. Due to its nature, once DA-HOC model can obtain the repeating pattern for the seasonal feature, the prediction is more accurate. The SD-HOC seasonal feature focuses on the regularly repeated event during a short period. Domain adaptation method for the seasonal feature will learn from the short amount of new testing data and then correlate them with the original model to find the similarity. Duration of each repeating is essential and will be translated to the new model. The full algorithm for the seasonal feature is shown in Algorithm 2. With DA-HOC model, data pattern from the seasonal feature from source domain is extracted and from its pattern, the seasonal feature from target domain can be deciphered.

Algorithm 2 Finding a repeated pattern sequence inside seasonal feature

procedure Repeated_Sequence(SF_t^{CD}, SF_t^{DA})
  $len \leftarrow 0$ ▷ len: Length for SF_t^{CD}
  $a \leftarrow SF_t^{DA}[len]$ ▷ a: Start Point
  for each node $i \in SF_t^{DA}$ **do**
    $len{+}{+}$
    $SF_t^{CD} \leftarrow SF_t^{CD} + SF_t^{DA}[i]$
    if $a = SF_t^{DA}[i]$ **then**
      if DTW(SF_t^{CD},$SF_t^{DA}[i+1..i+len]$) > 95 **then**
        $SF\_Fin_t^{DA} \leftarrow SF_t^{CD}$
        break
      end if
    end if
  end for
  return $SF\_Fin_t^{DA}$
end procedure

4.2.3 Domain Adaptation Method for the Irregular Feature (IF_t^{DA}). The irregular feature is the residual component from the raw data minus both the trend and seasonal features. DA-HOC irregular feature is similar to trend feature without Pearson product-moment Correlation Coefficient. We implement the similar domain adaptation method from Section 4.2.1.

$$MMD[\mathcal{F}, IF_S^{DA}, IF_T^{DA}] := \sup_{f \in F} \left(\frac{1}{m} \sum_{i=1}^{m} f(IF_{Si}^{DA}) - \frac{1}{n} \sum_{i=1}^{n} f(IF_{Ti}^{DA}) \right) \tag{8}$$

For DA-HOC irregular feature model, we correlate it like trend feature with weight value due to the possibility of different in scale between the source and the target components.

$$IF_t^{DA} = \beta_0 + \beta_1(IF_t^{DA}) + \beta_2(IF_t^{DA})^2 + ... + \beta_n(IF_t^{DA})^n + \gamma \tag{9}$$

4.2.4 Domain Adaptation Method for the Zero Pattern Adjustment Feature (ZF_t^{DA}).
The SD-HOC ZPA feature is invented to adjust the condition where the room is vacant and minimising false positives. Domain adaptation method for ZPA includes a adapted ZPA border adjustment for the start point and the end point that have been declared in the original model with a new weight mechanism. This weight mechanism needs to be implemented as the new testing data will not be enough to understand the complete structure of the new environment and the original model for ZPA can provide a minimal contribution. The weighting calculation is shown in Equation 10.

$$w = \frac{1}{2}\left[\frac{\sum_{i=1}^{n} ZFStart_t^{CD}}{n} + \frac{\sum_{i=1}^{n} ZFEnd_t^{CD}}{n}\right] \tag{10}$$

w	Weight value
$ZFStart_t^{CD}$	ZPA Starting Point for SD-HOC
$ZFEnd_t^{CD}$	ZPA Ending Point for SD-HOC
n	Total number of dataset
i	Counter value

once the weight value has been obtained, the new $ZFStart_t^{DA}$ and $ZFEnd_t^{DA}$ can be acquired using equation 11.

$$ZF[Start/End]_t^{DA} = w \cdot ZF[Start/End]_t^{CD} \tag{11}$$

$ZF[Start/End]_t^{DA}$	ZPA Starting/End point for DA-HOC
$ZF[Start/End]_t^{CD}$	ZPA Starting/End point for SD-HOC
w	Weight value

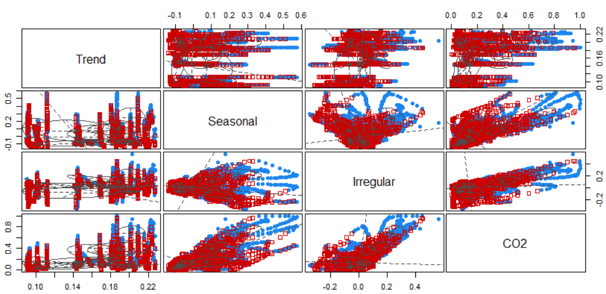

Figure 7: The correlation plot between trend, seasonal, irregular and CO_2 value between vacant (red) and non-vacant room (blue).

4.3 Post Adjustment DA-HOC

Post adjustment phase contains final calibration and model evaluation. The final calibration includes plotting each feature from the previous subsection and correlate each value with their vacant and non-vacant condition using Gaussian Mixture Models (GMM) as shown in Figure 7.

From Figure 7, we can conclude that the correlation between seasonal and irregular factors resulting in clear separation between vacant and non-vacant so we can group the prediction accordingly. If each correlation mix vacant and non-vacant together like the trend and the seasonal factors above, the final calibration step is finished.

For model evaluation, we decided to use two baselines. The first one is SVR and the second one is SD-HOC. From Figure 8, it shows data from both source domain (D_S) and target domain (D_T). Target domain is divided into three different parts. The first one, training data for the baseline (TB) is only used for baselines. The second part, the learning task from target domain (L_T) and the length duration is short. The third part is the prediction target (PT). PT

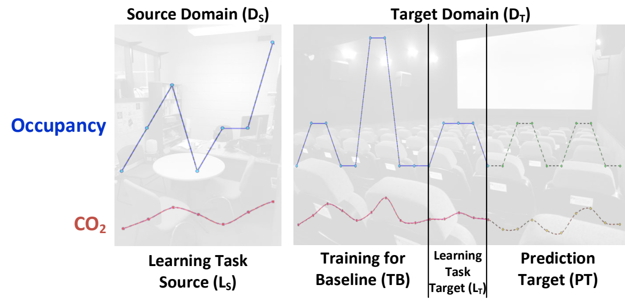

Figure 8: Source Domain and Target Domain.

is the duration where we predict the number of people for every minute.

For each baseline, they are both trained with and without TB as in the real world scenario, it is not easy to have a long duration of training data set. We have SVR(-TB) and SD-HOC(-TB) for baseline that train using L_T dataset only, SVR(+TB) and SD-HOC(+TB) for baseline that train using both TB and L_T dataset and our proposed DA-HOC model with model trained with other domain dataset (L_S) and improved using short duration of L_T.

We used two baselines methods to ensure that the new proposed domain adaptation algorithm performs. Domain adaptation research is hard to compare and it is almost impossible to compare the prediction accuracy based on the numbers only. For this reason, we decide to add another dataset before L_T so the baseline could have a proper training dataset. We called this new dataset as training for baseline (TB). Because of this, it is expected that any baseline with TB will have a better accuracy prediction compared to any algorithm with domain adaptation. The cross-domain analysis will perform worse than same domain analysis. The question that we want to solve is, how worse is the result compared to the same domain algorithm? Furthermore, we subsequently run the same baseline algorithm without TB. If our domain adaptation algorithm can perform better than baseline algorithm without TB result, it provides a promising research direction, as producing a new model from another domain is propitious to increase the prediction accuracy given the lack of data from the target domain.

5 EXPERIMENTS AND RESULTS

For the source domain, we decided to gather ambient data on one academic office belonging to one staff member at RMIT University, Australia. This room is chosen for DA-HOC source domain because a controlled experiment can be conducted for an extended period of data collection.

For the target domain, we utilised a large dataset of volatile organic compounds collected with a mass spectrometry in a cinema theatre from Germany [25]. We extracted and used the CO_2 channel from the dataset for the purpose of our study. This cinema theatre dataset is utilised due to its nature of having fluctuating numbers of people throughout the day. The numbers of people in the audiences can reach up to three hundred and can decrease to zero within two hours.

Our experiment is divided into two stages, binary occupancy prediction and occupancy counting prediction. Binary occupancy predicts whether the room is vacant or occupied. For the single office room, 89.38% of the dataset shows when the room is unoccupied. For cinema theatre dataset, it shows 67.89%. Occupancy counting prediction focuses on predicting the number of people in the room.

Table 1: Detailed statistical information of the datasets for both small and large rooms.

Information	Single office room	Cinema theatre
Min. # of occupants	0 people	0 people
Max. # of occupants	5 people	279 people
Avg. # of occupants	0.16 people	20.69 people
Mode # of occupants	0 people	0 people
Min. CO_2	378 ppm	388.59 ppm
Max. CO_2	1002 ppm	3518.06 ppm
Avg. CO_2	464.82 ppm	690.95ppm
Mode CO_2	453 ppm	424.26 ppm
Number of entrance	1	3
Door condition	Default state is closed	Closed
Windows condition	Closed at all times	Closed at all times
Sensor used	NetAtmo weather station	Mass Spectrometry

5.1 Experiment Settings and Parameters

5.1.1 Source Domain Location. A commercial off-the-shelf Netatmo urban weather station (Range: 0-5000 ppm, accuracy:±50 ppm) was used to collect ambient CO_2 data. The duration of data gathering is two months, from May to June 2015. The dataset is uploaded to a cloud service for integration purposes. Due to the small room characteristic (3x4x3.5m), the time lag is 0 as we assume that there is a negligible period between exhaling process and sensor reading. The time window that used for this dataset is 5-min.

We did manual labelling for the number of people as we were doing controlled experiment. Netatmo urban weather station was chosen because it has a good range of CO_2 sensor. The device was put on a window with the human nose's height because at this level of height CO_2 concentration is changed first. Further detail information about this room dataset in elaborated in Table 1.

5.1.2 Target Domain Location. The cinema dataset was collected between December 2013 and January 2014 [25]. The cinema's maximum capacity is 300 people. Due to the size of the cinema room, the dataset was collected using mass spectrometry machinery installed on the air ventilation system. The air flows from the screening room via the ventilation system to the mass spectrometer for data analysis.

For cinema theatre dataset, we use 5-min time window for data analysis. The cinema theatre capacity is up to 300 people and for this experiment, we run the line of best fit for time lag 0 to time lag 60. The lowest NRMSE is at time lag 32 and we use time lag 32 as time lag baseline. Further detail information about this cinema dataset in elaborated in Table 1.

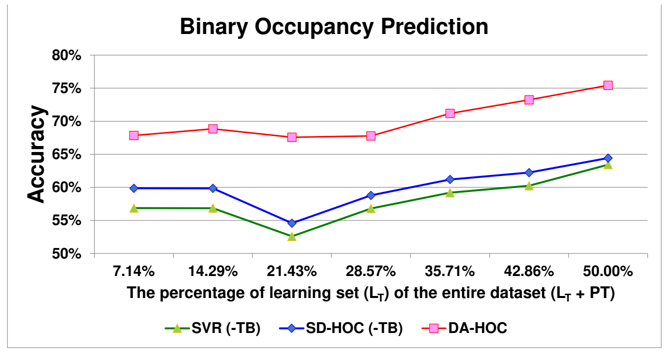

Figure 9: Binary Occupancy Prediction Result.

5.1.3 Experiment Tool. Waikato Environment for Knowledge Analysis (Weka), MATLAB and R were utilised to help us perform this experiment. Weka is used for machine learning algorithms and data analysis. MATLAB is utilised in building some models and R is for data integration, analysis and visualisation. We imported the data from R into Microsoft Excel for data analysis and visual output enhancement.

5.2 Evaluation and Baseline

For the evaluation of time lag, we implement the SD-HOC preprocessing method [3]. The time lag for the cinema theatre is 32 minutes. This value is derived from the size of the room and the location of the mass spectrometer. Time lag means that there is a time

Table 2: SVR (with TB), SD-HOC (with TB), SVR (without TB), SD-HOC (without TB) and DA-HOC Human Binary Occupancy Prediction Accuracy Result.

L_T - PT $\left(\frac{L_T}{L_T+PT}\right)$	SVR (+TB) [$TB + L_T$ ->PT]	SD-HOC (+TB) [$TB + L_T$ ->PT]	SVR (-TB) [L_T ->PT]	SD-HOC (-TB) [L_T ->PT]	DA-HOC [$L_S + L_T$ ->PT]
1 - 13 (7.14%)	81.62%	88.58%	56.85%	59.85%	**67.85%**
2 - 12 (14.29%)	80.93%	88.30%	56.84%	59.84%	**68.84%**
3 - 11 (21.43%)	80.44%	89.51%	52.58%	54.58%	**67.58%**
4 - 10 (28.57%)	84.19%	90.38%	56.79%	58.79%	**67.79%**
5 - 9 (35.71%)	87.52%	91.04%	59.19%	61.19%	**71.19%**
6 - 8 (42.86%)	90.00%	93.89%	60.24%	62.24%	**73.24%**
7 - 7 (50.00%)	91.28%	95.84%	63.42%	64.42%	**75.42%**

Table 3: SVR (with TB), SD-HOC (with TB), SVR (without TB), SD-HOC (without TB) and DA-HOC Human Occupancy Counting Prediction Accuracy Result.

L_T - PT $\left(\frac{L_T}{L_T + PT}\right)$	SVR (+TB) [TB + L_T ->PT]	SD-HOC (+TB)	SVR (-TB) [L_T ->PT]	SD-HOC (-TB)	DA-HOC [L_S + L_T ->PT]
1 - 13 (7.14%)	71.52%	73.85%	50.45%	51.45%	**59.45%**
2 - 12 (14.29%)	70.26%	72.00%	51.84%	52.84%	**59.84%**
3 - 11 (21.43%)	70.10%	72.10%	50.19%	51.19%	**59.19%**
4 - 10 (28.57%)	71.83%	72.84%	51.29%	53.29%	**59.29%**
5 - 9 (35.71%)	73.46%	74.23%	50.02%	51.02%	**62.02%**
6 - 8 (42.86%)	74.54%	76.13%	52.51%	53.51%	**62.51%**
7 - 7 (50.00%)	75.87%	77.18%	53.75%	55.75%	**63.75%**

delay between when the person enters the room, the measurement of the CO_2 and the reaction from our model. CO_2 needs to travel from the occupant's respiration output to the mass spectrometer's sensor.

There are total four baselines that we used for this experiment. SVR algorithm that model is trained with TB dataset and without TB dataset and SD-HOC algorithm that model is trained with TB dataset and without TB dataset. We called them as SVR(-TB), SVR(+TB), SD-HOC(-TB) and SD-HOC(+TB). We run each experiment for both vacant prediction and occupancy counting. Vacant prediction means that each algorithm only predicts whether the room is empty or occupied. Occupancy counting means that each algorithm needs to predict the exact number of occupants for each PT time frame. Vacant prediction accuracy should be higher than occupancy counting prediction accuracy. We decide to run the experiment multiple times for a different number of days for both L_T and PT. The longer the duration of L_T results in a higher accuracy of each algorithm.

If the prediction value is within ±5 people from the ground truth, we count this as true positive. We calculate the prediction accuracy as the number of true positive values divided by the number of PT records in the dataset.

5.3 Experiment Result

Starting from day one in the learning task target data (L_T), we split the data into two sections. We used the L_T data from day 1 to predict the next 13 days (PT) and compared this baseline result with our DA-HOC method. This result is line one of Table 2. With L_T data from 2 days predicted 12 days (PT) and compared this in line two. This step is repeated until learning and prediction data have 50-50 split.

For Figure 9 and 10, we calculate $\frac{L_T}{L_T + PT}$ and present this as a percentage to visualise the comparison of binary occupancy and occupancy counting prediction for the varying dataset predictions from $\frac{1}{1+13}$ (equals to 7.14%) to $\frac{7}{7+7}$ which corresponds to 50% split between learning task target and prediction target data.

The binary occupancy prediction accuracy results are shown in Table 2 and Figure 9. The accuracy of SD-HOC(+TB) ranges from 88.58% - 95.84% (see Table 2 column 2) and is the highest compared to other algorithms. The accuracy of SVR(+TB) is the second best (see Table 2 column 1). DA-HOC is the most accurate algorithm (compare Table 2 column 5 with column 3 and 4) if the TB dataset is not available. DA-HOC's performance is better by 12.29% in comparison to SVR and 10.14% in comparison to SD-HOC. Overall, SD-HOC is more accurate than SVR and this result agrees with previous research [3]. The accuracy increases for the larger learning dataset, higher L_T.

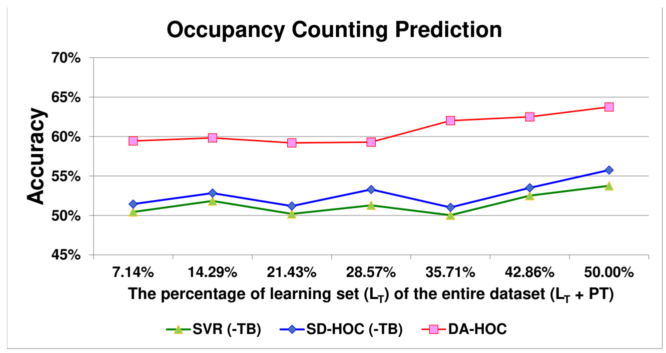

Figure 10: Occupancy Counting Accuracy Result.

The occupancy counting prediction accuracy results are shown in Table 3 and Figure 10. Any algorithms with TB as part of their training dataset result in higher accuracy. Without using TB dataset, DA-HOC has the highest accuracy averaging from 59.45% - 63.75% with ±5 people error tolerance (see Table 3 column 5). Even though the accuracy seems low, predicting the number of the occupant from 0 to 300 with more than 60% accuracy is acceptable.

There are two periods for the empty room prediction in the cinema theatre. The first one is from midnight to the morning when the cinema theatre is closed. Our ZPA method covers this interval. The second timespan is between cinema sessions where the previous screening audience leaves the cinema before the next screen audience enters. Some of our predictions returned negative occupancy values (up to -19 persons). Due to this inaccuracy, we forced our model to convert each negative prediction into zero occupancy. The binary occupancy accuracy prediction is higher than the occupancy counting accuracy prediction. This result is expected as predicting whether the room is vacant or not is easier than predicting the exact number of people inside. The significance of this research is any large room utilisation prediction can be conducted with a negligible amount of training data.

6 CONCLUSION, LIMITATION AND FUTURE WORK

Human occupancy counting research is useful in many areas, such as space and room utilisation, energy consumption reduction, human comfort and security. In some cases, having a proper training dataset to build a robust human occupancy counting model could not be obtained. DA-HOC is the latest state-of-the-art domain adaptation technique to predict indoor human occupancy with a minimum amount of training data set and leverage the prediction model knowledge from other content to be used as the baseline model. Domain adaptation allows us to transfer the classification knowledge into a new domain.

DA-HOC's binary prediction accuracy up to 75.34% with only minimum training data is encouraging. Even though DA-HOC performs on average 20% less accurate for binary prediction and 15% for occupancy counting, DA-HOC is the best model when we could not secure the right amount of historical data for the target domain. DA-HOC is also the optimal option for making occupancy predictions for a newly constructed building, as the building will have no historical data to draw on.

The limitation of this research is that we have tested the model in one domain. Hence, applying this to multiple locations requires an adaptive approach. The algorithm needs to be recalibrated based on the size of the room for the new target domain during model preprocessing. Another limitation is our occupancy label for the cinema came from the number of sold tickets. If there are customers who purchased tickets, but did not attend the movie, our occupancy label is not 100% accurate. However, we assume this condition to only occur within the error boundary of 2-3%.

Domain adaptation and transfer learning research are still in their infancy and more research is focusing on this area. For future work, we plan to increase DA-HOC prediction accuracy and implement DA-HOC for real-time data sets. Furthermore, the DA-HOC model can also be applied not only to human occupancy, but also other domains that utilise domain adaptation analysis. Other future work includes incremental learning of DA-HOC with unsupervised domain adaptation without any labelling. The unsupervised technique reduces the cost of gathering the labels and finding the ground-truth. Progressive learning models can learn and will gradually improve their prediction accuracy over time.

ACKNOWLEDGMENTS

The authors would like to thank Joerg Wicker from University of Mainz for providing the cinema dataset used in this paper. This research is supported by the Australian Government Research Training Program Scholarship and two RMIT and Siemens Sustainable Urban Precinct Project (SUPP) grants: "iCo2mmunity: Personal and Community Monitoring for University-wide Engagement towards Greener, Healthier, and more Productive Living" and "The Greener Office and Classroom".

REFERENCES

[1] Hisham Al-Mubaid and Syed A Umair. 2006. A new text categorization technique using distributional clustering and learning logic. *IEEE Transactions on Knowledge and Data Engineering* 18, 9 (2006), 1156–1165.

[2] Irvan Bastian Arief-Ang, Flora Dilys Salim, and Margaret Hamilton. 2016. Human occupancy recognition with multivariate ambient sensors. In *Pervasive Computing and Communication Workshops (PerCom Workshops), 2016 IEEE International Conference on.* IEEE, 1–6.

[3] Irvan Bastian Arief-Ang, Flora Dilys Salim, and Margaret Hamilton. 2017. SD-HOC: Seasonal Decomposition Algorithm for Mining Lagged Time Series. In *the Fifteenth Australasian Data Mining Conference, Melbourne, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 172.*

[4] Chandrayee Basu, Christian Koehler, Kamalika Das, and Anind K Dey. 2015. Perccs: person-count from carbon dioxide using sparse non-negative matrix factorization. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing.* ACM, 987–998.

[5] John Blitzer, Mark Dredze, Fernando Pereira, et al. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, Vol. 7. 440–447.

[6] Corinna Cortes and Mehryar Mohri. 2011. Domain adaptation in regression. In *International Conference on Algorithmic Learning Theory.* Springer, 308–323.

[7] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. 2007. Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning.* ACM, 193–200.

[8] Hal Daume III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research* 26 (2006), 101–126.

[9] Saandeep Depatla, Arjun Muralidharan, and Yasamin Mostofi. 2015. Occupancy estimation using only wifi power measurements. *IEEE Journal on Selected Areas in Communications* 33, 7 (2015), 1381–1393.

[10] U. DOE. 2010. *Building Energy Databook.* Technical Report.

[11] Prabal K Dutta, Anish K Arora, and Steven B Bibyk. 2006. Towards radar-enabled sensor networks. In *Information Processing in Sensor Networks, 2006. IPSN 2006. The Fifth International Conference on.* IEEE, 467–474.

[12] Nazli Farajidavar, Teofilo de Campos, and Josef Kittler. 2014. Transductive transfer machine. In *Asian Conference on Computer Vision.* Springer, 623–639.

[13] E. Fortet and E. Mourier. 1953. Convergence de la re'paration empirique vers la re'paration the'orique. *Ann. Scient. E'cole Norm. Sup.* 70 (1953), 266–285.

[14] Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Hongjun Lu, and Philip S Yu. 2006. Text classification without negative examples revisit. *IEEE transactions on Knowledge and Data Engineering* 18, 1 (2006), 6–20.

[15] Sozo Inoue and Xincheng Pan. 2016. Supervised and Unsupervised Transfer Learning for Activity Recognition from Simple In-home Sensors. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services.* ACM, 20–27.

[16] David Kale, Marjan Ghazvininejad, Anil Ramakrishna, Jingrui He, and Yan Liu. 2015. Hierarchical active transfer learning. In *Proceedings of the 2015 SIAM International Conference on Data Mining.* SIAM, 514–522.

[17] Brian Kulis, Kate Saenko, and Trevor Darrell. 2011. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on.* IEEE, 1785–1792.

[18] David Lopez-Paz, Jose M Hernández-lobato, and Bernhard Schölkopf. 2012. Semi-supervised domain adaptation with non-parametric copulas. In *Advances in neural information processing systems.* 665–673.

[19] Weining Lu, Bin Liang, Yu Cheng, Deshan Meng, Jun Yang, and Tao Zhang. 2017. Deep model based domain adaptation for fault diagnosis. *IEEE Transactions on Industrial Electronics* 64, 3 (2017), 2296–2305.

[20] Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2010), 1345–1359.

[21] Sinno Jialin Pan, Vincent Wenchen Zheng, Qiang Yang, and Derek Hao Hu. 2008. Transfer learning for wifi-based indoor localization. In *Association for the advancement of artificial intelligence (AAAI) workshop.* 6.

[22] Kanoksri Sarinnapakorn and Miroslav Kubat. 2007. Combining subclassifiers in text categorization: A dst-based solution and a case study. *IEEE Transactions on Knowledge and Data Engineering* 19, 12 (2007), 1638–1651.

[23] Stephan Spiegel. 2016. Transfer Learning for Time Series Classification in Dissimilarity Spaces. In *Proceedings of AALTD 2016: Second ECML/PKDD International Workshop on Advanced Analytics and Learning on Temporal Data.* 78.

[24] Matthew E Taylor and Peter Stone. 2009. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research* 10, Jul (2009), 1633–1685.

[25] Jörg Wicker, Nicolas Krauter, Bettina Derstorff, Christof Stönner, Efstratios Bourtsoukidis, Thomas Klüpfel, Jonathan Williams, and Stefan Kramer. 2015. Cinema Data Mining: The Smell of Fear. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 1295–1304.

[26] Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. 2008. Top 10 algorithms in data mining. *Knowledge and information systems* 14, 1 (2008), 1–37.

[27] Ke Yan, Lu Kou, and David Zhang. 2016. Domain Adaptation via Maximum Independence of Domain Features. *arXiv preprint arXiv:1603.04535* (2016).

[28] Qiang Yang and Xindong Wu. 2006. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making* 5, 04 (2006), 597–604.