PerCCS: Person-Count from Carbon dioxide using Sparse Non-negative Matrix Factorization

Chandrayee Basu
University of California at Merced
Merced, CA, USA
basu.chandrayee@gmail.com

Christian Koehler
Carnegie Mellon University
Pittsburgh, PA, USA
christiankoehler81@icloud.com

Kamalika Das
UARC, NASA Ames Research Center
Moffett Field, CA, USA
kamalika.das@nasa.gov

Anind Dey
Carnegie Mellon University
Pittsburgh, PA, USA
anind@cs.cmu.edu

ABSTRACT
Occupancy count in rooms is valuable for applications such as room utilization, opportunistic meeting support, emergency responses in buildings and efficient heating and cooling operations. Few buildings, however, have the means of knowing occupancy beyond simple binary presence-absence. In this paper we present the *PerCCS* algorithm that explores the possibility of estimating person count from CO_2 sensors already integrated in everyday room air-conditioning infrastructure. *PerCSS* uses task-driven Sparse Non-negative Matrix Factorization (SNMF) to learn a non-negative low-dimensional representation of the CO_2 data in the preprocessing stage. This denoised CO_2 acts as the predictor variable for estimating occupancy count using Ensemble Least Square Regression. We tested the algorithm to estimate 15 minutes average occupancy count from a classroom of capacity 42 and compared its performance against existing methods from the literature. *PerCSS* estimates occupancy with a normalized mean squared error (NMSE) of 0.075 and outperformed our comparative methods in predicting occupancy count with 91 % and 15 % for exact occupancy estimation, when the room was unoccupied and occupied respectively, whereas the competing methods failed completely.

Author Keywords
Building energy efficiency; Machine Learning.

ACM Classification Keywords
H.1.m. Information Systems: Miscellaneous

INTRODUCTION

UbiComp'15 Adjunct, September 7–11, 2015, Osaka, Japan.

DOI string from ACM form confirmation

Knowledge of occupancy count can be leveraged in several smart building applications including smart meeting space allocation, emergency evacuation, opportunistic meetings, assessment of collaboration-space usage and energy efficient and comfortable heating-cooling operation. Smart meeting space allocation systems attempt to predict how best to allocate a finite set of meeting rooms to match desired room requests. Many of these systems attempt to solve this predictive resource allocation problem using energy-related cost functions [22]. In simulations, Majumdar *et al.* found that, of several factors, matching the room capacity with the desired meeting size resulted in the lowest daily energy use from all the meeting rooms.

In addition to predictive meeting allocation, it may also be important to track attendance at a meeting or seminar to determine whether they should be canceled due to low attendance or moved to bigger rooms due to higher than expected attendance. Under such circumstances, obtaining an occupancy count using only an infrastructure-integrated system can minimize human labor and be used to automatically update room schedules. Currently, rooms are scheduled for meetings and classes assuming attendance of all registered participants, resulting in inefficient space scheduling and usage.

Another use of occupancy count is for opportunistic meetings. Opportunistic meetings are a form of informal communication mediated by close physical proximity of workers. Such meetings are important in R&D organizations for initiating collaborations [18]. Horvitz *et al.* [12] proposed COORDINATE, a complex sensor fusion system that fostered such meetings by informing coworkers of each others presence and availability. COORDINATE could forecast the presence, availability and meeting status of a person from calendar and status of appointment features. Occupancy count can enhance the functionality of COORDINATE-like systems by providing information about whether a user is alone in her room or has visitors and can also validate system forecasts.

Building maintenance is a major contributor to society's overall energy consumption [1]. The age of a building, the building's envelope, the kind of HVAC system that is installed in the building, and organizational choices such as supporting a 24/7 workplace all contribute to the overall cost of building maintenance. Traditional temperature control systems use estimates of when a building is occupied to define a static temperature control schedule. Knowing the actual room usage and occupancy pattern can significantly influence how efficiently a control system can balance energy savings considerations and thermal comfort [28, 16, 3].

Existing HVAC systems are controlled to support maximum occupancy. For example, the ventilation airflow rate is maintained as a linear function of maximum occupancy. In variable air volume systems, the cooling airflow rate is maintained at 30% of the maximum ventilation airflow rate [33]. A temperature setpoint-based cooling control therefore means higher volumes of cool air blasted in larger rooms than rooms of smaller capacity. It works well for offices, which usually have less than five occupants, but for large shared spaces (*e.g.*, conference rooms or classrooms) without a designated owner, the actual occupancy can largely vary depending on the current use. For sparsely occupied rooms, this blasting of cool air can lead to local discomfort. Recent research has shown that knowledge of occupancy count alone can save 42% of annual air conditioning energy in commercial buildings [7] by controlling the minimum ventilation rate based on actual occupancy rather than the maximum designed occupancy. Another simulation study found that in large commercial buildings, precise occupancy count-based cooling and heating control can save 2 to 3 times more energy than binary occupancy-based control for particular climate zones [33]. The projected savings on a national scale is 17.8%.

Despite the obvious benefits of having a reliable occupancy count, very few building systems have a measure of occupancy beyond simple binary presence-absence. However, much research has been dedicated to occupancy estimation, most of which use multi-sensor fusion [23, 19, 20] and/or intrusive instrumentation like imaging [14, 32]. Researchers have used multi-sensor fusion to infer small occupancy counts (20 or less). These sensors measure temperature, humidity, capacitance, sound, electrical interference and even water, electrical load, ventilation air flow rate and pressure. Other systems are tracking based and use Bluetooth or WiFi beacons, door-way crossing sensors [26] and thermal array sensors. All of these above approaches of occupancy sensing call for additional infrastructure. Moreover, none of these methods have been employed to estimate large occupancy counts in classrooms or lecture halls.

Occupancy count has also been treated as a prediction problem, where future occupancy is inferred from existing occupancy data. Most applications of predictive methods have been limited to binary occupancy and individual user tracking scenarios [28, 16]. Other methods require building related information such as floor plans, thermal properties of building materials and operational characteristics of the ventilation systems such as ventilation air flow rate and CO_2 concentration in the supply air.

In our work, we estimate occupancy count using only a single sensor, a CO_2 sensor. CO_2 sensors monitor only a single environmental variable, CO_2 concentration measured at one point in the room. They are an integral part of the room air-conditioning infrastructure in most buildings and thus obviate the need for additional infrastructure. Unlike most prior research conducted on office spaces with small and relatively static occupancy, our research is focused on spaces of high and variable occupancy like classrooms and conference rooms.

We investigate a new algorithm for performing occupancy estimation, which we call ***PerCCS****: Person Count from CO_2.* using Sparse non-negative matrix factorization (SNMF). SNMF is a task-driven approach for denoising the CO_2 data. This processed CO_2 data is then used as a predictor in an ensemble least square regression for occupancy estimation. The method is referred to as *task-driven* because the denoising of the predictor is performed iteratively so as to minimize the above regression error. Denoising is required due to the fact that CO_2 represents an aggregate of multiple generating factors, of which occupancy count is just one and hence CO_2 data has higher fluctuations than occupancy count. This SNMF denoising method has not been explored in the domain of smart spaces to the best of our knowledge.

We compare the performance of our method with algorithms from the literature that have achieved the highest accuracy in estimating occupancy count from only CO_2 concentration or in combination with other sensor data such as sound and humidity. Note that different units have been used to define the accuracy of occupancy estimation. In the binary occupancy detection case, accuracy refers to number of correct detections [19]. For a regression problem such as ours, we report accuracy in terms of the Normalized Mean Square Error (NMSE) and mean ℓ_1 error.

We also consider a baseline scenario where we use raw CO_2 data without denoising. In order to evaluate the performance of our method, we collected ground truth occupancy counts and CO_2 data from a classroom with a capacity of 42. Using 13 days of data, we demonstrate that our method is able to estimate zero occupancy during 91% of unoccupied periods and exactly track the occupancy during 15% of the occupied periods, unlike the baseline method and SVR which could not predict exact occupancy at all. On average, *PerCCS* overestimates the occupancy count by 1.0 person compared to the baseline error of 1.3 and SVR error of 2.54 persons. Moreover *PerCCS* has an overall NMSE of 0.075, while the baseline and SVR have NMSEs of 0.16 and 0.51, respectively. We conclude that *PerCCS* is a significant improvement over these existing methods.

The benefits of our work are two-fold. In addition to our novel approach that outperforms existing approaches, our work extends past work by leveraging *only a single sensor* that is already available in most buildings, and predicts occupancy in *large rooms with variable occupancy*.

RELATED WORK

The goal of our work is to accurately estimate occupancy count from carbon dioxide measurements. We do so by applying SNMF based regression to this new problem domain. We divide our discussion of the related literature into two categories: i) environmental sensing that supports the estimation of occupancy count and ii) algorithms from the literature used in occupancy counting from environmental sensors.

Occupancy estimation using multi-sensor

The simplest and the most ubiquitous occupancy sensor is Passive Infrared (PIR) which provides binary occupancy information. Erickson *et al.* showed that while some energy savings can be achieved with binary information, approximate occupancy count provides higher potential for building controls and energy savings. Several methods have been proposed for detecting, counting, tracking, and identifying people inside buildings using a combination of different sensing modalities and machine learning approaches [30].

Many of these methods are either privacy-intrusive or require additional instrumentation that can be costly or difficult to deploy. Others use an indirect means of counting people like tracking of devices that people carry [3]. Even though they allow for a very precise occupant count, they do not scale to large buildings since both the rooms as well as the occupants need to be instrumented. Modern technology such as iBeacons and WiFi access-points might make this concept more feasible, but still require the distribution of additional sensors in each room. Furthermore they, along with smartphone localization, have the larger privacy drawback.

In this paper we investigate passive, room-level counting of people by using ubiquitous CO_2 monitors. CO_2 is a measurable physical quantity that has a direct relationship to human occupancy. Teixeira *et al.* [29] refer to these physical quantities as static intrinsic traits. Static intrinsic traits are directly affected by human presence and do not require any additional activity or input or addtional device to be carried by user beyond mere presence (unlike, for example, motion detectors or accelerometers), thus respecting occupants' privacy.

Many such intrinsic traits like temperature, capacitance, carbon dioxide, humidity, even water usage and electric loads have been used by researchers to estimate the occupancy count in a room. Even though these methods have shown promise, some of them are not practical in the long run. For example, temperature sensors will require a high fidelity model of the buildings thermal capacity and will have interference from heat generated by several other unrelated sources in the indoor environment like electrical appliances. Alternative forms of thermal sensors, like thermopile sensor arrays require appropriate installation and significant calibration [10].

Shape and weight also fall under the category of static intrinsic traits, both of which are good for tracking and identification, but they require additional instrumentation such as cameras or pressure sensitive surfaces. Cameras require sophisticated image processing that is both computationally costly and introduces multiple privacy concerns. Weight sensors, on the other hand, might not differentiate between people leaving and entering a room and thus compute a wrong occupancy count, unless at least a pair of sensors are deployed similar to doorway monitoring, but this requires additional infrastructure.

Erickson *et al.* have contributed largely to the literature on occupancy estimation, specifically in the testing of several of occupancy sensing technologies as occupants move through spaces. The technologies range from single passive infrared sensors [14, 10, 8, 13, 32, 6] to more sophisticated thermopile and camera arrays [4, 9].

In [24], occupancy count estimation is performed by making an assumption that the CO_2 generation rate per person is a constant of 0.01 SCFM (cfm standardized by temperature and pressure condition of the surrounding gas). The method also relies on measurements of CO_2 concentrations in both the space and the supply air as well as the supply air flow rate. The authors used steady state and transient equations for predicting occupancy from the above measurements and reported an accuracy of estimation within 2 occupants during walk-through counts in a room of capacity 25. The author noted that transient CO_2 measurements resulted in impractical occupancy count that needed to be "damped" by restricting the allowed change to 1 person/15 seconds. The motivation of the paper was to present the usefulness of real-time occupancy information on outdoor airflow control without sufficient elaboration on the results that can be leveraged for comparison. The aforementioned work is the only one in the smart space literature that has tried to predict occupancy from CO_2 data alone, however, it required measurements at several points beyond a single room sensor.

We selected carbon dioxide based occupancy counting for several reasons. CO_2 sensors are an integral part of infrastructure for demand-controlled ventilation. Furthermore CO_2 based occupancy counting is not prone to accumulated counting errors, as are sensors that detect state transitions between occupied and unoccupied. While some researchers argue against the slow build up time of carbon dioxide due to ventilation, a preliminary feasibility study that we conducted found that depending on the room ventilation air flow rate, the CO_2 build up time varies between 10 - 20 minutes. Hence a CO_2 based occupancy estimation system may have a reaction time of at most 20 minutes.

Algorithms for Occupancy estimation

A number of research papers have discussed the use of CO_2 sensing for occupancy detection, sometimes restricting to the problem of binary presence-absence detection. The majority of existing research has used multi-modal environmental sensing for occupancy count estimation. To date the highest prediction accuracy, estimated as the percentage of correct predictions, has been reported by Lam *et al.* [19]. Using information gain for feature selection, the authors found that CO_2, the second order change of CO_2, the difference between indoor and outdoor CO_2 levels, and the 20 minutes moving average of CO_2 are the most relevant features for explaining occupancy variation. Besides CO_2, sound data was found to bear a strong correlation with occupancy count compared to other environmental data. In this study, the highest prediction accuracy of 75 % was achieved using Hidden Markov Models, followed by Support Vector Regression at 70 %. The authors argued that HMMs can leverage the latent correlation between current and past occupancy and CO_2 dynamics for prediction. However, the maximum occupancy of the test bed at any given time was only *four* with occasional fluctuations which HMMs failed to address.

Support Vector Machines (SVMs) address the non-linearity in the relationship between CO_2 and occupancy count and also account for variations in the room occupancy patterns. SVMs can also accommodate discrete variables like count. Erickson *et al.* [10] modeled room occupancy as a multi-variate Gaussian distribution even though occupancy count is a discrete data. While [10] used only the temporal pattern of occupancy in this model, [7] implemented three Markov Chain models to leverage the spatio-temporal correlation in occupancy pattern. Their models use inter-room occupancy correlation from image data to predict future occupancy count with 80% accuracy, and a false positive rate of 13%. The rationale for modeling inter-room occupancy is justified by the use of image sensors that record the movement of a person from one room to another, e.g., transition between two adjoining spaces. The number of states exponentially grows with large room sizes and more interconnected rooms, even after ignoring repeated state transitions. Furthermore, the amount of data required to ensure reasonable prediction accuracy is much higher in such state space models compared to regression. Since the study closest to our approach obtained the best results using HMMs and SVMs [19], we decided to compare the performance of our algorithm to these methods.

BACKGROUND

Understanding the problem space

CO_2 exhaled by a person is an intrinsic trait and varies from person to person. Equations 1 and 2 are empirical models of CO_2 exhaled by a person and partially explains why CO_2 concentration (measured in ppm) in room air does not linearly grow with the number of people in the room.

$$V_{CO_2} = \frac{0.00276 A_D.M.RQ}{0.23RQ + 0.77} \tag{1}$$

$$A_D = 0.203 H^{0.725} W^{0.425} \tag{2}$$

A_D is the DuBois surface area of a person in m^2, M is the metabolic activity equivalent or MET value: 1.25 for students in school. H is the height of a person in meters and weight is body mass in kilograms, RQ is the respiratory quotient typically taken as 0.83. A_D is variable from person to person. One of the co-authors body surface area is 1.489 m^2, while that of an average American of age 20-30 is 1.8 m^2. The corresponding V_{CO2} are 0.0042 and 0.0052 L/s respectively. For a room with a 850 cfm (cubic feet per minute) ventilation rate, the above difference in the A_D means a 3 ppm difference in steady state CO_2 concentration, which may be less than the sensor error. For a smaller room, however, the above surface area difference may amount to a concentration difference as high as 100 ppm.

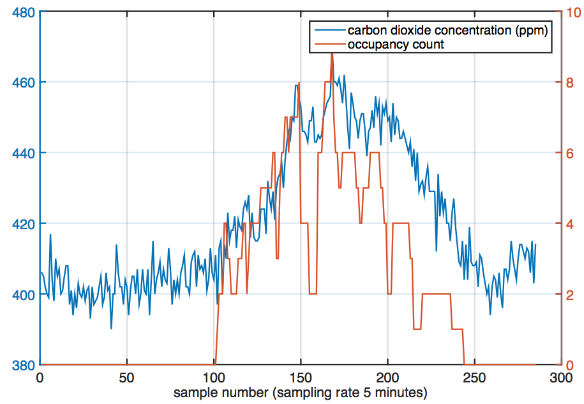

Figure 1. Distribution of occupant count and CO_2 for one day in a laboratory with a capacity of 15 people

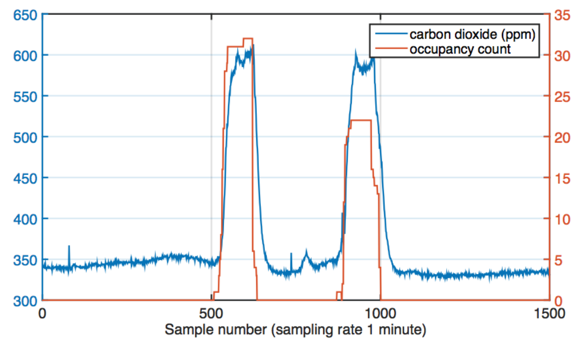

Figure 2. Distribution of occupant count and CO_2 for one day in a classroom with a capacity of 42 people

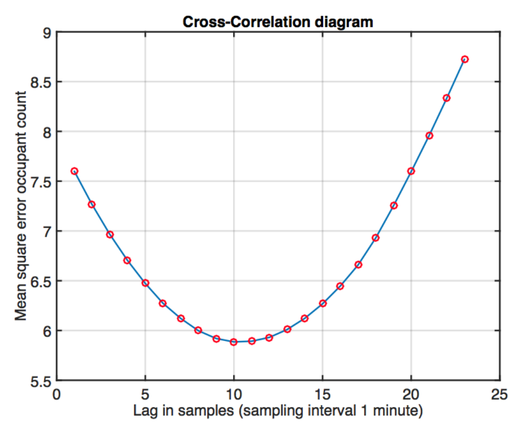

Figure 3. Ordinary Least Square Regression Mean Squared Error between predicted and actual occupancy minimum at 10 minutes lagged predictor data

Feasibility study

In order to ascertain the feasibility of inferring occupancy count from CO_2 we collected occupancy count and CO_2 data in a lab space of capacity 15 and a classroom of capacity 42. For our feasibility study, we selected two rooms to understand the difference in CO_2 dynamics from room to room. We installed our own CO_2 sensor in the lab which is in an older part of the building and does not have a CO_2 sensor. This CO_2 sensor has a maximum sampling rate of 5 minutes. The 42-person capacity classroom we used also serves as our final test bed, and its ventilation system came already equipped with a CO_2 sensor that samples every minute. Figure 1 shows the distribution of occupancy and CO_2 in a lab with a capacity of 15 people. The two variables are well correlated as the lab is located in a basement with constant ventilation air flow rate and no external facade. However in larger spaces like classrooms (capacity>40) open windows and doors facing corridors with glass linings can impact such correlations (see Figure 2). Moreover CO_2 exhaled by people slowly reaches a static and uniform concentration in the room air, so that the full change must be captured by the sensor located high up on the room wall, at different distances from the sources. The above dynamics depend on the room volume, amount of air mixing and density of occupants in the room.

Agarwal *et al.* [2] mentioned that while occupancy may be inferred from CO_2, the slow evolution time of CO_2 as a result of occupancy may prevent real-time occupancy detection. Moreover, Lam *et al.* observed that a 20 minutes moving average of CO_2 is one of the most informative features for occupancy inference [19].

In order to understand the CO_2 dynamics in room air as a function of occupancy, we performed i) a cross-correlation analysis between CO_2 and occupant count, and ii) a linear regression between CO_2 and differentially lagged instantaneous counts. The measurements were taken in both the rooms of capacity 15 and 42. Our preliminary feasibility study shows that depending on the room ventilation air flow rate and the room capacity, the CO_2 build up time varies. For example, in the 15-person capacity lab, the build up time is 20 minutes, whereas in the 42-person capacity large classroom, the build up time is 10 minutes (see Figure 3). In Figure 3 the x axis is the chosen lag in time steps applied to occupancy data and the y axis is the mean squared error of linear regression between lagged occupancy of the classroom and CO_2 data. The NMSE is lowest at 10 time steps lagged occupancy, which is 10 minutes at a 1 minute sampling rate.

CO_2 concentration as a predictor of occupancy is both noisy and redundant since it represents an aggregate of multiple generating factors of which room occupancy count is just one. The temporal pattern of occupancy in our testbed is fairly regular over the course of the day, say at an hourly resolution, however quite irregular at a time scale of 1 minute, especially during the time when many people enter or leave the space. Such rapid changes in occupancy cannot be captured using CO_2 data, because of CO_2's slow dynamics.

In this paper, we focus on occupancy estimation at 15 minutes resolution, sufficient or practical for energy management purposes. This is because any change in supply air to the room in response to occupancy change will manifest in the room air slowly. For example in case of temperature control in residential buildings, Koehler *et al.* [16] found that homes in their study took an average of 59 minutes to heat up to the set point temperature from the time the heating was initiated. In our test bed building we observed that any temperature set point change triggered by a change in binary occupancy takes approximately 5 minutes per degree change to take effect. In other words controlling air supply volume flow based on occupancy changes faster than 15 minutes will not be effective. However, once the occupancy count has reached a steady state, our method should be able to initiate reactive air conditioning based on this steady state value.

We also examine the ability of our algorithm to predict zero occupancies. This is because minimizing false positives is of paramount importance from an energy efficiency standpoint. The resultant problem of predicting occupancy from CO_2, therefore, has the following characteristics: *Non-linear relationship*, *Time lag*, *Noisy predictor data with some redundancy* and *Target variable has several zero values*.

Our problem can be cast as one of source separation, where CO_2 is an aggregate of multiple generating factors including room occupant count, ventilation airflow rate, outdoor CO_2 concentration (if the room has infiltration and/or natural ventilation) and contributions of plants. We assume that each of the these contributing factors have additive effects on the CO_2 development in a space. We, therefore chose to denoise the CO_2 data prior to regression analysis.

Regression with denoising and disaggregation

Non-negative matrix factorization (NMF) is an approach for source-separation applied to problems such as speaker identification and energy disaggregation. In NMF a source signal is expressed as a linear combination of several source components, like music composed by the addition of notes. The composing sources are together called a dictionary. When the number of composing signals are more than the data dimensionality, the dictionary is said to be over-complete. The composing sources may not be present in all parts of the composite signal. In other words the weights of the linear models may sometimes be zero. The weight matrix is sparse in that case. This is particularly true for over-complete dictionaries.

The NMF method that allows us to learn the sparse weights is called Sparse Non-negative matrix factorization (SNMF). Sparsity can also be applied when we perform dimensionality reduction. Here the number of constituent signals in the dictionary is lesser than the data dimension, but sparsity of the weights may still be required for accurate representation of the composite signal. For example, in our study, a classroom may have several unoccupied periods, where occupancy does not contribute to room CO_2 concentration. Originally SNMF was used by Schmidt and Ollson (2006) [27] for speech separation of multiple speakers. The authors experimented with the sparsity coefficient and the number of dictionary elements and concluded that a larger number of dictionary elements improves speech separation performance.

Kolter *et al.* (2011) [17] applied the similar principle of single channel source separation to the energy disaggregation problem: identifying which devices in a building are consuming energy. Dictionaries were trained on individual energy end uses, and the weights of the learnt dictionaries were updated. The authors implemented an SNMF formulation of the dictionary learning problem. For cases of disaggregation, where individual components of energy consumption cannot be measured, but factors affecting those components (contexts) are available, Wytock and Kolter (2014) [31] proposed a contextually supervised source separation method. In order to account for arbitrary delays between the actual energy components and their contexts, they used a sliding window. This problem is similar to ours, except that instead of predicting the components responsible for energy use, we want to predict one of the contexts for CO_2, namely occupancy.

A related approach is Task-driven dictionary learning, proposed by Mairal *et al.* (2012). This dictionary learning approach accommodates other objective functions of dictionary learning besides minimizing reconstruction error. The aforementioned problem of source separation is an example of discriminative dictionary learning, where the objective is to maximize separation while reducing the aggregate reconstruction error. Another case of a task-driven dictionary learning is where least square regression is used for signal recovery from a noisy measurement of the signal [21].

In matrix factorization based source separation problems, denoising and separation can be improved with knowledge of the noise structure. For example, in the energy disaggregation problem, if the goal is to identify the contribution of a kitchen appliance alone, the contribution of the rest of the devices could be considered as a structured noise. [17] implemented a method to verify that the learnt basis functions, when updated, do actually represent the original sources. While this approach is promising, for our setting, we only know one of the composing signals of CO_2, occupancy, which we are trying to learn. Hence we cannot make use of the noise structure in our algorithm. However we select the number of the composing signals, more formally the dictionary size and the coefficient of sparsity based on domain knowledge in room heating and air conditioning.

Building on the past work in source-separation for composite signals, we use SNMF to obtain a low-dimensional representation of CO_2 and use the resultant, presumably less redundant and less noisy data as a predictor for occupant count. We now describe our algorithm in detail.

PERCCS ALGORITHM

Sparse Non-negative Matrix Factorization with Regularization

The task in this paper is to obtain an occupant count Y as a function of the lower dimensional representation of CO_2 data X. A task-driven approach ensures that the learnt dictionary minimizes the regression error between Y and the representation of X. Let X be a real-valued matrix such that $X \in \mathbb{R}^{n\times m}$. In NMF, we express X as a product of two matrices $W \in \mathbb{R}^{n\times k}$ and $H \in \mathbb{R}^{k\times m}$, where $k < min(m, n)$. Therefore, $X \approx WH$ is a matrix factorization problem, that reduces the dimension of the input matrix X. One of the factors is called the weight, W, and the other is the basis matrix or the dictionary, H. NMF focuses on boundaries of the distribution of the data. The NMF representation characterizes all data as lying within a compact convex region. The goal is to select bases H so as to maximize the compactness of the solution and also to enclose as much data as possible within the bases. An approximation error results from an inability of the solution bases to enclose all the data points in the matrix X. The bases and the associated weights are learnt iteratively to minimize the following objective function.

$$min_{W,H}\frac{1}{2}\|X - WH\|_F^2 s.t. W \geq 0, H \geq 0 \tag{3}$$

The above objective function is the ℓ_2 divergence. KL divergence $D(X||WH)$ is an information theoretic error that accounts for the length of the data vectors when computing the divergences, and hence is a more reliable and popular cost function for NMF. Several other algorithms have been proposed for NMF [15], some of which include Block Principal Pivoting [15], Hierarchical Alternating Least Squares [5] and more recently block coordinate descent [25]. Unlike other projection methods like PCA, NMF can generate sparse results due to the strict positivity of the bases and the weights. However, bases fewer than the data dimension can lead to less accurate solutions. This drawback can be overcome by using an over-complete dictionary H ($k > m$) and applying sparsity constraints to the weights matrix W. Sparsity is imposed in the form of the ℓ_1 norm, which is the sum of absolute values of the weights. Sometimes regularization, as in the case of elastic net formulation [34], can prevent the weights from assuming arbitrarily large values by constraining the ℓ_2 norm of W. The resulting objective function for sparse NMF is given by equation 4 with positivity constraints on W and H.

$$\frac{1}{2}\|X - WH\|_F^2 + \alpha\|W\|_2^2 + \beta\|W\|_1 \tag{4}$$

α and β are the regularization coefficients and the sparsity coefficient respectively. We used a Block Principal Pivoting (BPP) solver for Sparse and regularized NMF to learn the dictionary and weights from raw CO_2 data. We also tested multiplicative updates using both KL divergence and ℓ_2 divergence. The performances of all the solvers were comparable in terms of accuracy. BPP allowed for elastic net formulation and was faster than other solvers.

In the “Background” section, we presented an analysis of CO_2 dynamics resulting from occupancy. In order to account for the slow evolution of CO_2 as a function of occupancy, we assume that each 15 minutes span of occupancy changes affect the CO_2 readings over the following several minutes, henceforth referred to as *time length*. Let the CO_2 readings in ppm be $\mathbf{x} = x_1, x_2, ..., x_N$ where the sampling interval is 1 minute and N is the total number of samples.The corresponding occupant count data is $\mathbf{y} = y_1, y_2, ..., y_N$. The matrix X and the target vector Y, defined earlier in this section, have the following structures.

$$X = \begin{bmatrix} x_1 & x_2 & .. & x_{15} & .. & x_{30} \\ x_{16} & x_{17} & .. & x_{30} & .. & x_{45} \\ .. & .. & .. & .. & .. & .. \\ x_{N-29} & x_{N-28} & .. & x_{N-15} & .. & x_N \end{bmatrix}$$

$$Y = \begin{bmatrix} mean(\mathbf{y}_{1:15}) \\ mean(\mathbf{y}_{16:30}) \\ .. \\ mean(\mathbf{y}_{N-29:N-15}) \end{bmatrix}$$

In the above matrices the *time length* is 30 minutes and the occupancy prediction resolution is 15 minutes. Thus, assuming N is a multiple of 30, the number of rows n in X and Y are $\frac{N}{15}$. Using the BPP formulation, we learn the weights $W \in \mathbb{R}^{n \times 15}$ from training data $X \in \mathbb{R}^{n \times k}$, where k is the *time length*. We used a Boosted Least Square regression (LSQ) to compute Y as a function of learnt W. In a task-driven approach to SNMF parameter selection, we iterated through four parameters: *time length*, *dictionary size*, *sparsity coefficient* and *regularization coefficient* and selected the parameter values for the lowest normalized mean square error of regression. Parameter selection was performed through cross-validation explained later in the paper.

During task-driven dictionary learning, we optimize the dictionary size and the sparsity of W in order to minimize the ℓ_2 error of occupancy count estimation from CO_2. The estimation process trains several Least Square regression models that map occupancy as a linear function of low dimensional representation of CO_2, a regression method known as Ensemble Least Square.

PerCCS algorithm

At the core of the *PerCCS* algorithm is a two step process. Step 1 is a task-driven data denoising process, applied to CO_2 data, which we label as task-driven Sparse Non-negative Matrix Factorization (SNMF) with regularization and time-shifted predictor variables. Step 2 uses this lower dimensional CO_2 signal to infer 15 minute average occupancy using the Ensemble Least Square Regression. The predictor data is time-shifted to account for delay between occupancy change and the resultant CO_2 concentration change, applied through the parameter *time length*. The basic structure of the algorithm is as follows:

1. During the data preprocessing phase, take all the CO_2 and occupancy data up to now and construct the matrix X and vector Y as shown in the previous subsection.
2. During the modeling phase, learn the parameters of the SNMF, the dictionary H and the lower-dimension representation (weights) W of X that minimizes the estimation error for 15 minutes average occupancy. Use an Ensemble Least Square Regression (LSQ) to map the representations to the past occupancy count.
3. During the prediction phase, take all the CO_2 data until now and construct the matrix X'. Compute its low-dimensional representation W' using the SNMF parameters and learnt dictionary H. Estimate average occupancy for the past 15 minutes $\hat{Y}$ from W' using the already learnt LSQ model.

EVALUATION

Competing Methods

We compared the performance of our *PerCCS* algorithm with that of Hidden Markov Models and Support Vector Regressions, as these methods had the highest prediction accuracy in the literature. Given that one of the main purposes of *PerCCS* was to denoise the data appropriately so as to minimize the occupancy estimation error, we also present a baseline scenario where the same Ensemble LSQ is used without *task-driven* denoising. Similarly for HMMs and SVMs as well, we used the raw feature matrix X, in place of its lower-dimension representation W. In the original study by Lam *et al.* (2009), the authors used second order change in CO_2, moving average CO_2 and the difference between indoor and outdoor CO_2, in addition to acoustics as the feature set. For fair comparison with our algorithm, we chose to use only indoor CO_2 concentration. The 20 minutes moving average of CO_2 used by [19] performs filtering and induces delay in the predictor features, thereby accounting for redundancy and the slow impact of occupancy on CO_2. However, from our initial feasibility study in two different rooms, we found that this delay varies from room to room between 10 and 20 minutes. Therefore, using a 20 minutes window of averaging may not be optimal for all scenarios. In *PerCCS*, the *time length* parameter takes care of the varying delay and representation learning removes redundancy through appropriate dictionary size selection. The feature selection is performed using cross-validation. More details about this can be found in the "Testing" section.

Performance Metric

We use the normalized mean squared error (NMSE) of the regression and mean of ℓ_1 error. NMSE is a variant of ℓ_2 error between actual and predicted results of a model, defined as $NMSE = \frac{1}{n\sigma_y^2}\sum_{i=1}^{q}(\hat{y}_i - y_i)^2$, where y is the actual occupancy, $\hat{y}$ is the predicted occupancy, σ_y^2 is the variance of y and n is the total number of samples. The extent to which the estimated count deviates from actual occupancy is captured by the mean of ℓ_1 error. We do not report accuracy in terms of percentage correct detection as used in Lam *et al.* [19] because accuracy in this form is more appropriate for binary presence and does not allow for small deviations from actual occupancy in large shared spaces, which may be tolerable in certain applications.

Zero occupancy is valuable information for space control, room management and space usage evaluation. Therefore, we also considered the performance of the algorithms during unoccupied periods. Further we analyzed the performance of the algorithms for periods of transient occupancy, high and low occupancy.

Data Acquisition

We collected ground truth occupant count data for 13 days in summer 2014 for a classroom of capacity 42 in a large building within the university campus. The data acquisition was event triggered, in that we manually counted the people entering and exiting our test bed from 7 am to 9 pm. This data was used to construct an occupancy based dataset with 1 minute granularities The CO_2 data for this classroom at a 1 minute sampling rate was obtained from a building management BACNet server. We recorded the status of the classroom doors and windows as closed or open and the position of the CO_2 sensor relative to the door, occupant density and distance from the sensor, as this contributes to CO_2 dynamics in space. This manual data collection was performed by two undergraduate students.

Over a 13 day period, 84% of the time the test bed was unoccupied. Rest of the time the classroom was only partially occupied with a maximum recorded occupancy of 34. During occupied periods, the average was 17.34 people (std. dev. 10.6). Our data followed a dual peak daily pattern, as expected in classrooms with an unoccupied period during the lunch break. Even though the classroom had a design capacity of 42, the peak occupancy count varied significantly, during the first half of the day in particular. We observed a maximum occupancy of 16 over the first six days and then very small occupancy for the following two days. During the last four days of monitoring, the occupancy in the first half of the day varied between 27 and 34. The occupancy fluctuates frequently before the class begins. Occasionally one or two persons left the classroom to come back right before a class started. Besides classes, small groups of students or single students used the classroom informally for meetings and phone calls. We noted that for an average of 20 minutes before a class, the occupancy gradually changed from zero to 34. Other transient periods are in between classes when some students leave and other students come in at the same time. These transients sometimes can be as short as 5 or even 10 minutes. Such transients are short for affecting the control of thermal and ventilation systems. Furthermore it would also be impractical to change the control set points of these systems just for the beginning of the class. However, the 20 minutes gradual build up of occupancy does offer an appropriate window to reset the system from unoccupied to occupied mode. The goal of our occupancy count system would be to maintain comfortable temperature and air flow rate for the small groups of people informally using the classroom and when a class is partially occupied during a full class hour. We therefore chose to estimate 15 minutes average occupancy from CO_2.

Testing

We partitioned the 13 days of CO_2 and occupancy data into training and test sets and performed 20-fold cross-validation on the training data of 863 samples spanning a period of 9 days. We performed a parametric study of SNMF and SVR for each fold and reported the mean NMSE of 20 folds for each of the SNMF, SVR, HMM and baseline methods. We explored the significance of dictionary size (number of dictionary elements) from domain knowledge, degree of sparseness, regularization and time length of CO_2 dynamics for our model. The *time length* parameter accounts for slow dynamics of CO_2, as explained in the "*PerCCS* algorithm" section. Y is the mean occupancy of 15 minutes. We tested the effect of current occupancy count on CO_2 dynamics for the next 7 time steps, with each time step being 15 minutes. The dimension of X, therefore, varies between 15 and 120. The reduced dimension of W after performing SNMF on X is governed by the dictionary size. The parameters for SVR are model cost function, and error tolerance affecting the number of support vectors. Similar to the work of Lam *et al.* [19] we use a Gaussian kernel for SVR. The baseline method uses the same 15 minutes average Y and X matrix without denoising.

Metric	**Baseline**	***PerCCS***	**SVR**
NMSE	0.16	0.075	0.51
Mean ℓ_1 error total (persons)	1.3	1.0	2.54
Std. dev. ℓ_1 error total	3.4	2.1	5.5
Mean ℓ_1 error OP (persons)	5.1	3.2	7.1
Std dev ℓ_1 error OP	4.1	3.4	10.7
Zero occupancy detection (% UP)	0	91	0
Exact occupancy detection (transience) (% OP)	0	15	0

Table 1. Algorithm performances for two metrics. OP and UP stand for occupied period and unoccupied period respectively.

RESULTS

One of our first observations was that while regression approaches like SNMF-regression and SVR perform satisfactorily with only 9 days of sparse training data, using an HMM requires a greater number of samples per state to achieve reasonable accuracy. Our test dataset consisted of 73 data points of occupied periods and 310 data points of unoccupied periods after averaging over 15 minutes. Moreover some of the higher occupancy values were observed only 1 - 5 times in the test data and some values in the test data did not have a corresponding match in the training data. After testing the HMM, we found that the predictions were biased towards low values and performed worse than random due to skewed data. In order to increase the number of samples per state we could bin the data, but that would be an approximation already. While HMMs have been found to be accurate for a small occupancy of 4, they do not provide a scalable solution, as the data requirement grows with the maximum occupancy of the space. Discriminative models such as as regression are known to perform better with small datasets. This means that the required training period can be smaller for our algorithm. We choose not to report the HMM results further.

Table 1 presents the comparison of performance of our method against baseline (column 1) and SVR (column3) in terms of NMSE, overall mean ℓ_1 error and the standard deviation of ℓ_1 error and during occupied period only. SVR with the Gaussian kernel has the highest error, followed by the baseline method. The baseline model over-estimates the occupancy compared to SNMF-regression overall (1.3 versus 1.0). Further, focusing on periods when the room is occupied, *PerCCS* considerably outperforms the baseline and SVR. This could be attributed to high transient peaks in the CO_2 concentration, not necessarily contributed by occupancy, that is not removed without denoising. Moreover, the standard deviation of ℓ_1 error is higher for the baseline and

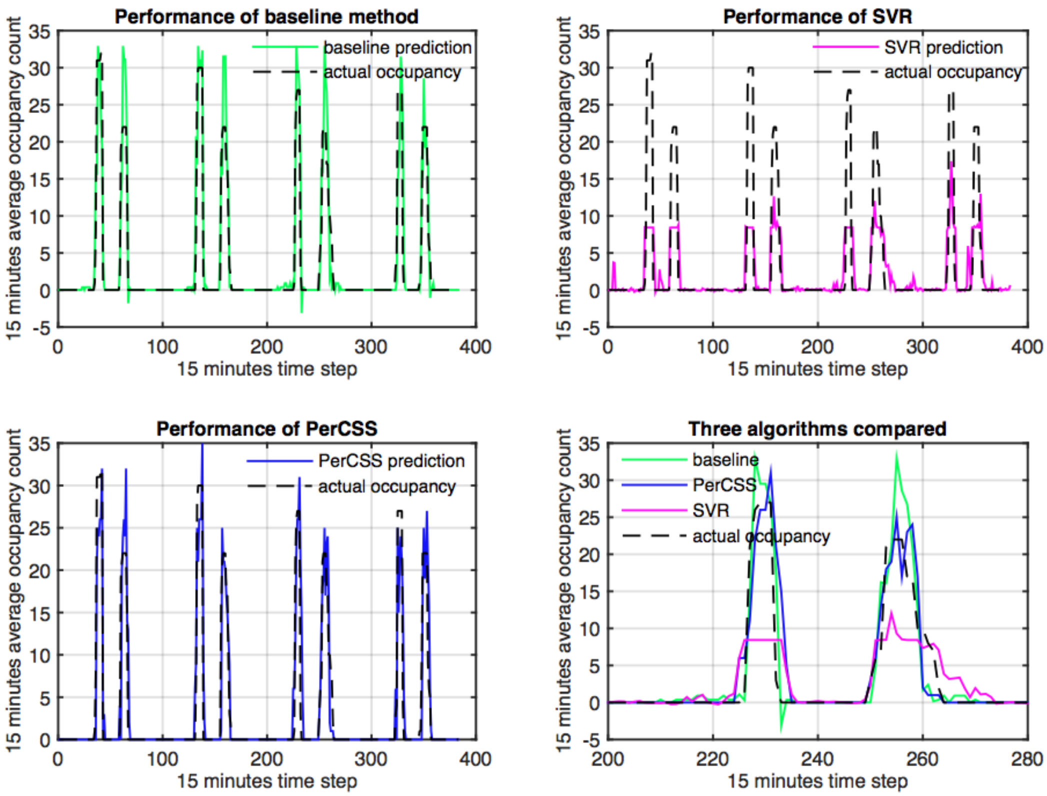

Figure 4. Prediction results of occupant count from three algorithms during the validation period

SVR approaches (see Table 1). SNMF-regression, on the other hand, captures the fluctuations in occupancy better than SVR, especially during the occupied period. SVR also fails to predict zero occupancy, which SNMF-regression models accurately. This is because the low dimension representation W learnt using SNMF is always positive and possibly represents the sparsity of the occupancy to some extent.

A demonstration of the results of prediction of the three methods are displayed in Figure 4. The NMSE results in increasing order are 0.075 (*PerCCS*), 0.16 (baseline) and 0.51 (SVR). Also, note that the during occupied periods, the mean of the ℓ_1 error in occupancy estimation are 3.2 people as opposed to 5.1 and 7.1 for the baseline and SVR, respectively.

We also examined the performance of the 3 algorithms during zero occupancy. SVR shows a mean error of 1 with negligible standard deviation. While both the baseline and *PerCCS* are able to predict zero occupancy with an ℓ_1 error < 1, 91% of the unoccupied periods is estimated correctly by *PerCCS*, unlike baseline or SVR, which have almost a 100% incorrect prediction. We also studied how well our algorithm is able to estimate fluctuations in occupancy during the occupied periods. We report this is terms of percentage of occupied periods when the occupancy count is estimated exactly. This is 15% for *PerCCS*, but baseline and SVR were never able to predict occupancy count exactly i.e., 0%.

The purpose of the task-driven dictionary learning was to learn the dictionary and weight matrix parameters in such a way that minimizes the occupancy estimation error. Prior research indicated concern about latency of occupancy estimation from CO_2 due to slow dynamics of the latter. We found that occupancy has some observable effect on one time step ahead (in our case, 15 minutes) for the CO_2 dynamics, as seen across several time series. In other words using a *time length* of 30 produced better results than that of 15. This means that the average occupancy of the current 15 minutes time step has the highest correlation with the next 15 minutes of CO_2. This aligns with our initial finding from the feasibility study for the same test bed, where occupancy was best predicted by CO_2 data with 10 minutes lead. Using 20 folds cross-validation, the minimum NMSE was obtained for a dictionary size of 5.

DISCUSSION

In *PerCCS* we learnt a lower-dimensional representation from noisy CO_2 data and used it as a predictor to estimate 15 minutes average occupancy count with Ensemble Least Square Regression method.

We found that small dictionary size produced lower NMSE, which means that a small number of contributing factors can explain the evolution of CO_2 in our test classroom. This supports our initial domain knowledge, that mostly four indoor parameters like occupant count, ventilation airflow rate, outdoor CO_2 concentration (if the room has infiltration and/or natural ventilation) and plants' contribution contribute to measured CO_2 concentration in room air. Moreover we are concerned about only one of the contributing factors, i.e., occupancy count, for which the patterns can be captured with a small number of basis functions. The rest of the factors can in fact be aggregated into a single structured noise representation for the purpose of our regression. Thus, even though the reconstruction error of SNMF decreases with higher dictionary sizes, the same does not improve the regression performance. Low sparsity is probably an outcome of the small dictionary size, when few basis functions are able to explain the sparsity structure of most of the CO_2 data.

We compared the performance of this model with that of LSQ regression without data denoising and SVR in terms of NMSE, mean and standard deviation of ℓ_1 error during occupied and unoccupied period. We found that SVR has the lowest performance as this model generalizes more and is unable to capture the fluctuations in CO_2, probably because of several zero values in the data. *PerCCS* predicted 15 minutes average occupancy fairly accurately (1.0 person error on average) with an NMSE of 0.075.

We also observed that our algorithm performs better than both the baseline method and SVR during unoccupied periods as well with 91% accuracy. *PerCCS* can also better capture transient occupancy without overshooting like the baseline method or over-generalizing like the SVR approach. The *task-driven* denoising plays a key role in capturing the sparsity of occupancy in low-dimensional representations of CO_2 as we had hypothesized. Next we evaluate how the performance of *PerCCS* in terms of accuracy and latency, will affect the potential smart environment applications like smart meeting space allocation, opportunistic meeting support, design evaluation of collaborative environments and energy-efficient and comfort-sensitive heating and cooling operation. We also point out the limitations of our system based on the characteristics of our testbed.

In current building management systems, lights are turned off or temperature setups or setbacks are applied in rooms that have been detected as unoccupied for more than 10-15 minutes, typically using Passive Infrared (PIR) sensors which are fast but unreliable. *PerCCS*, can accurately (91%) detect unoccupied periods with 15 minutes latency. If CO_2 sensors are deployed for binary occupancy-based building operations, then the current 10-15 minutes inaction period after occupancy detection should be adjusted to allow for the built-in latency of our system. The mean ℓ_1 error in occupancy count during the unoccupied period in our testbed was found to be less than +1, an unlikely occupancy count for rooms with scheduled meetings. Therefore, rooms can be declared empty with sufficient certainty based on our detection system.

PerCCS was able to estimate the exact occupancy 15% of the time with an NMSE of 0.075 and mean ℓ_1 error of 3.2 persons as opposed to 7.1 persons using SVR and 5.1 persons in the baseline. We observed that the highest errors occurred during the periods when actual occupancy had exceeded 20. The effect of this deviation on the allocation of large meeting rooms may be small. While the goal of this paper was to estimate occupancy count in large shared spaces, we also studied the scalability of the method in scenarios like opportunistic meetings, where it is important to distinguish between the presence of one person and multiple persons in a room. Low occupancy was rare in our dataset. Occupancy is always slightly overestimated or underestimated during periods when only one person is present in the room. Occupancy ranged between 2-5 in only 9 instances in our dataset, of which 8 instances had deviations of only 1 to 2 people from the actual.

Occupancy state	**Conditioning**
Occupancy detection	Binary, PIR sensors
Occupied	Temperature set by occupants
Unoccupied (heating)	Temperature setpoint+4 F
Unoccupied (cooling)	Temperature setpoint-4 F
Ventilation (occupied)	Set for maximum occupancy

Table 2. Control systems in test bed.

The limitations of the *PerCCS* system can be categorized as sensor-specific and testbed-specific. The sensor-specific limitation of 15 minutes latency due to slow CO_2 dynamics should be taken into account in any control system leveraging *PerCCS*. The control strategy for HVAC, heating and cooling in the testbed building in our study is shown in Table 2. The HVAC system in its current state does not use the CO_2 data. The constant airflow rate allows the CO_2 concentration in this room to fluctuate according to the occupancy. However, in rooms where ventilation is controlled to limit the CO_2 to a certain level [11], the predictive capability of a CO_2-based occupancy count system may be compromised near maximum occupancy levels irrespective of room size. While we do not have a suitable testbed at our disposal to study the performance of *PerCCS* in rooms with demand-controlled ventilation, future studies should be conducted to ensure further scalability of our system.

CONCLUSION

Knowledge of occupancy count is valuable for applications such as efficient temperature and ventilation control, room utilization, opportunistic meeting support and emergency responses in buildings. Few buildings, however, have the means of knowing occupancy beyond binary presence-absence. Existing work has leveraged additional infrastructure, multi-sensor fusion and/or intrusive instrumentation like imaging, and have only been applied to the estimation of a small occupancy count (less than 20). The goal of our work was to explore the feasibility of modeling and predicting 15 minutes average occupant count in large classrooms as a function of CO_2 concentration using a new algorithm, *PerCCS*. *PerCCS* uses only a single sensor commonly found in rooms, and extends occupancy count to classrooms. It uses task-driven sparse non-negative matrix factorization with regularization to reduce the redundancy and noise in measured CO_2. We found that *PerCCS* can achieve a low Normalized Mean Square Error of 0.075 and an average over-estimation of 1.0 person with a latency of 15 minutes, both significantly better than a baseline Ensemble Least Square regression without denoising and Support Vector Regression. *PerCCS* outperformed the baseline and existing methods in predicting occupancy count with 91 % and 15 % exact occupancy estimation when the room was unoccupied and occupied respectively, where the baseline and existing methods failed completely. On average *PerCCS* deviated from actual occupancy by 1 person, better than both the baseline and SVR. Moreover, the parameters of *PerCCS* can be learnt for different rooms with a few weeks of training data.

FUTURE WORK

One of the limitations of our work is the paucity of data. During cross-validation we already noted that inadequacy of data may generate higher variance in errors across folds. We plan to collect more data from a greater variety of rooms to test *PerCCS* and the compared methods. A natural extension of our work would be to determine when to automatically update the dictionary from new raw CO_2 data. *PerCCS* in its current form is a reactive system, slow for temperature control. However, with more data, as we confirm that the low-dimensional representation of CO_2 learnt using SNMF reflects the occupancy structure of the space well, we may be able to use CO_2 as a direct proxy for occupancy, and then use its trend for predictive temperature control.

REFERENCES

1. Buildings Energy Databook, 2012. `http://buildingsdatabook.eren.doe.gov/TableView.aspx?table=1.1.3`.
2. Agarwal, Y., Balaji, B., Gupta, R., Lyles, J., Wei, M., and Weng, T. Occupancy-driven energy management for smart building automation. In *BuildSys 2010*, ACM Press (2010), 9–18.
3. Balaji, B., Xu, J., Nwokafor, A., Gupta, R., and Agarwal, Y. Sentinel: Occupancy based hvac actuation using existing wifi infrastructure within commercial buildings. In *ACM SenSys 2013* (2013), 17:1–17:14.
4. Beltran, A., Erickson, V. L., and Cerpa, A. E. Thermosense: Occupancy thermal based sensing for HVAC control. In *Proceedings of the Fifth ACM Workshop on Embedded Sensing Systems for Energy-Efficient Buildings (BuildSys 2013)*, ACM, Ed. (2013), 11:1–11:8.
5. Cichocki, A., Zdunek, R., and Amari, S.-i. Hierarchical als algorithms for nonnegative matrix and 3d tensor factorization. *Independent Component Analysis and Signal Separation* (2007), 169–176.
6. Erickson, V., Achleitner, S., and Cerpa, A. E. POEM: Power-efficient occupancy-based energy management system. In *Proceedings of the Twelfth ACM/IEEE International Conference on Information Processing on Sensor Networks (IPSN 2013)*, ACM/IEEE, Ed. (2013), 203–216.
7. Erickson, V., Carreira-Perpinan, M. A., and Cerpa, A. E. Occupancy modeling and prediction for building energy management. *ACM Transactions on Sensor Networks 10*, 42 (2014), 23–33.
8. Erickson, V., and Cerpa, A. E. Occupancy based demand response HVAC control strategy. In *Proceedings of the Second ACM Workshop on Embedded Sensing Systems for Energy-Efficient Buildings (BuildSys 2010)*, ACM, Ed. (November 2010), 7–12.
9. Erickson, V. L., Beltran, A., Winkler, D. A., Esfahani, N. P., Lusby, J. R., and Cerpa, A. E. TOSS: Thermal occupancy sensing system. In *Proceedings of the Fifth ACM Workshop on Embedded Sensing Systems for Energy-Efficient Buildings (BuildSys 2013)*, ACM, Ed. (2013), 35:1–35:2.
10. Erickson, V. L., Lin, Y., Kamthe, A., Brahme, R., Cerpa, A. E., Sohn, M. D., and Narayanan, S. Energy efficient building environment control strategies using real-time occupancy measurements. In *Proceedings of the 1st ACM Workshop On Embedded Sensing Systems For Energy-Efficient Buildings (BuildSys 2009)*, ACM, Ed. (November 2009), 19–24.
11. Fisk, W. J., and De Almeida, A. T. Sensor-based demand-controlled ventilation: a review. *Energy and buildings 29*, 1 (1998), 35–45.
12. Horvitz, E., Koch, P., Kadie, C. M., and Jacobs, A. Coordinate: Probabilistic forecasting of presence and availability. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, Morgan Kaufmann Publishers Inc. (San Francisco, CA, USA, 2002), 224–233.
13. Kamthe, A., Carreira-Perpinan, M. A., and Cerpa, A. E. Enabling building energy auditing using adapted occupancy models. In *Proceedings of the Third ACM Workshop on Embedded Sensing Systems for Energy-Efficient Buildings (BuildSys 2011)* (2011), 33–38.
14. Kamthe, A., Jiang, L., Dudys, M., and Cerpa, A. Scopes: Smart cameras object position estimation system. In *Proceedings of the Sixth European Conference on Wireless Sensor Networks (EWSN 2009)*, Springer-Verlag, Ed. (2009), 279–295.
15. Kim, J., and Park, H. Fast Nonnegative Matrix Factorization: An Active-Set-Like. *SIAM Journal on Scientific Computing 33*, 6 (2011), 3261–3281. `http://dx.doi.org/10.1137/110821172`.
16. Koehler, C., Ziebart, B. D., Mankoff, J., and Dey, A. K. TherML: occupancy prediction for thermostat control. In *UbiComp '13*, ACM (2013), 103–112.
17. Kolter, Z., Batra, S., and Ng, A. Y. Energy disaggregation using discriminative sparse coding. In *Proceedings of NIPS 2011*, Curran Associates, Inc. (2011), 1153–1161.
18. Kraut, R. E., Fish, R. S., Root, R. W., and Chalfonte, B. L. Informal communication in organizations: Form, function, and technology. *Human reactions to technology: Claremont symposium on applied social psychology* (1990), 145–199.

19. Lam, K., Hynck, M., Dong, B., Andrews, B., Chiou, Y., Zhang, R., D., B., and Choi, J. Occupancy detection through an extensive environmental sensor network in an open-plan office building. *IBPSA Building Simulation* (2009), 1452–1459.

20. Liao, C., and Barooah, P. An integrated approach to occupancy modeling and estimation in commercial buildings. In *American Control Conference (ACC), 2010* (June 2010), 3130–3135.

21. Mairal, J., Bach, F., and Ponce, J. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34*, 4 (2012), 791–804.

22. Majumdar, A., Albonesi, D. H., and Bose, P. Energy-aware meeting scheduling algorithms for smart buildings. In *Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, BuildSys '12, ACM (New York, NY, USA, 2012), 161–168.

23. Meyn, S., Surana, A., Lin, Y., Oggianu, S., Narayanan, S., and Frewen, T. A. A sensor-utility-network method for estimation of occupancy in buildings. In *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on* (Dec 2009), 1494–1500.

24. Mumma, S. A. Transient occupancy ventilation by monitoring co_2. In *ASHRAE IAQ 2004: Critical Operations: Supporting the Healing Environment Through IAQ Performance Standards* (2004), 21–23.

25. Potluru, V. K., Plis, S. M., Le Roux, J., Pearlmutter, B. A., Calhoun, V. D., and Hayes, T. P. Block coordinate descent for sparse nmf. In *First International Conference on Learning Representations (ICLR 2013)* (May 2013).

26. Ranjan, J., Yao, Y., and Whitehouse, K. An rf doormat for tracking people's room locations. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM (New York, NY, USA, 2013), 797–800.

27. Schmidt, M. N., and Olsson, R. K. Single-channel speech separation using sparse non-negative matrix factorization. In *Ninth International Conference on Spoken Language Processing* (2006), 2614–2617.

28. Scott, J., Bernheim Brush, A. J., Krumm, J., Meyers, B., Hazas, M., Hodges, S., and Villar, N. PreHeat: controlling home heating using occupancy prediction. In *UbiComp '11*, ACM (2011), 281–290.

29. Teixeira, T., Dublon, G., and Savvides, A. A Survey of Human-Sensing: Methods for Detecting Presence, Count, Location, Track, and Identity, 2010. `http://www.eng.yale.edu/enalab/publications/human_sensing_enalabWIP.pdf`.

30. Trumler, W., Bagci, F., Petzold, J., and Ungerer, T. AMUN: autonomous middleware for ubiquitous environments applied to the smart doorplate project. *Journal of Advanced Engineering Informatics 19*, 3 (2005), 243–252.

31. Wytock, M., and Kolter, Z. Contextually supervised source separation with application to energy disaggregation. In *AAAI Conference on Artificial Intelligence, 2014* (2014), 486–492. `http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8629`.

32. Zaveri, S., Ross, S., Erickson, V., Kamthe, A. U., Liu, T., and Cerpa, A. E. Building energy management systems actuated using wireless camera sensor network. In *Proceedings of the Third ACM Workshop on Embedded Sensing Systems for Energy-Efficient Buildings (BuildSys 2011)* (2011), 39–40.

33. Zhang, J., Liu, G., Lutes, R., and Brambley, M. Energy savings for occupancy- based control (obc) of variable-air-volume (vav) systems, January 2013. `http://www.pnnl.gov/main/publications/external/technical_reports/PNNL-22072.pdf`.

34. Zou, H., and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67* (2005), 301–320.