Building Simulation 2009
Eleventh International IBPSA Conference
Glasgow, Scotland
July 27-30, 2009

OCCUPANCY DETECTION THROUGH AN EXTENSIVE ENVIRONMENTAL SENSOR NETWORK IN AN OPEN-PLAN OFFICE BUILDING

Khee Poh Lam[1], Michael Höynck [2], Bing Dong[1], Burton Andrews[2], Yun-Shang Chiou[1], Rui Zhang[1], Diego Benitez[2] and Joonho Choi[1]
[1]Center for Building Performance and Diagnostics, Carnegie Mellon University, Pittsburgh, PA, U.S.A. 15213
[2]Research and Technology Center, Robert BOSCH LLC, Pittsburgh, PA, U.S.A.15212

ABSTRACT

Contemporary office buildings commonly experience changes in occupancy patterns and needs due to changes in business practice and personal churns. Hence, it is important to understand and accurately capture the information of such trends for applications in building design and subsequent building operations. Detection of occupant presence has been used extensively in built environments for applications such as demand-controlled ventilation and security, and occupancy profiles are widely used in building simulations. However, the ability to discern the actual number of people in a space is often beyond the scope of current sensing techniques. This paper presents a study to develop algorithms for occupancy number detection based on the analysis of environmental data captured from existing sensors and ambient sensing networks. Both wireless and wired sensor networks are deployed in the Robert L. Preger Intelligent Workplace (IW) at Carnegie Mellon University, comprising six different types of sensors. An average of 80% accuracy on the occupancy number detection was achieved by Hidden Markov Models during testing periods. The findings also offer encouraging possibilities for incorporating the algorithms into building management systems for optimizing energy use while maintaining occupant comfort.

INTRODUCTION

A fundamental goal of energy efficient and high performance buildings is to facilitate a comfortable, healthy and productive environment for the occupants while maintaining minimum energy consumption. Information regarding the number of occupants in a building space is a key component to achieving this task and is useful for numerous applications such as lighting control or demand-controlled ventilation.

Current approaches to occupancy detection take place mostly in commercial buildings through the use of passive infrared (PIR) motion detectors. However, motion detectors have inherent limitations when occupants remain relatively still. The use of probabilistic models offers improved capability of detecting occupant presence (Dodier et al. 2006, Page et al. 2008). However, the fundamental dependence on motion still remains. Moreover, motion detectors alone only provide information regarding the presence or absence of people in a space rather than the number of occupants, information which is highly useful for building control tasks such as demand controlled ventilation (Emmerich, 2001). Video cameras have been used in this regard (Stanislay et al., 2006 and Trivedi et al., 2000); however, video capture raises privacy concerns and requires large amounts of data storage. Other work has focused on the use of carbon dioxide (CO_2) sensors in conjunction with building models for estimating the number of people generating the measured CO_2 level (Federspiel 1997, Wang et al. 1998). Sufficient models, though, are often not easy to obtain and extensions to complex or open spaces may be difficult. Recent research on so-called smart environments involves the use of a diverse set of sensors to monitor and infer human activity in a building. Examples include the Georgia Tech Aware Home (Lesser et al., 1999), the MIT Intelligent Room (Torrance, 1995), the University of Colorado Boulder Neural Network Adaptive Home (Mozer, 1998), and the University of Texas at Arlington MavHome (Cook et al, 2004, Youngblood et al., 2007). Most of these works focus on behavioural modeling or mobility tracking and do not exploit additional sensing capability for the detection of occupancy numbers. Furthermore, these test environments are most commonly residential buildings. In general, occupancy detection that fully exploits information available from low cost, non-intrusive, environmental sensors is an important yet little explored problem in office buildings_To investigate the use of ambient sensors for detecting the number of occupants in an office building, a comprehensive, ubiquitous, environmental sensing test-bed was deployed in the Robert L. Preger Intelligent Workplace (IW) at Carnegie Mellon University. The overall goal of this test-bed is to integrate state-of-the-art IT systems as well as sensing, actuating, and controls technologies to achieve energy efficiency while providing a healthy and productive environment. This test-bed includes distributed sensors for a variety of environmental parameters such as CO_2, carbon-monoxide (CO), total volatile organic compounds (TVOC), small particulates (PM2.5), acoustics, illumination, motion, temperature, and humidity. The contribution of the test-bed lies in the magnitude and diversity of the sensor infrastructure deployed as well as the ability to capture data continuously with very little human intervention. While the aim of the study described here is on the detection of the number of occupants in the building space, this test-bed is an ideal testing environment for a large variety of building technology research areas such as human-centered environmental control, security and energy efficient and sustainable green buildings. In particular, the derived occupancy information can be used as an input for both validating building simulation models and simulating new building or control designs on realistic occupancy profiles.

The paper is organized as follows. The ambient sensing infrastructure of the sensor network, the environmental parameters measured, and the underlying database setup are described. This is followed by an analysis of the correlation between the measured environmental parameters and occupancy level to be used as an indication of which ambient sensors provide the most information regarding the number of people in the space. The most relevant features are then used in conjunction with several machine learning techniques, providing results for the detection of occupancy levels from the sensing network.

SENSOR NETWORK DEPLOYMENT IN THE IW

The IW, depicted in Figure 1, is an open plan office space with sixteen rooms (bays) and one conference room. It provides accommodation for five faculty members, twelve PhD students and two staff members. Many visitors frequent the IW every day, and classes are held in the conference room. The entire indoor environment can be considered heavily dynamic.

Sensing network

The sensing infrastructure deployed in the IW is divided into three separate sensor network systems: a wired sensors gas detection sensor network, which measures CO_2, CO, TVOC, outside temperature, dew point, and small particulates (PM2.5); a wireless ambient-sensing network, which measures lighting, temperature, relative humidity (RH), motion and acoustics; and an independent CO_2 sensor network. This multitude of ambient information captured by the network aims at capturing the different methods of interaction possible between occupants and their working environment, namely the emission of heat, the emission of pollutants such as CO_2 and odor, and the generation of sound (Page et al., 2007). Such a diverse set of sensors also allows for investigation into which environmental parameters have the greatest correlation with occupancy levels. The sensor layout is illustrated in Appendix 1. This experimental setup employs an on-line transaction processing (OLTP) / on-line analytical processing (OLAP) database management structure. All sensing systems (gas detection, CO_2, wireless sensor and camera networks) have their own respective OLTP databases that update the information of the sensors and sensing data continuously. To integrate the sensing data information for data analysis, a separate, standalone OLAP database was established as a data warehouse

Figure 1. Robert L. Preger Intelligent Workplace

In addition to the sensing networks, a video camera system is deployed with a video camera in each of several IW bays. Captured videos can be analyzed by user-assisted software to determine the number of occupants in the space at a given time. This information is used for ground truth occupancy profiles in our analysis.

Data collection

Data collection in the IW for this work took place during two continuous periods and in two bays. The time periods are 1) January 29th to March 7th, 2008; and 2) March 17th, 2008 to April 4th, 2008. Occupancy data was recorded on weekdays from 8:00 am to 6:00 pm from the two bays with the most frequent occupancy activity, bays 13 and 10. Table 1 lists the details of each dataset and the label for the dataset used throughout the rest of the paper.

Table 1 Data collection periods

Dataset	Bay	Starting date	Ending date	# Data points in total
B13_P1	13	01/29/08	03/07/08	21528
B13_P2	13	03/17/08	04/04/08	7705
B13_P3	13	03/27/08	04/03/08	1156
B10_P1	10	01/29/08	03/07/08	20702
B10_P2	10	03/17/08	04/04/08	7555
B10_P3	10	03/27/08	04/03/08	1157

FEATURE SELECTION

We first explore which features of the environmental sensing network provide the most useful information in the detection and prediction of the occupancy number. To this end, we use the notion of information gain, which is a measure of the amount of uncertainty of the input of a system given the value of the output. We present here a brief overview of the methodology and results of the feature selection analysis; a full report of the details can be found in (Lam et al., 2009).

Information gain

Mathematically, the relative information gain between two random variables *x* and *y* is defined as (Mitchell, 1997)

$$RIG(y,x) = \frac{IG(y,x)}{H(y)} \cdot 100\% \tag{1}$$

where the mutual information *IG* is

$$IG(y,x) = H(y) - H(y \mid x) \tag{2}$$

and the entropy *H(y)* is a measure of the inherent uncertainty of the random variable *y*:

$$H(y) = \sum_{i=1}^{n} -p(y_i)\log_2 p(y_i) \tag{3}$$

with *n* indicating the total number of values the random variable *y* can take. High entropy corresponds to high uncertainty and vice versa. We use information gain in this study to assess the correlation between occupancy and different sets of features derived from the sensor data. In general, the feature set is comprised of the following features computed for each ambient sensor: the original output, first order difference, second order difference and difference between the indoor and outdoor values. For CO_2 and acoustics, a 20 minute moving average value is also considered. We employ a tool (Anderson and Moore, 1998) that uses an exhaustive search algorithm to check all possible feature combinations from the feature space and then select the most informative combination of features based on the relative information gain.

Results from feature selection

Table 3 shows an example of the feature selection analysis on CO_2 data for a particular bay. The features investigated are shown in Table 2.

Information gain is computed for increasing numbers of input features and, for each iteration, feature combinations yielding the highest information gain are noted (indicated by the check marks in Table 3). This analysis is repeated for each bay, and the number of selections of each feature is totalled to obtain the most informative features for a given sensor. For instance, the three most informative features for CO_2 are found after totalling the selections across all bays to be *CO_2_Out*, *CO_2_FD2* and *CO_2_MA_20min*.

A similar analysis was conducted combining the three most informative features for a given sensor with those from other sensors. A detailed analysis can be found in (Lam, et al., 2009).

Table 2 Investigated features of CO_2

Feature	Description
CO_2_FD	1st order difference of CO_2: CO_2(t(i))-CO_2(t(i-1))
CO_2_FD2	1st order shifted difference of CO_2 (CO_2(t(i))-CO_2(t(i-2))
CO_2_SD	2nd order difference of CO_2: CO_2_FD(t(i))–CO_2_FD(t(i-1))
CO_2_Diff	1st order difference of CO_2 difference between indoor and outdoor: CO_2_Diff(t(i))-CO_2_Diff(t(i-1))
CO_2_Diff_FD	*CO_2_Diff_SD* (2nd order difference of CO_2 difference between indoor and outdoor: CO_2_Diff_FD(t(i))-CO_2_Diff_FD(t(i-1))
CO_2_MA_20min	20 minutes of moving average of CO_2 measurement

Table 3 Information gain with different number of features as output for CO_2 for the period B13_P1

CO_2	*CO_2_FD*	*CO_2_FD2*	*CO_2_SD*	*CO_2Out*	*CO_2_Diff*	*CO_2_Diff_FD*	*CO_2_Diff_SD*	*CO_2_MA_20min*	RIG(%)
		√		√	√				20
		√		√	√		√		28
		√		√		√	√	√	40
		√	√	√	√	√	√	√	52
√		√	√	√	√		√	√	60
√		√	√	√	√	√	√	√	67
√	√	√	√	√	√	√	√	√	67

Summarizing the results, thermal performance parameters such as temperature and relative humidity are dominated more by the building heating, cooling, and ventilation systems. The selected features giving the largest information gain are found to be: CO_2, *CO_2_Diff*, *CO_2_FD2* and *CO_2_MA_20min* acoustics, acoustics_FD2 and PIR. These features are used as inputs to the occupancy estimation methods discussed below. Note that the occupancy estimation methods were also evaluated with additional feature combinations, and those yielding the best results were consistent with the results of Lam et al. (2009).

OCCUPANCY DETECTION ANALYSIS

Occupancy Estimation Methodology

In this section, three popular machine learning technologies including Support Vector Machines (SVM), Neural Networks (NN) and Hidden Markov Models (HMM) are introduced as possible techniques for studying the occupancy detection.

Support Vector Machine

Support vector machines, developed by Vapnik and his co-workers in 1995, have been widely applied in classification, forecasting and regression of random data sets. Their practical success can be attributed to solid theoretical foundations based on Vapnik-Chervonenkis Theory (Cherkassky, 2004). The detailed theory and principles can be found in (Vapnik, 1995). One of the primary features of SVM is to map non-linear functions in a low dimensional space to a higher dimensional space through the use of a kernel function. Most previous reported studies used a Gaussian function as the kernel model for regression analysis. A SVM with a Gaussian kernel is applied to this sensor network data. The LibSVM toolkit developed by Chang and Lin (2001) is then used to train and test the data sets. In order to avoid over-fitting, a ten-fold cross validation was conducted on the data sets.

Neural Network

An Artificial Neural Network (ANN) is an interconnected group of artificial neurons that uses a mathematical or computational model for information processing based on a connection approach to computation.

An ANN of two hidden layers with different combinations of neuron numbers in each hidden layer was tested on the data from the IW. Figure 2 shows the structure of the ANN. The neural network applied in this study is used for creating, training, and simulating a fully-connected, feed-forward network. Fully-connected means that each node is connected to all other nodes in the adjacent layers, and feed-forward indicates that information is passed in a single direction from the input to the output nodes.

The learning algorithm employed is the back-propagation, generalized delta method. In this algorithm, the value of the output of the NN is compared to a target value to determine an error. The weights associated with the connection between nodes are then adjusted in a backward direction from the output layer to the input layer in order to minimize this error.

The ANN was implemented using the Matlab Neural Network toolbox. The input layer inputs the most important features obtained from the results of feature selection. The Log Sigmoid function is used as the transfer function in all hidden layers, and a linear function is used in the output layer. Because neural networks are not guaranteed to reach a global solution, training is repeated 10 times, and the output results are averaged.

Hidden Markov Model

A hidden Markov model is a statistical model in which the system being modeled is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters from the observable parameters. The extracted model parameters can then be used to perform further analysis, for example, for pattern recognition applications. A HMM can be considered to be the simplest dynamic Bayesian network.

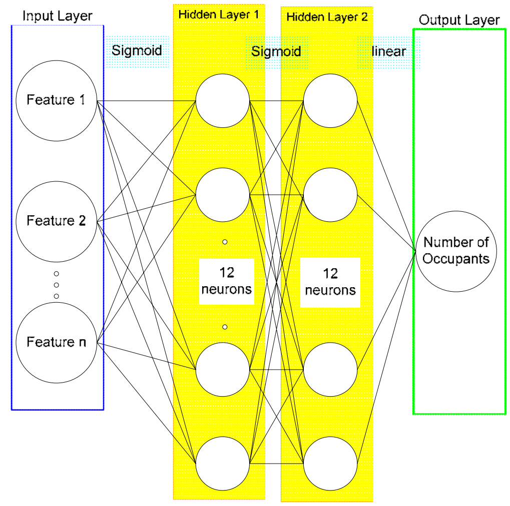

Figure 2 Structure of 2-hidden layer Neural Network

In this study, the occupancy number is considered to be a hidden state and the most important features from the sensor network as observations as shown in Figure 3. Unlike the NN approach, the HMM method explicitly accounts for temporal correlations between occupancy levels and environmental parameters in consecutive time steps. This temporal information has the potential to greatly improve prediction.

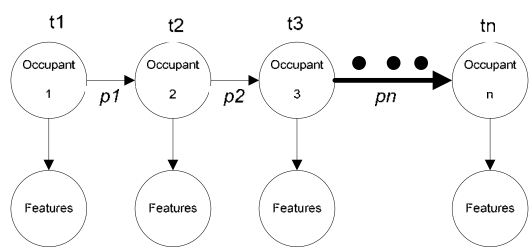

Figure 3 Structure of HMM

To train the HMM, the forward and backward algorithm is applied. The update rule is (Rabiner, 1989):

(1)Initialize: $\alpha_1(X_1) = P(X_1)P(O_1|X_1)$ (4)

Where O1..n are observed sensor values.

(2)For i=2 to n,

$$\alpha_i(X_i) = \sum_{x_{i-1}}^{n} P(O_i|X_i)P(X_i|X_{i-1} = x_{i-1})\, \alpha_{i-1}(x_{i-1}) \quad (5)$$

$X_i$and X_{i-1} are the number of occupancy in time t and time t-1.

(3) Initialize: $\beta_n(X_n) = 1$ (6)

(4) For i =2 to n,

$$\beta_i(X_i) = \sum_{x_{i+1}}^{n} P(O_{i+1}|x_{i+1})P(x_{i+1}|X_{i+1})\, \beta_{i+1}(x_{i+1}) \quad (7)$$

(5) Finally, $P(X_1|O_{1\ldots n}) \propto \alpha_i(X_i)\beta_i(X_i)$ (8)

Where

α_i is a forward factor;
β_i is a backward factor;
X_i the ith state;
O_i the ith observation;

The final estimation is obtained from Equation (8), which is the maximum probability based on the current sensor observations and previous occupant number.

Occupancy estimation results

Results from NN and SVM

Figures 4 and 5 show the results from the SVM and ANN analysis, respectively. Data for one day (March 21) was used for testing, and the remaining dates were used for training. The x axis corresponds to time in terms of the number of samples (sampling time is once per minute), and the y axis the number of occupants in the space. The blue line is the actual occupancy profile and the red dotted line is the estimation.

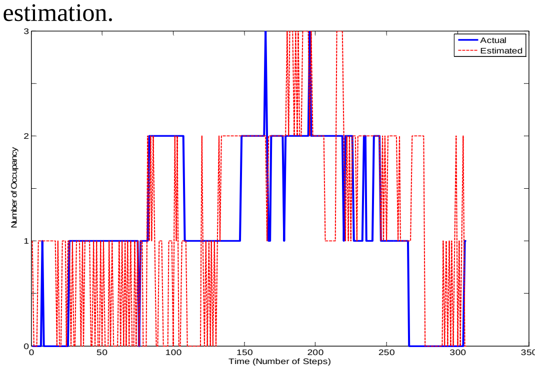

Figure 4 Occupancy Estimation Results of Bay 13_P2 on March 21 with SVMs of 73% accuracy

Both SVMs and ANN generate rather noisy occupancy estimates with frequent fluctuations. This can in part be attributed to the SVM and ANN assumption that each data point is independent and identically distributed, which is not always accurate. This is particularly true with respect to parameters such as CO_2 because of the strong temporal correlations inherent in CO_2 measurements. The HMM approach is more well suited to account for these temporal correlations because of the dynamic Markov properties.

Results from Hidden Markov Model

Figure 6 shows the result of the HMM estimate on the same date, March 21. Compared to the results from SVM and ANN, the estimate profile is much smoother and reasonable. The estimated occupancy profile tracks very well with the actual profile with an accuracy of 75% (number of correctly estimated points divided by the total number of points).

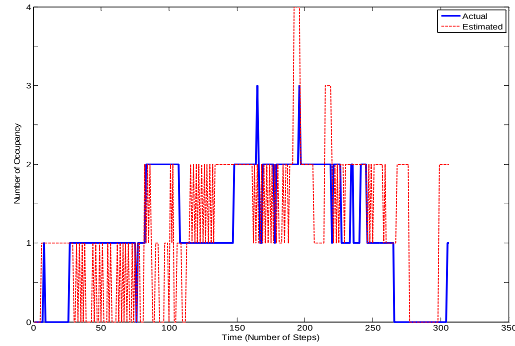

Figure 5 Occupancy Estimation Results of Bay 13_P2 on March 21 with ANN of 75% accuracy

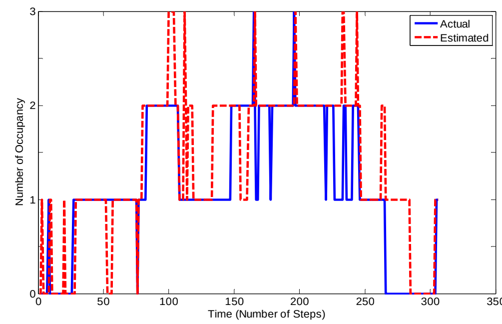

Figure 6 Occupancy Estimation Results of Bay 13_P2 on March 21 with HMM of 75% accuracy

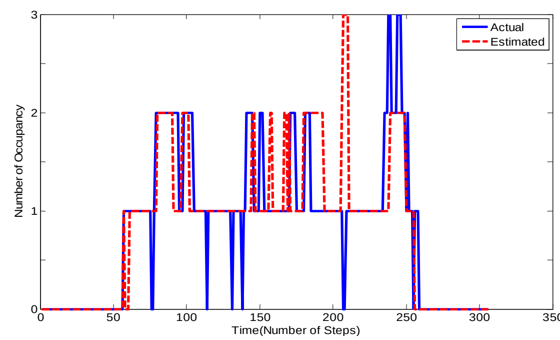

Figure 7 Occupancy Estimation Results of Bay 10_P2 on March 21 with HMM of 70% accuracy

Figure 7 shows the results from bay 10 on the same date. There are several spikes in the true occupancy that the HMM does not detect; these spikes represent a sudden change in the number of occupants in a particular bay for a short duration, for example, a student dropping by a bay for under a minute. Figure 8 shows the results on Bay 13 for a testing date of March 06, 2008 with training on the remaining days of the P1 test period. The model successfully detects periods where nobody is in the space but sometimes with a slight delay that is due to the 20-minute moving average of CO_2 that is used as one of the features. The total accuracy is around 60%. Figure 9 shows the result from Bay 10 on January 29.

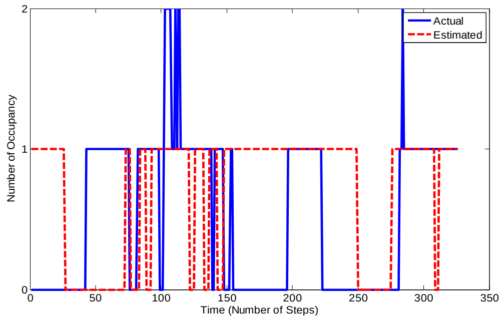

Figure 8 Occupancy Estimation Results of Bay 13_P1 on March 06 with HMM of 60% accuracy

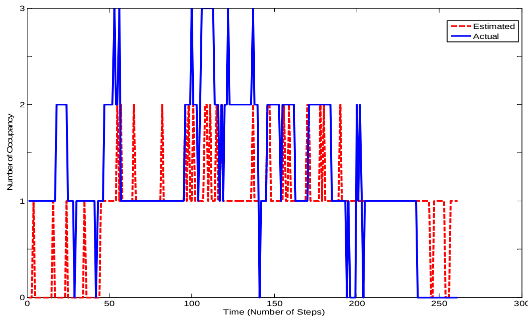

Figure 9 Occupancy Estimation Results of Bay 10_P1 on January 29 with HMM of 58% accuracy

We next tested the HMM estimation approach on longer time periods of one week. Figures 10 and 11 show one week estimation results from bays 13 and bay 10, respectively. The testing period is from period P3 as shown in Table 1 and the training period is obtained by combining the P1 and P2 periods. In total, there are 1156 data points. Accuracies for Figures 10 and 11 are 70% and 65%, respectively. While these numbers appear somewhat low, the profiles illustrate that the estimations track changes in occupancy fairly well. The estimated profiles also present a "smoothed" version of the true occupancy profile.

In summary, the HMM successfully describes the major changes in occupancy while ignoring abrupt fluctuations of short duration. From the perspective of an occupancy-based control scheme, this behaviour is sufficient because the abrupt changes are rather insignificant. Also, it should be noted that the definition of accuracy here is a one-to-one correct mapping for the estimated and actual occupancy numbers. An alternative approach that leads to improved accuracy and a still meaningful result for occupancy-based control is to estimate occupancy ranges (e.g., 0 occupancy, 1-2 occupants, 3-4, etc.).

CONCLUSION

This paper presents the challenges and experience gained from deploying a large-scale sensor network in a test-bed open office environment. The environment closely represents a "real-world' scenario where a plethora of different IT-based systems are typically found in contemporary buildings. Each system requires its own set-up procedures and sensor calibrations as well as communication protocols for both the wired and wireless networks.

Three machine learning methods were investigated for the estimation of occupancy numbers for a typical daily schedule. Our results indicate that, due to the characteristics of the open office plan, CO_2 and acoustic parameters have the largest correlation with the number of occupants in the space. Complications arise when using acoustics, however, because of the affect of sound in adjacent office bays. A Hidden Markov approach to occupancy detection results in estimation accuracy similar to that of a Neural Network approach. However, the HMM model more realistically describes an occupancy presence profile due to its ability to discount sudden brief changes in occupancy levels as well as maintain a constant level during static occupancy periods. Both the daily and weekly results show HMM achieves reasonable tracking of an actual occupancy profile.

Future studies will focus on HVAC control design and operation such as ventilation control based on the results of detected number of occupants. Additionally, although our experience showed reasonable occupancy estimation accuracies with training data sets of 2-4 weeks, further exploration of sufficient training set sizes is needed. Generalization of learned models to other environments (e.g., different buildings or seasons) is also an area of future research.

ACKNOWLEDGMENTS

This work is supported in part by the Bosch Research and Technology Center, Pittsburgh, PA.

REFERENCES

Anderson, B. and Moore, A. 1998. AD-trees for Fast Counting and for Fast Learning of Association Rules. *Knowledge Discovery from Databases Conference.*

Cherkassky V., Ma Y. 2004. Practical selection of SVM parameters and noise estimation for SVM regression, *Neural Networks*, Vol.17, pp113-126.

Chang, C. C. and Lin, C. J. 2001, LIBSVM: a library for support vector machines. *Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm*.

Diane J.C. and Sjal K.D. 2005. *Smart Environments, Chapter 9*. John Wiley & Sons, Inc.

Dodier, R.H., Henze, G.P., Tiller,D.K., Guo, X. 2006.Building occupancy detection through sensor belief networks. *Energy and Buildings*, Vol.28 (9), pp 1033-1043.

Emmerich S.J. and Persily A.K. 2001. "State-of-the-art review of CO2 demand controlled ventilation technology and application", *NISTIR* 6729.

Federspiel, C.C. 1997. Estimating the inputs of gas transport processes in buildings, *IEEE Transactions on Control Systems Technology* Vol.5, pp 480–489.

Lesser V., Atighetchi M., Benyo B., Horling B., Raja A., Vincent R., Wagner T., Ping X., and Zhang S.X. 1999. The intelligent home testbed. *In Proceedings of the Autonomy Control Software Workshop.*

Lam, K.P. Höynck, M. Zhang, R., Andrews, B. Chiou, Y.S. Dong, B. and Benitez, D. 2009. Information-theoretic environmental features selection for occupancy detection in open offices. *Proceedings of Building Simulation 2009* Glasgow, submitted.

Mitchell, T.1997. *Machine Learning*, McGraw-Hill.

Mozer M. 1998. The neural network house: An environment that adapts to its inhabitants. In Proceedings of the *AAAI Spring Symposium on Intelligent Environments*, pp 110–114.

Rabiner, L.R. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, Vol.77 (2), pp. 257–286.

Page, J.Robinson, D. Morel, N. and Scartezzini, J. L. 2007. A generalised stochastic model for the simulation of occupant presence. *Energy and Building*.

Stanislav F., Carlos G., Mark P. and Rahul S. 2006. Distributed Localization of Networked Cameras, *In the Fifth International Conference on Information Processing in Sensor Networks.*

Trivedi, M. Huang K. and Mikic, I. 2000. Intelligent environment and active networks. *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics.*

Torrance M.C. 1995. Advances in human-computer interaction: The intelligent room. *In Working Notes of the CHI 95 Research Symposiu.* .

Vapnik V.N. 1995. *The Nature of Statistical Learning Theory*, Springer.

Wang, S.W. and Jin, X.Q. 1998. CO_2-based occupancy detection for on-lin outdoor air flow control. *Indoor+Built Enviroment.* 1998. Vol. 7 pp 165-181.

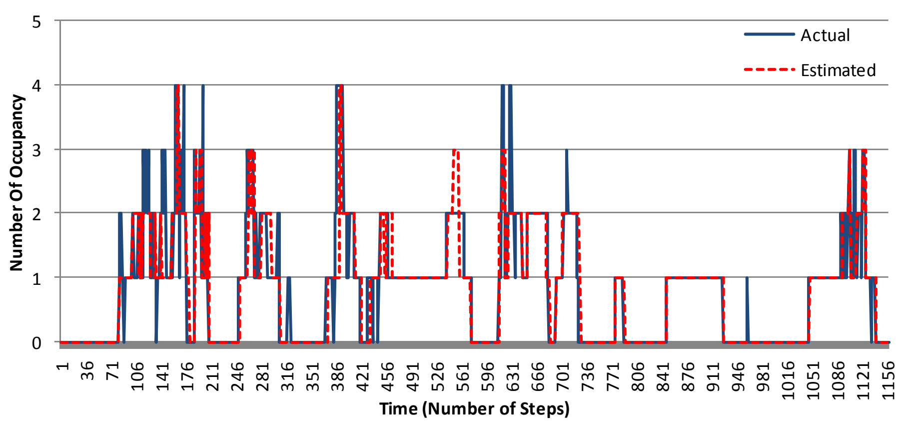

Figure 10 Occupancy Estimation Results of Bay 13 from Results from dataset B13_P3 from March27 to April03 with HMM of 70% accuracy

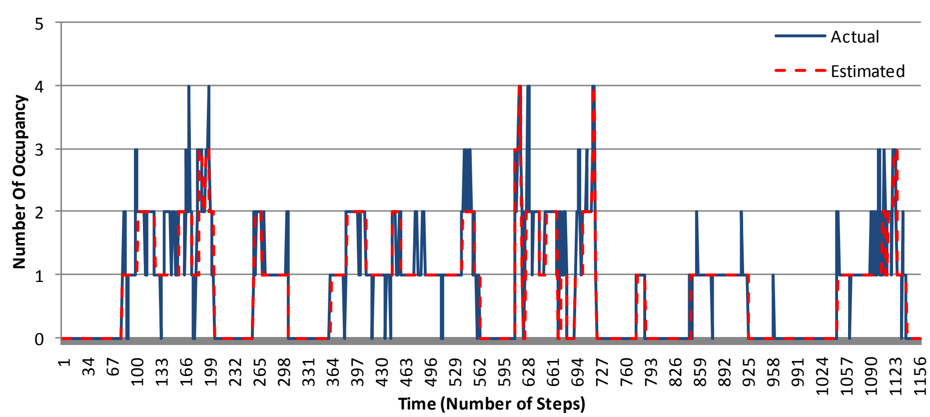

Figure 11 Occupancy Estimation Results of Bay 10 from Results from dataset B10_P3 from March27 to April03 with HMM of 65% accuracy

APPENDIX 1 Sensor network layout of the Intelligent Workplace, Carnegie Mellon University

ITEST Sensor Key	
C	CO2 Sensor
(red square)	CO, CO2, TCOV Sensor
(green square)	Gas Sensor Network Router
(blue diamond)	Temperature, RH, Acoustics Lighting, Motion Sensors
P_x	Power Supply (x = voltage)
D_x	Data Acquisition Unit
S	Data Server
C_x	Camera
P	Pressure Sensor
M	Motion Sensor