Residential House Occupancy Detection: Trust-based Scheme Using Economic and Privacy-aware Sensors

Jun Jiang[a], Chenli Wang[b], Thomas Roth[b], Cuong Nguyen[b], Patrick Kamongi[b], Hohyun Lee[†c], and Yuhong Liu[†a]

[a]Computer Science and Engineering Department, Santa Clara University, Santa Clara, CA, USA
[b]National Institute of Standards and Technology, Gaithersburg, MD, USA
[c]Mechanical Engineering Department, Santa Clara University, Santa Clara, CA, USA
{jjiang2, hlee, yhliu}@scu.edu, {chenli.wang, thomas.roth, cuong.nguyen, patrick.kamongi}@nist.gov

***Abstract*—Internet of Things (IoT) technologies (e.g., power-efficient occupancy-based energy management systems) are increasingly deployed in commercial buildings to reduce building energy consumption. However, the sensors involved in such systems are rarely adopted in residential houses due to their relatively high costs and users' privacy concerns. Low-cost and non-intrusive IoT sensors have been proposed for residential houses for use with machine learning algorithms. Furthermore, such sensors may be triggered very infrequently due to their non-intrusive nature, and it can take several days/weeks to collect sufficient training data. There is a research gap in accurately detecting occupancy information in residential houses with limited training data. This paper proposes a trust-based occupancy detection scheme, which achieves high detection accuracy based on limited training data collected by non-intrusive, low-cost sensors. First, rather than directly taking raw sensor data as inputs, the semantic meanings (i.e., human activity sequences) are extracted from the data based on the order of triggered sensors. Second, the extracted human activity sequences are fed into the proposed trust-based sequence matching scheme for further occupancy detection. Comprehensive experimental results show that, when compared to existing occupancy detection algorithms, the proposed scheme can reliably achieve higher accuracy, especially when only limited training data is available.**

***Index Terms*—Energy Efficient Occupancy Detection, Semantic human activities, Trustworthy decision making, IoT Sensors**

I. Introduction

Internet of Things (IoT) technologies enable data collected by diverse sensors to be shared through networks to facilitate smart decision making [1]. Recently, IoT based occupancy detection, which is defined as the detection of human presence in a specific space, has achieved great success in indoor intelligent systems with various functions, such as health monitoring [2], [3] and smart control [4], [5]. Recent surveys [6]–[8] report that user comfort and energy saving are significantly improved by deploying an occupancy-driven control system. One prior study [9] suggests that accurate occupancy information can help reduce the energy consumption of a building by 10% to 40%.

†: Corresponding author

Many of the existing works for commercial buildings are based on expensive and long-term payback period systems [10], [11], making them hard to be adopted by cost-sensitive residential houses. Moreover, according to [12], privacy is defined as a protected condition, where explicit or inferable information about a person should not be released without permission. The risks of privacy infringement associated with the usage of some intrusive sensors (e.g., video and audio sensors) that may collect highly sensitive data from individual users have raised great concerns, further hindering the adoption of occupancy detection systems in residential houses. The utilization of low-cost, non-intrusive sensors (e.g., temperature sensor and motion sensor) is an essential approach to facilitate the adoption of occupancy detection systems in residential houses, but this area is not well studied in literature.

Most existing occupancy detection studies rely on either intrusive sensors with sophisticated deployment or a sufficient amount of data. The two main types of existing detection algorithms are knowledge-based algorithms and machine learning based algorithms. Knowledge-based algorithms mainly detect occupancy based on the changes of environmental variables, such as CO_2 concentration [13], [14], room temperature and humidity levels [15]. However, accurate measurement of these environmental variables often requires a sophisticated deployment of specific sensor types to precise locations in the building, which makes measurement difficult in a diverse range of residential houses with different floor plans, orientations and materials. Machine learning based algorithms [16], [17] are gaining increasing popularity since they can achieve good performance without relying on specific deployment of sensors and in-depth domain knowledge. These schemes, nevertheless, require sufficient training data to make accurate decisions, which may take considerable time for data collection in residential houses where changes in occupancy are infrequent.

This work proposes the use of low-cost, non-intrusive sensors with limited training data for occupancy detection. However, using the limited amount and types of human activities captured in residential houses by the low-cost, non-intrusive sensors results in several major challenges in training data, including imbalance, limited size and lack of human

activity diversity. (1) The collected data are highly imbalanced because most sensor signals are reflecting "silent time" when human activities are not captured. To address the highly imbalanced data, sensor data is pre-processed to determine the windows of time with human activity in the house. A human activity sequence extraction module extracts "human activities" from the sensor data and organizes the "human activities" into "event windows", so that the later data analysis can ignore the silent time. (2) Achieving high detection accuracy through the limited training data requires full utilization of limited information and not adding extra noises in each human activity sequence. This requirement may lead to human activity sequences of variable length, which cannot be used as a direct input for most machine learning based algorithms. This work addresses this challenge by proposing a sequence matching algorithm, which allows the activity sequences with variable length as input data. (3) Due to the diversity of human activities, some new activity sequences may not be captured by the limited training data, leading to high indeterminacies. To overcome this challenge, this work proposes a trust-based scheme, which effectively reduces data "indeterminacy" by considering label confidences, common activity ratio and the significance of different activities.

The main contributions of this paper are as follows.

First, rather than directly using the imbalanced raw sensor data collected by low cost and non-intrusive sensors, the human activity sequence extraction method is proposed to extract the semantic meanings of the sensor data as "human activity sequences". This effectively resolves the data imbalance issue in the raw sensor data. Furthermore, it makes the proposed scheme independent from the underlying sensor types and thus more easily applicable to different residential houses.

Second, a sequence matching algorithm, which can handle the human activity sequences with variable lengths, is proposed to fully utilize limited training data for occupancy detection. Experiment results show that the proposed algorithm achieves high detection accuracy by avoiding information loss caused by sequence truncating and additional noises caused by sequence padding.

Third, a novel trustworthy sequence matching algorithm is introduced, which considers "indeterminacy" as an important aspect in the decision making process. Experiment results show that the introduction of trust factors significantly improves the occupancy detection accuracy, especially when the training data size is limited.

The rest of this paper is organized as follows. Section II discusses the state-of-the-art related works. Section III introduces the proposed scheme. The experiment set-up and data collection are presented in Section IV. The experiments results are discussed in Section V, followed by a conclusion in Section VI and acknowledgment in Section VII.

II. Related Work

A. Occupancy Detection Sensors

Different types of sensors have been proposed and adopted for occupancy detection. In [13], the authors deploy CO_2 concentration sensors for data collection in five residential and office rooms with different structures and sizes. In [18], the authors monitor CO_2 concentration in the $44\ m^3$ conference room by using fifteen CO_2 concentration sensors. In [19], the authors propose a plug-and-play occupancy detection method based on the trajectory data collected by six indoor sensors, including temperature, humidity, noise, passive infrared (PIR), volatile organic compound (VOC) and CO_2 sensors, in a single person office. The authors in [20] collect the data by deploying fifty Building Level Energy Management System (BLEMS) sensor nodes in a three-story educational building. Each BLEMS sensor node consists of lighting, sound, reed, CO_2, temperature and PIR sensors, which can report three types of variables, including instant variables, count variables and average variables. In addition, authors in [21], [22] use one and four cameras, respectively, to collect the occupancy information. However, the involvement of these intrusive sensors may expose the sensitive data (e.g., faces, talks, and private activities) from individual users. In addition, some sensors are also expensive, limiting their adoption by residential houses.

B. Occupancy Detection Schemes

Many knowledge-based algorithms have been proposed to determine occupancy status based on physical relationships between human activities and the environment. For example, dynamic mass balance equation [13], steady-state mass balance equation [14], partial differential equations and ordinary differential equations [18] are proposed to model the occupancy status. Moreover, some studies generate occupancy detection rules based on specific domain knowledge. In [19], user-defined thresholds and the tolerated lag between sensor data are adopted to set different rules to estimate the probability of occupancy status for impulse-based sensor data (i.e., PIR and noise sensors). In [23], a threshold-based approach is utilized to detect occupancy where the threshold is determined by observing the variations of CO_2 concentration levels in state of occupation and absence. This type of approach heavily relies on domain knowledge to formulate the relationship between occupancy status and environmental changes. More importantly, these algorithms may only work with specific types of sensors, limiting their generality to be applied in diverse residential houses with different types of sensors.

Recently, machine learning based algorithms are frequently adopted for occupancy detection as these algorithms can be generally applied on data collected from diverse sensors. Support Vector Machine (SVM) with different kernel functions, such as the radial basis function (RBF) kernel [14], and the transfer kernel learning (TKL) [24], have been proposed for occupancy detection. In addition, decision tree (DT) [25], [26], Bayesian network based approach [17], Artificial Neural Network (ANN) [27], Logistic Regression (LR) [28], and other machine learning models are also applied for occupancy detection. In [29], the authors compare the performances of three occupancy detection algorithms, including Linear Discriminant Analysis (LDA), Classification And Regression Trees (CART), and Random Forest (RF). In [30], the authors use truncation of the raw sensor data as the input of the machine learning models. Specifically, this work significantly

improves the accuracy of several popular machine learning approaches, including decision tree, random forest, KNN and SVM models by introducing a two-layer detection scheme. According to a latest review paper [7], although machine learning based algorithms achieve highly accurate occupancy detection results, most of them require sufficient training data collected from multiple types of sensors, leading to an increased training time and system complexity.

As sensor data collected over time can be considered as time sequences, a few studies also adopt sequence matching algorithms for occupancy detection. In [31], the CO_2 concentration is measured and recorded as time series. Distances between segments of the time series are calculated based on different models, such as Euclidean, city-block, Cranberra and angular distance. Sequence matching based on such distances is adopted to make occupancy decisions. However, as the raw sensor data collected in residential houses may generate very long sequences containing various noises, directly applying sequence matching based approaches on such data often yields very limited accuracy and efficiency, and is therefore adopted by very few studies. Therefore, this work proposes to first pre-process the raw sensor data to extract their semantic meanings (i.e., human activity sequences), and then match these sequences based on different trust factors.

Trustworthy computing, which has been widely adopted in wireless sensor networks [32], reputation based online social media [33], biomedical applications [34], and Vehicular Ad Hoc Network (VANET) [35], provides a probability-based solution to dynamically estimate indeterminacies. For example, Beta trust model [33], fuzzy models [36] and Dempster-Shafer theory [37] are often used for trust evaluations. Trust-based schemes have been recently adopted in cyber-physical systems to evaluate the indeterminacies of system components or sensor signals [38], [39]. In this work, there are three types of potential indeterminacies for the accurate occupancy detection: (1) the data collected by only a few low cost and non-intrusive sensors often lack reliability; (2) the limited size of training data may contain sparse representations of each activity sequence, resulting in indeterminacies in the label accuracy; and (3) the limited training data may not necessarily include all possible activity sequences, causing indeterminacies in the labeling of testing sequences that never appear in the training data. Trust model can calculate the most trustworthy activity sequence in training data by introducing different trust factors to minimize indeterminacies. To the best of the authors' knowledge, this is the first work using trust models to minimize indeterminacies in occupancy detection.

III. Proposed Scheme

A. System Overview

This work proposes to perform occupancy detection in residential houses based on limited training data. Figure 1 shows the inputs, outputs and major modules of the proposed scheme.

The input is the raw data collected by a few low-cost, non-intrusive sensors, which can be either continuous or binary values. The experimental setup uses four temperature sensors and one motion sensor. The data collected from temperature sensors have continuous values from the measured temperature at the sensor location. Depending on the specific sensor location, the measured temperature variations can provide information about certain types of human activities, such as door opening and water usage. Meanwhile, the data collected from the motion sensor are binary values indicating whether motions has been detected.

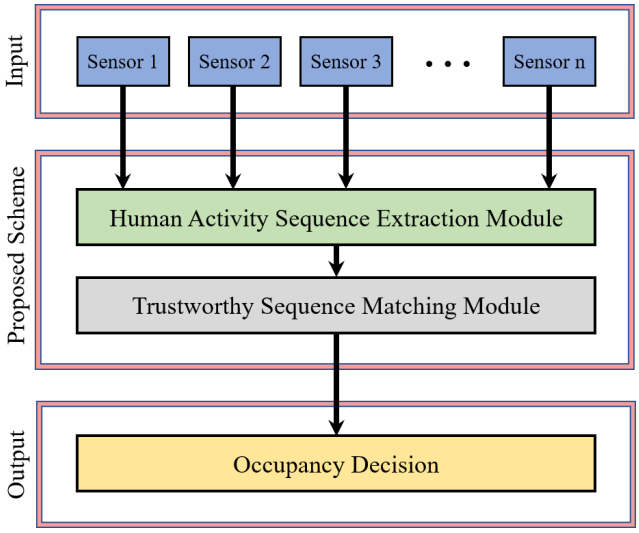

Fig. 1: The Proposed Occupancy Detection Scheme

These raw sensor data are fed into the human activity sequence extraction module, which is the first module of the proposed scheme. This module aims to extract the semantic meanings (i.e., human activity sequences) from the raw sensor data. The first step is to identify specific types of human activities (e.g., water usage, door handle touch) based on the raw sensor data. Each specific combination of these activities may indicate a specific occupancy status or change in occupancy. Therefore, a second step orders these detected activities into event windows according to the order of their occurrence in time. The activities occurring in each event window will form an activity sequence. These sequences are further trimmed to avoid redundant information.

The second module is a trustworthy sequence matching module, which takes the detected human activity sequences as inputs to determine the occupancy status. For matching a new activity sequence to an existing activity sequence, the matching decisions are made based on the likelihood of these two sequences sharing the same label.

The outputs of the system are occupancy decisions, which are based on "someone entering the house", "someone leaving the house" and "someone in the house".

Furthermore, deploying occupancy detection systems in residential houses often raise privacy concerns from two aspects: (1) whether sensitive visual or audio information is recorded, and (2) whether personal identifiable information, such as name and age, is collected. The proposed scheme addresses these privacy concerns as follows. First, to prevent private talks and activities from being recorded, intrusive sensors, such as web-cameras [22], [40] and voice-detection sensors [26], [41],

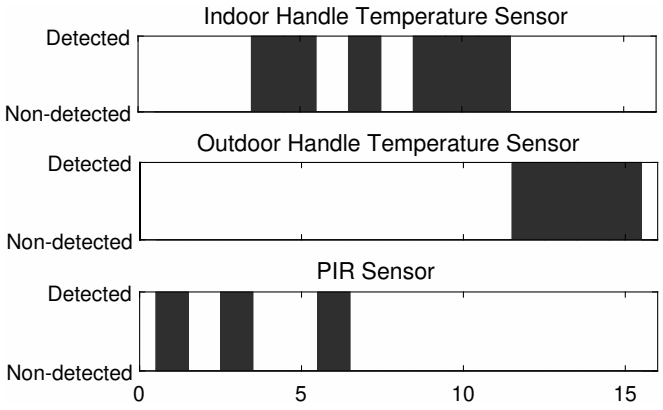

Fig. 2: Example of Binary Sequences Retrieved from Raw Sensor Data

are excluded. Second, personal identifiable information is not collected in this work. Third, the collected data are not stored in their raw format, but in the form of window-based activity sequences, which does not include the specific time/date of the collection. As a result, the identity of a specific resident cannot be uniquely determined from the collected activity sequences.

B. Human Activity Sequence Extraction

There are three main challenges to effectively and efficiently making accurate occupancy detection based on the raw sensor data. First, since the limited human activities can be detected, these sensors are not triggered for most of the time in residential houses, leading to high data imbalances. Second, temperature amplitudes measured by sensors can be easily influenced by environmental factors (e.g., weather and seasons) and human factors (e.g., specific human body temperatures), requiring additional data processing. Third, the raw data collected from different sensors may have diverse formats, such as binary values, continuous values or discrete values, and are difficult to handle by a single model.

To address these challenges, the proposed scheme extracts the semantic meanings from sensor raw data through the human activity sequence extraction module. In particular, this module mainly focuses on the time duration when human activities are detected; normalizes the collected temperature amplitude changes by eliminating the impact of environmental and human factors; and converts different types of sensor signals into a consistent format as binary values to indicate if a specific type of human activity occurs. As a result, the later occupancy decisions based on the extracted human activities can be independent of the underlying sensors, and thus can be generally applied in diverse residential houses.

1) Human Activity Detection: This work adopts two types of sensors, temperature sensors and passive infrared sensor (i.e., PIR). The prior work [42] reveals that temperature changes can indicate human activities. For example, a human "hand touch" on a door handle can lead to changes in the door handle's temperature. Besides, since water temperature from ground is different from the temperature of water staying within pipes, the water pipe temperature will change when someone is using the water. Therefore, four temperature sensors are installed to collect the temperatures of the outdoor and indoor handles of the main entrance, the water temperature of the faucet and toilet pipes, respectively. In addition, one PIR sensor is installed above and inside the main entrance to record "human movement" activities close to the main entrance.

However, the collected sensor data are in different formats, which requires separate processes to retrieve human activities. The data collected by temperature sensors are continuous signals indicating the real time temperature. To retrieve human activities, a transient and steady-state heat transfer model developed in prior work [42] normalizes temperature data to exclude environmental factors and human factors, and reliably detects human activities when the normalized temperature change exceeds a threshold value for a certain duration. The output is a sequence of binary values with timestamps, where '1' indicates that a human activity (e.g., hand touch or water usage) has been detected. Furthermore, based on their locations, the four temperature sensors can detect the "outdoor touch", "indoor touch", "faucet usage", and "toilet usage" activities, respectively. On the other hand, the PIR sensor directly outputs a sequence of binary values with timestamps, indicating if a "human movement" activity has been detected within the area covered by the sensor. Figure 2 shows an example of binary sequences retrieved from the raw data of three different sensors. The y-axis and x-axis of the sub-figure in Figure 2 represent whether human activity is detected by the sensor and timestamp, respectively. The black bar represents the detection of human activity by specific sensor. The width of the black bar indicates the duration of human activity being detected.

2) Event Window: As discussed earlier, a significant portion of the sensor data is collected when no human activities are captured. Consequently, the output binary sequence from the human activity detection sub-module is dominated by zeros, causing high data imbalances. To address this issue, the data is split into event windows, which mainly focus on the time duration when human activities occur.

Specifically, each event window starts from a detected human activity, and lasts until no more human activities are detected in a certain duration. In this work, this duration is 30 seconds since in a residential house, two activities detected 30 seconds apart are likely to be non-consecutive activities. Then each event window may contain one or multiple activities, as an active user may sequentially trigger multiple sensors in a short time period. For example, when someone enters the house, they may touch the outdoor handle first, the indoor handle next, and then trigger the PIR sensor.

In addition, occupancy changes are related to not only the types of human activities occurring in an event window, but also the sequential order in which the different human activities occur. For example, when a person leaves the house, the sensor on the indoor handle is triggered before the sensor attached to the outdoor handle. Therefore, the activities occurring in each event window are ordered based on their timestamps to form a sequence of human activities. For example, assume that all the activities shown in Figure 2 occur in one event window, the aggregated activity sequence can be marked as 'PPIIPIIIIOOOO', where 'I', 'O' and 'P' represent an indoor handle touch activity, an outdoor handle

touch activity and a movement activity, respectively.

3) Removing Consecutive Repetitive Activities: In an event window, one activity may sometimes be detected multiple times consecutively (i.e., without other activities detected in between). This can be caused by different reasons. For example, according to different users' door opening habits, the touch duration and intensity may vary, which may possibly lead to multiple consecutive detection of the touch activities. Also, according to different configurations of the PIR sensor sensitivity level, a user's movement with a longer duration may trigger the PIR sensor multiple times. However, as long as these repetitive activities are detected in a short duration (e.g., no more than a few seconds) and no other activities are detected in between, the repetitive occurrence of one activity will not change the occupancy status. Therefore, such consecutive repetitive activities are removed from the activity sequence. As a result, the activity sequence 'PPIIPIIIIOOOO' shown in Figure 2 will be converted into 'PIPIO'.

In summary, by introducing the human activity sequence extraction module, the underlying heterogeneous and imbalanced raw data are converted into more general human activity sequences as inputs for the next trustworthy sequence matching module.

C. Trustworthy Sequence Matching

This section discusses the trustworthy sequence matching, which makes occupancy decisions based on the above extracted human activity sequences. Specifically, there are two challenges. The first challenge is that the extracted human activity sequences have variable lengths. Most machine learning models require fixed length inputs, and using padding or truncation techniques to generate fixed length inputs may change the physical meanings of human activity sequences or lose important information. The second challenge is the various indeterminacies caused by limited training data size. In particular, the detection of human activities from limited raw sensor data may not be fully reliable, especially for the cases when the environmental temperature changes significantly (e.g., extremely cold or hot days) or when multiple occupants have very different body temperature (e.g., very low or very high). In addition, due to the diversity of human activity sequences, some activity sequences may only occur once or twice in the limited training data (i.e., data sparsity), which leads to indeterminacies in the accuracy of ground truth labels. Furthermore, the limited training data may not necessarily include all possible activity sequences, causing indeterminacies in the labeling of testing sequences that never appear in the training data.

Therefore, the trustworthy sequence matching module is needed to reliably match new sequences to existing ones without any length requirements. More importantly, the trustworthiness of different matching options is evaluated and the one with the minimum indeterminacies is selected. Specifically, three aspects are considered as (1) similarity represented by the longest common sub-sequence shared by two sequences, (2) differences represented by significant human activities in non-overlapping part of activity sequences, and (3) label confidence represented by accuracy of labels for existing activity sequences. The first two factors focus on the correlations between the new activity sequence and existing activity sequences, while the last factor focuses on the probability of an existing activity sequence to be assigned its current label.

1) Common Activity Ratio: As two sequences sharing more common activities are more likely to share the same label, the first trust factor is the common activity ratio, which is mainly determined by the longest common subsequence shared between two sequences. A higher ratio between two sequences indicates higher confidence that these two sequences share the same label. Specifically, activities are only counted as common if they follow the same order because the order of a series of human activities may indicate certain physical meanings. For example, given two activity sequences 'IPO' (i.e., indicating leaving events) and 'OPI' (i.e., indicating entering event), although both sequences contain 'I', 'P', and 'O', the number of common activities is counted as one due to their completely reversed order. In addition, the activities in the longest common subsequence do not need to occupy consecutive positions in the original sequence. For example, the longest common subsequence of the activity sequence 'FPIPO' and the activity sequence 'PIO' is 'PIO'. The non-overlapping activities 'F' (i.e., faucet usage) and 'P' (i.e., human movement) will be considered later by the next trust factor.

The common activity ratio is calculated as follows,

$$f_c = \frac{L_0}{L_1} \tag{1}$$

where L_0 is the length of the longest common subsequence shared by the two sequences. It is normalized by L_1, the length of the longer sequence between the two, so that the value of f_c is in the range of $[0, 1]$. For example, given two activity sequences 'PIPIO' and 'PIPO', the longest common sequence 'PIPO' leads to $f_c = \frac{4}{5}$.

2) Significance of Non-overlapping Activities: While the first factor considers the commonly shared part between two sequences, the second factor mainly focuses on the non-overlapping part. Specifically, it considers whether any activities in the non-overlapping part are statistically "significant" to merit different labels for the two sequences. If more significant human activities appear in the non-overlapping part, there may be greater differences between the two, and thus lower probability for them to share the same label. However, quantitatively evaluating the significance of each activity in the non-overlapping part is a challenging task. The significance value of an activity is calculated based on information entropy, which is often used to measure the indeterminacy of variables [43].

Specifically, given an existing sequence with label j and a new sequence, assume that activity i is one activity in the non-overlapping part of the two sequences. First calculate the probability of activity i to be assigned label j as follows,

$$p(x_{i,j}) = \frac{N_{i,j}}{N_i} \tag{2}$$

where N_i represents the total number of activity sequences containing activity i; and $N_{i,j}$ represents the number of

activity sequences containing activity i that are labeled as j. Then the entropy of activity i is calculated as follows.

$$H(x_{i,j}) = -\,p(x_{i,j}) \times \log_2 p(x_{i,j}) - (1 - p(x_{i,j})) \times \log_2(1 - p(x_{i,j})) \tag{3}$$

Here, the entropy value has different meanings for different $p(x_{i,j})$ values. When $0.5 \leq p(x_{i,j}) \leq 1$, a higher entropy value represents a higher indeterminacy that activity i leads to label j, indicating that a higher significance value should be assigned to i. On the other hand, when $0 \leq p(x_{i,j}) \leq 0.5$, a higher entropy value represents a higher indeterminacy that activity i leads to other labels, indicating that a lower significance value should be assigned to i. Therefore, significance value of a sensor activity i (i.e., $S_{x_{i,j}}$) is defined as follows.

$$S_{x_{i,j}} = \begin{cases} H(x_{i,j}) - 1, & for \quad 0.5 \leq p(x_{i,j}) \leq 1 \\ 1 - H(x_{i,j}), & for \quad 0 \leq p(x_{i,j}) < 0.5 \end{cases} \tag{4}$$

$S_{x_{i,j}}$ is a continuous value in the range $[-1, 1]$. When activity i never occurs in sequences with label j, the associated $p(x_{i,j})$ is zero. When activity i always occurs in sequences with label j, the associated $p(x_{i,j})$ is one. When $p(x_{i,j}) = 0$, Equation 4 resolves to $S_{x_{i,j}} = 1$ indicating the highest significance in causing different labels for the two sequences. When $p(x_{i,j}) = 1$, Equation 4 resolves to $S_{x_{i,j}} = -1$, indicating that activity i negatively contributes to causing different labels for the two sequences. As the value of $S_{x_{i,j}}$ increases, the two sequences based on activity i are more likely to share different labels. The $S_{x_{i,j}}$ is further normalized as follows.

$$S'_{x_{i,j}} = \frac{S_{x_{i,j}} + 1}{2} \tag{5}$$

where $S'_{x_{i,j}}$ is ranging between [0, 1].

Next, as the normalized significance value $S'_{x_{i,j}}$ negatively influences the trust value for assigning the same label j, $1 - S'_{x_{i,j}}$ is used to calculate the trust value of activity i. Furthermore, since the activities in the non-overlapping part of activity sequences are independent, the significant value of all non-overlapping activities are concatenated to calculate the second trust factor f_n:

$$f_n = \prod_{x_i=1}^{t} (1 - S'_{x_{i,j}}) \tag{6}$$

where t represents the total number of activities in the non-overlapping part of two activity sequences. If the two sequences are exactly the same, the value of trust factor f_n is obviously 1. The activity with larger value of $S'_{x_{i,j}}$ in the non-overlapping part leads to a smaller value of trust factor f_n. Furthermore, the more activities involved in the non-overlapping part, the lower probabilities that the two sequences share the same label, and thus the lower trust factor value would be.

3) Label Confidence: Higher trust values from the first two factors show the correlation between the new and existing activity sequences. Beyond these two factors, the confidence on whether the existing activity sequence should be assigned its current label is also important. Due to the complexity of human activities and the limited size of training data, some outlier activity sequences may occur and lead to non-unique labels for identical activity sequences. Therefore, the third factor models confidence for an activity sequence to be assigned to the most possible label.

TABLE I: The Physical Meanings of Different Trust Factors

Symbol	Equation	Physical Meaning
f_c	$\frac{L_0}{L_1}$	The common activity ratio between two activity sequences
f_n	$\prod_{x_i=1}^{t}(1 - S'_{x_{i,j}})$	The significance of non-overlapping activities
f_l	$\frac{\alpha+1}{\alpha+\beta+2}$	The label confidence

In particular, such confidence is modeled through a Beta trust model [33], which is a Bayesian based statistical model. Given an activity sequence, multiple independent observations can be made about its labels as either j or other labels. Each time, if the probability of observation for label j is θ, where $0 \leq \theta \leq 1$, the probability of observing label j for k times and other labels for m times is shown as follows.

$$P(k, m|\theta) = \binom{k+m}{k} \theta^k (1 - \theta)^m \tag{7}$$

θ is assumed to follow a Beta distribution [44], which is a family of continuous probability distributions defined on the interval $[0, 1]$ and parameterized by two positive shape parameters, denoted by a and b. Then,

$$P(\theta = y; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1 - y)^{b-1} \tag{8}$$

where Γ denotes the gamma function. The probability expectation value of the Beta distribution is shown as follows.

$$E(\theta) = \frac{a}{a+b} \tag{9}$$

Note that, when there are no observations, $a = 1$ and $b = 1$.

Therefore, given a specific activity sequence, the confidence of assigning each possible label j is calculated as follows,

$$f_l = \frac{\alpha + 1}{\alpha + \beta + 2} \tag{10}$$

where α represents label j's occurrence number; and β represents other labels' occurrence numbers. When the label j occurs more frequently than other labels for a specific activity, the larger value of trust fact f_l, which indicates the higher accuracy for this activity with label j, can be calculated.

The three trust factors are summarized in Table I. By combining the above three factors, the overall trust value of mapping the new activity sequence to an existing activity sequence is calculated as follows.

$$T = w_c f_c + w_n f_n + w_l f_l \tag{11}$$

Here, w_c, w_n and w_l are the weights for the three trust factors respectively. The best matching activity sequence can be chosen as the one with the highest trust value. Then its label will be used as the label for the new activity sequence.

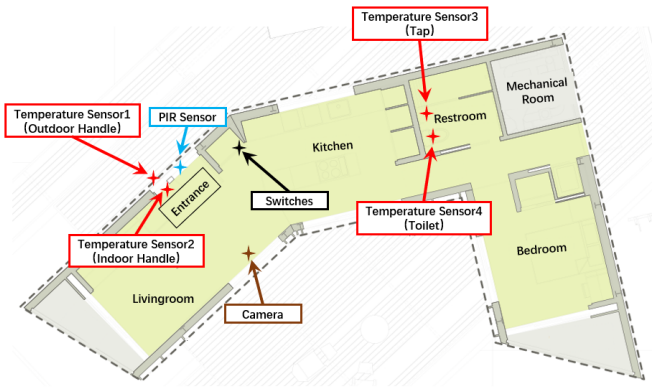

Fig. 3: Floor Plan of Laboratory and Sensor Distribution

IV. Experiment Set-up and Data Collection

A. Experiment Set-up

The experiments in this work are conducted in the Solar House laboratory of Santa Clara University, whose area is 65 m^2 and maximum capacity is five. The laboratory room consists of a living room, a kitchen, a bedroom, a bathroom and a mechanical room. In the experiments, it is assumed that occupants close the door when entering or leaving the room. The floor plan of the laboratory room and the sensor distribution are shown in Figure 3. There are five non-intrusive, low cost sensors, including four temperature sensors and one PIR sensor. A typical occupancy detection system [28], [45] often requires deploying 10 to 15 sensors in around 30 m^2 to 50 m^2 area. In addition, the costs of the temperature and PIR sensors adopted in this work are very cheap. For example, a CO_2 sensor, which is a typical sensor adopted in existing studies [28], [45], has a retail price as around $70 to $100. The price is over 70 times higher than that of the temperature/PIR sensors adopted in this work.

In particular, two temperature sensors are installed on the inside and outside handles of the entrance door. Two temperature sensors are installed on the water pipe surface of the faucet and toilet in the restroom. A PIR sensor is deployed on the indoor frame of the lab entrance to collect activities near the inner door.

Furthermore, five switches are installed on the wall near the entrance. Each switch records the ground truth (i.e., leaving or entering the lab) of a specific occupant. A camera is deployed near the entrance as a backup to double check the ground truth accuracy for the case where someone forgets to turn on/off their switch. Please note that this camera is not required by proposed scheme and is only used for ground truth validation. The temperature data are measured and recorded by LabView [46] with an NI Compact DAQ. The motion data and ground truth are collected with a system-on-a-chip, Beaglebone Black [47].

B. Data Collection

The data and ground truth in the lab room are collected through 40 days from 10/18/2019 to 01/22/2020, excluding holidays and some accidents (e.g., power outages). After feeding the collected sensor data to human activity sequence extraction module, there are 324 activity sequences, including 36 unique activity sequences, as shown in Table II. The first column of Table II shows the component of each unique activity sequence with its corresponding label, and the second column shows the corresponding number of occurrences. For example, in the second row, the activity sequence 'PIP' represents the order of triggered sensors, which is 'PIR sensor → Indoor Handle sensor → PIR sensor'. Moreover, the activity sequence 'PIP' has two labels, namely '1' and '-1', which appear two times and eighty times respectively.

TABLE II: Collected Activity Sequence with Corresponding Label and Occurrence Times

Activity Sequences[1](Label[2])	Occurrence Times
PIP (1, -1)	2: (1), 80: (-1)
P (2)	73
OP (1)	52
T (2)	17
TF (2)	16
IOP (1)	14
OPOP (1)	12
IOPOP (1)	9
PIPOP (-1), FTF (2)	7
F (2)	4
FT (2), OPIPOP (1), PIPIOP (-1), IPOP(1), PIPI (-1), FTFPT (2)	2
OPFT (1), FTFT (2), OPO (1), TPIP (-1), TFPIP (-1), FPIP (-1), OPIOP (1), TFT (2), PIOP (-1), PIPIPIPIPI (-1), TP (2), PTP (2), FTFPTPTP (2), FP (2), IOPO (1), OPTFPFPFPTPIPT (1), FTFPITPT (-1), IOPIPIOP (1), PFPTF (2)	1

[1] Sensor: P: PIR, I: Indoor Handle, O: Outdoor Handle, F: Faucet, T: Toilet
[2] Label: 1: Entering, -1: Leaving, 2: Someone in the house

V. Results Analysis

This section evaluates the performance of the proposed scheme. In particular, the effectiveness of the major modules of the proposed scheme is tested and then compared with other state-of-the-art algorithms. In all these experiments, F1-score is adopted as the performance metric, which can be regarded as the harmonic mean of precision and recall.

A. Performance of Human Activity Sequence Extraction

In this section, the necessity and effectiveness of human activity sequence extraction module are validated by three experiments. In the first experiment, the random forest model, one of the most effective machine learning algorithms adopted in existing works [48], is directly applied to the raw sensor data. The raw sensor data include the real temperature values measured by the temperature sensors, as well as the binary values collected by the PIR sensor indicating if a movement has been detected. The raw data is highly imbalanced as there are no human activities captured between 8:00 pm and 8:00 am in the lab. Although this time duration is specific for our lab, similar sequences can also be observed in residential houses when people go to work during daytime, or conduct no activities during sleeping time. As a result, only a very small portion of the raw data contains occupant related sensor signals.

In the second experiment, human activities are extracted from the raw sensor data and organized in event windows for

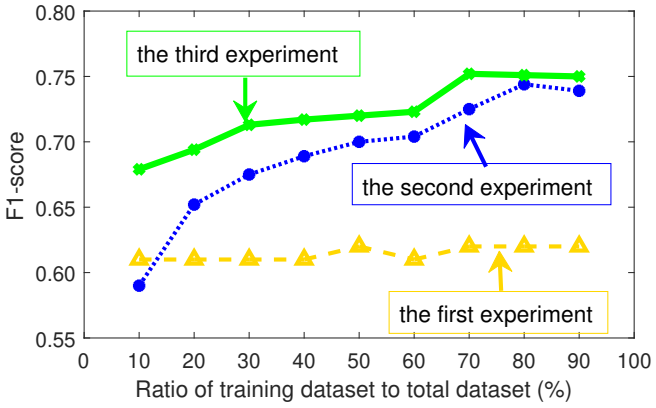

Fig. 4: Effectiveness validation of the human activity sequence extraction. The first experiment represents the random forest model with raw data. The second experiment represents the basic sequence matching with human activity detection and event window. The third experiment represents the basic sequence matching with human activity detection, event window and consecutive repetitive activities remove.

occupancy detection. Consecutive repetitive events detected by a single sensor are not removed from these windows. A basic sequence matching approach is applied to determine the occupancy as the human activity sequences with variable lengths cannot be fed into machine learning algorithms directly. The basic sequence matching does not consider any trust factors, and can only assign a label to a testing sequence when the testing sequence and an existing sequence are exactly the same. Otherwise, the testing sequence will be assigned an 'unknown' label.

In the third experiment, the full human activity sequence extraction module is introduced by further removing any consecutive repetitive events from the event windows. The same basic sequence matching is then applied to identify labels for testing sequences.

Figure 4 shows that the third experiment achieves the highest detection accuracy regardless of the training data size, significantly than that of the first experiment. It shows the effectiveness of the proposed human activity sequence extraction module.

The first experiment takes the highly imbalanced raw sensor data as inputs, in which the meaningful sensor data reflecting occupancy status is overwhelmed by a large amount of non-occupancy noises due to limited human activities detected. Such heavily imbalanced data can lead to low recall values, which further leads to a low F1-score. More importantly, such F1-score cannot be improved by increasing the size of the training data [49]. Therefore, the performance of the random forest model with raw data remains low.

Furthermore, by comparing the results of the second and third experiments, the effectiveness of removing consecutive repetitive activities is validated. In particular, when the training data size is small (e.g., 10% - 20%), the F1-scores of the second experiment are much lower than that of the third experiment. The reason is that when repetitive sensor events are allowed (i.e., the second experiment), the same activity may be expressed in different repetitive formats when it lasts for different duration, or when the sensor sensitivity level changes, resulting in more diverse patterns of activity sequences. When the training data size is very limited, such diversity will cause additional challenges to match a testing sequence to a training sequence, leading to significant accuracy drops. When the training data size gradually increases, the offset between the results of the two experiments becomes smaller.

B. *Effectiveness of Trust Factors*

In this section, the impact of each trust factor on the occupancy detection accuracy is evaluated. The data set used in this section is pre-processed by the human activity extraction module. The following experiments evaluate the performances of (1) schemes with only one trust factor, (2) schemes with two trust factors, and (3) schemes with all the three trust factors.

The experiment results are shown in Figure 5, Figure 6 and Figure 7. In each figure, the y-axis represents the F1-score, and the x-axis represents the training data size, ranging from using 10% to 90% of the collected data for training. Different curves in each figure represent schemes considering different trust factors. In particular, f_c, f_n and f_l represent the sequence matching schemes considering common activity ratio, significance of non-overlapping activities and label confidence, respectively.

1) Effectiveness of Single Trust Factor: Figure 5 shows the comparison of F1-score for schemes with only one trust factor. In addition, the basic sequence matching scheme (i.e., the 3rd experiment in Figure 4) is also shown in Figure 5 as a reference.

In Figure 5, the two schemes with f_c (common activity ratio) and f_n (significance of non-overlapping activities) achieve similar performances, indicating that both of these two trust factors effectively contribute to the correlation evaluations between the two compared activity sequences. When either of these two factors increases, there is a higher confidence that the two compared sequences will share the same label. Specifically, the scheme adopting only f_n yields better performance when the training data size is small (i.e., 10%). This is because limited training data can include only limited sequences, making it difficult to label testing sequences based only on commonly shared parts. Better decisions can be made by relying more on whether the non-overlapping part is significantly different or not. As the training data size increases, the performance of the f_c factor can exceed that of f_n, indicating that when more diverse sequences are trained, there is a higher chance for a testing sequence to be labeled correctly based on the commonly shared parts.

On the other hand, the basic sequence matching scheme yields much lower F1-score when compared to the scheme with f_c and the scheme with f_n. This is because this scheme can only label the testing sequence that exactly matches an existing training sequence. Moreover, as the training data size grows, this scheme yields higher F1-score since more sample sequences can be included in the training data, reducing the 'unknown' labels for testing sequences.

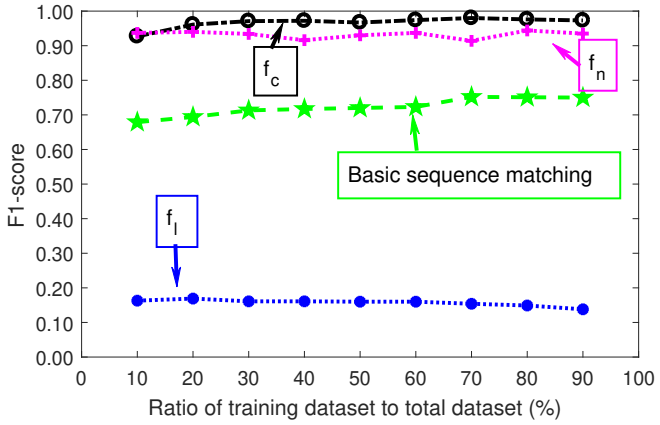

Fig. 5: Comparison of effectiveness of single trust factor

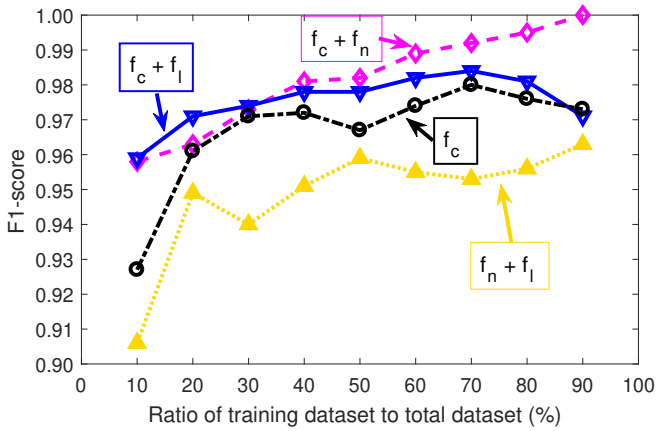

Fig. 6: Comparison of effectiveness of combination of two trust factors

Last, the scheme considering only the factor f_l achieves the lowest F1-score. This is because this factor mainly focuses on assigning the most trustworthy label to a sequence, while ignoring the relationship between the training and testing sequences. It also shows that the factor f_l cannot be used independently. Although the F1-score of scheme with single factor f_l stays low, the combinations of factor f_l and other factors can further achieve better performances, which are analyzed in the following experimental results.

2) Effectiveness of Combing Two Trust Factors: Next, the F1-score for all combinations of two trust factors are compared in Figure 6. The scheme with f_c, the highest performing single trust factor, is also included as a reference. When f_c is combined with f_n or f_l, the performances of the combined two factors are better than that of single f_c regardless of the training data size. This observation validates that on top of f_c, both f_n and f_l can provide additional information to match the testing sequences to more trustworthy sequences.

In addition, compared to the scheme with $f_c + f_n$, the scheme with $f_c + f_l$ achieves higher F1-scores when training data size is small (i.e., 10% - 30%), but lower F1-scores when the training data size is large. This is because f_n is calculated based on entropy, which tends to be more reliable when the training data size increases. While compared to f_n, f_l is less sensitive to the training data size.

On the other hand, because the combination of f_l and f_n does not consider the commonly shared part between two sequences, which was determined to be the most significant aspect to match two sequences in Figure 5, the combination of f_l and f_n does not perform as well as the other combinations.

3) Effectiveness of Combining Three Trust Factors: Figure 7 compares the combination of three trust factors and the best schemes in Figure 5 (i.e., f_c) and Figure 6 (i.e., $f_c + f_n$). It is obvious that the combination of three trust factors shows the best performance from small training data size to large training data size. It validates that the combination of three trust factors is more comprehensive and reliable than the best schemes with a single trust factor or the combination of any two trust factors. In other words, each of the three factors contributes to the overall accuracy from a different aspect.

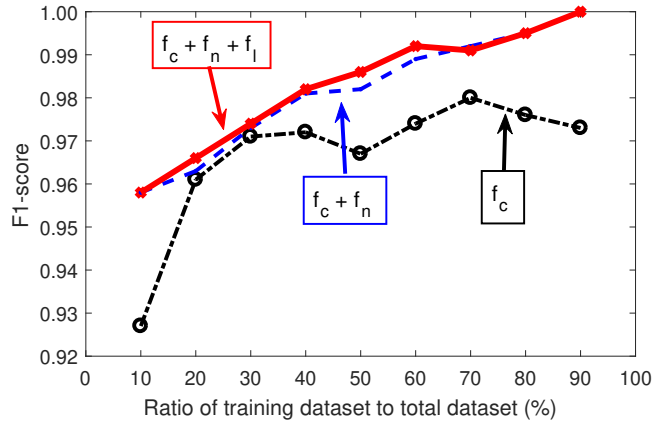

Fig. 7: Comparison of effectiveness of combination of three trust factors

C. Effectiveness of the Trust Factor Weights

This section aims to show how the weights of the three trust factors influence the F1-score. In particular, w_c, w_n and w_l represent the weights assigned to f_c, f_n and f_l, respectively, where the sum of w_c, w_n and w_l equals one. In this experiment, 50% data is used as the training data. The values of w_c, w_n and w_l are in the range of $[0.1, 0.8]$ with the step interval as 0.1. This experiment was run for eight rounds. For a specific round i, the value of w_c is fixed as $i * 0.1$, and all different combinations of w_n and w_l are tested, where $w_n + w_l = 1 - w_c$.

Figure 8 shows the F1-score for different weight combinations. The y-axis represents w_c (i.e., the weight of f_c) and the x-axis represents w_l (i.e., the weight of f_l). The weight of f_n (i.e., w_n) can be derived from $1 - w_c - w_l$. The gray scale bar on the right indicates the value of F1-score, where the darkest gray indicates the highest F1-score.

From Figure 8, observe that when the values of w_c, w_n and w_l are 0.4, 0.3 and 0.3 respectively, the proposed scheme achieves the highest F1-score for the selected training data set. In addition, when $w_c \in [0.2, 0.5]$, $w_n \in [0.1, 0.3]$ and $w_l \in [0.3, 0.7]$, the F1-score is higher than 0.984, which validates the robustness of the proposed scheme. For different training data sizes, the best combination of the three weights

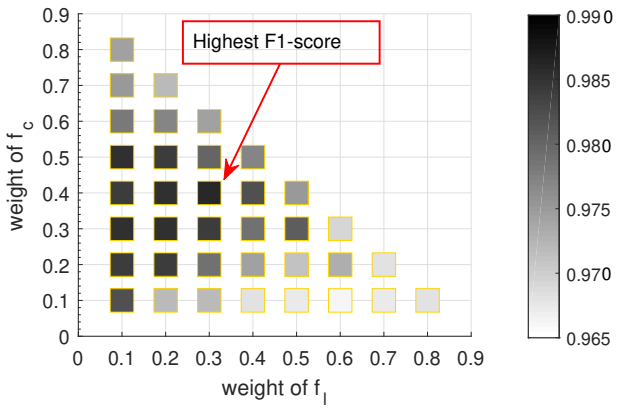

Fig. 8: F1-score performance of three factors combination with different weights

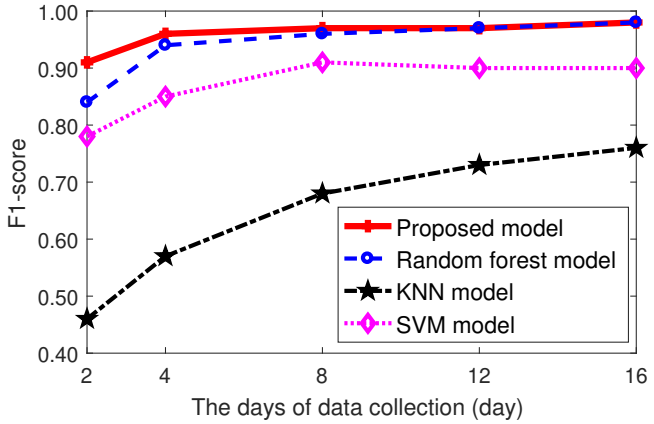

Fig. 9: F1-scores of machine learning models with truncated data and the proposed occupancy detection model

are very similar, indicating that the proposed scheme can achieve stable performances with different weights. Specifically, $w_c = 0.4, w_n = 0.3,$ and $w_l = 0.3$ is selected for the remaining experiments.

D. Overall Performance Comparison

This section compares the proposed occupancy detection scheme with the most popular and reliable models adopted in existing literature.

As directly applying machine learning algorithms on raw sensor data yields very low F1-score (e.g., the random forest model in Figure 4), an event window, which is the same event window from the human activity sequence extraction module of the proposed scheme, is introduced for all the algorithms. All the algorithms use the same training data, with the first few days of data collection used for training. Both truncation technique [50] and padding technique [51] are used for all machine learning models. These techniques are the most popular and effective means to ensure fixed length inputs for machine learning models.

1) Truncated Inputs: In this section, the input data is truncated into identical lengths and then fed into the machine learning models. Specifically, for each event window, only the first three activities detected by each sensor are selected (and any remaining activities are discarded) to produce a 5×3 matrix (i.e., 5 sensors, each with 3 activities). The value of a given element $x_{p,q}$ in the matrix, where $1 \leq p \leq 5,$ and $1 \leq q \leq 3,$ indicates the time when the q^{th} activity on sensor p is detected. These matrices are then fed into different machine learning models for occupancy detection.

The occupancy detection results, the F1-score and its standard deviation, are shown in Figure 9 and 10, respectively. The x-axis of both Figure 9 and Figure 10 represents the training data size, ranging from two days to sixteen days. Different curves in each figure represent occupancy detection models.

From Figure 9, it is observed that the proposed model achieves the highest F1-score, especially when the training data size is small. Specifically, when applying two days data as training data, the F1-score of the proposed model is higher than that of the random forest model, SVM model and KNN model by 8 %, 17 %, and 98 %, respectively. When the training data size is extended to four days, the F1-score of the proposed model is higher than that of the random forest model, SVM model and KNN model by 1 %, 14 %, and 70 %, respectively. With the increase of training data size, the performance of the random forest model is gradually approaching the proposed model.

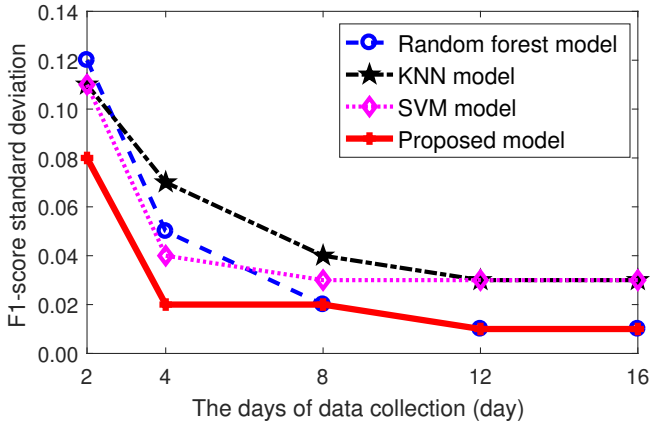

Fig. 10: F1-score standard deviation of machine learning models with truncated data and the proposed occupancy detection model

Figure 10 compares the F1-score standard deviation of the proposed model and machine learning based models. The proposed model yields the lowest F1-score standard deviation, especially for two days and four days training data set, demonstrating its high reliability.

2) Padded Inputs: In this section, the fixed input lengths are ensured through padding techniques. Specifically, the longest event window in terms of time duration is identified first. As the sampling frequency of each experiment is 1 Hz, an event window lasting for 60 seconds will include 60 samples for each of the five sensors. Assume that the longest event window lasts for l seconds, the goal is to construct an input matrix as a $5 \times l$ matrix from each event window. Each element $x_{p,q}$ in the matrix, where $1 \leq p \leq 5,$ and $1 \leq q \leq l,$ is a binary value, indicating whether an activity has been detected on sensor p at the q^{th} second. For event windows shorter than l, zeros are

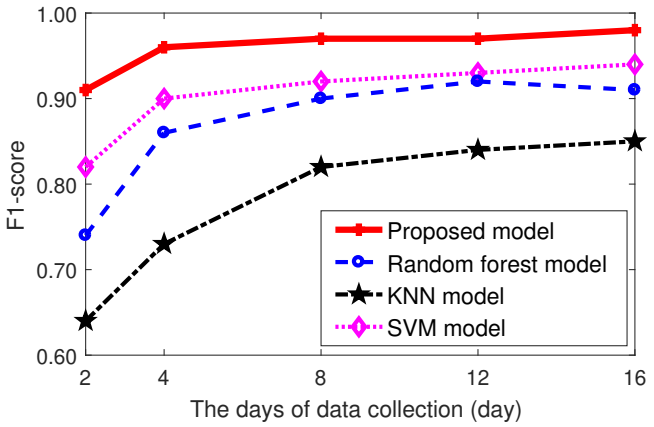

Fig. 11: F1-score of machine learning models with padded data and the proposed occupancy detection model

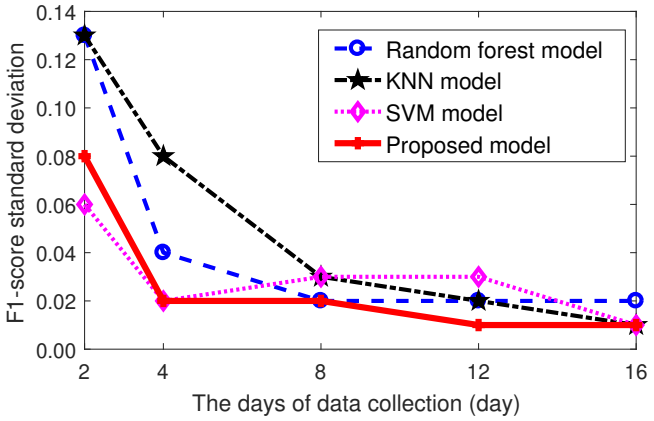

Fig. 12: F1-score standard deviation of machine learning models with padded data and the proposed occupancy detection model

added at the rightmost columns to construct a $5 \times l$ matrix.

Figure 11 shows the comparison of the F1-score of the proposed occupancy detection model and machine learning models with padding. It is easy to observe that the proposed model achieves the highest F1-scores. Particularly, when using two days data as training data, the F1-score of the proposed model is higher than that of the SVM model, random forest model and KNN model by 11 %, 23 % and 42 %, respectively. When the training data size is extended to four days, the proposed model achieves higher F1-score than SVM model, random forest model and KNN model by 7 %, 12 %, and 32 %, respectively. With the increase of training data size, the performance of the SVM model is gradually getting closer to the proposed model.

Figure 12 shows the standard deviation of the F1-score for the proposed occupancy detection model and machine learning models with padding. For two days training data, the SVM model has achieved the smallest standard deviation, only 2 % less than that of the proposed model. Although it shows the reliability of the F1-score for the SVM model, the highest F1 value is achieved by the proposed model. When the training data is extended to more than two days, the proposed model yields the smallest standard deviation, indicating its reliability.

In summary, from Figure 9 - 12, compared to other machine learning models, the proposed occupancy detection model reliably achieves the highest F1-score, especially when the training data size is small.

Table III compares the proposed model with existing models from different aspects. Compared with [13], which deploys an expensive CO_2 sensor to collect data as inputs, the accuracy of the proposed model is significantly higher. Compared with [25]–[28], which apply machine learning models on diverse data collected via a large number of expensive sensors, the proposed model with limited training data achieves higher accuracy. The accuracy of [21] is the same as that of the proposed model, which, however, may raise privacy concerns due to the introduction of a camera. Compared with the LDA model in [29], which achieves higher accuracy, the proposed model has significant advantages in the number of sensors per square meter and the cost of sensors. Table III validates that the proposed model is particularly appropriate to be applied in residential houses with limited number of low-cost and non-intrusive sensors.

VI. Conclusion

In smart residential houses, occupancy detection plays an important role in balancing energy saving and user comfort levels. In particular, a reliable and accurate occupancy detection mechanism is essential.

This paper proposes an occupancy detection system, using few low cost and non-intrusive sensors, extracts the semantic meanings of the sensor data as "human activity sequences", and applies a trustworthy sequence matching scheme to minimize indeterminacies in the decision-making process. Comprehensive experiments have been performed to validate the effectiveness of each module in the proposed scheme as well as its overall performance. Results show that the proposed scheme outperforms popular machine learning based occupancy detection schemes adopted in existing literature, especially in the cases where only limited training data is available.

Future work will further improve the accuracy of the proposed scheme (i.e., applying deep learning method to extract the weight of each trust factor). The input data without ground truth information will also be handled by proposed scheme in future work.

VII. Acknowledgment

Portions of this publication and research efforts are made possible through the support of NIST via federal award #70NANB19H136. Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States. Certain commercial products are identified in order to adequately specify the procedure; this does not imply endorsement or recommendation by NIST, nor does it imply that such products are necessarily the best available for the purpose.

TABLE III: Overall Comparison of Proposed Model and Existing Models

Reference		[13]	[21]	[25]	[26]	[27]	[28]	[29]	proposed model
Detection method		Mass balance equation based algorithm	Image-based method	DT	DT	ANN	LR	LDA	Trust-based scheme
Accuracy		95.80%	99.20%	98.20%	98.44%	99.06%	95.60%	99.33%	99.20%
Number of training data samples		N/A	N/A	19368	35947928	10807	10080	8232	4035
Ratio of training data set to total data set		N/A	N/A	90%	86%	53%	70%	46%	60%
Intrusive or Expensive Sensor	camera		✓						
	CO_2	✓		✓	✓	✓	✓	✓	
Non-intrusive and Economic Sensor	sound			✓	✓		✓		
	humidity			✓		✓		✓	
	light			✓	✓	✓	✓	✓	
	temperature			✓	✓	✓	✓	✓	✓
	motion			✓	✓				✓
Total number of sensors		1	1	8	7	N/A	10	4	5
Total coverage (m^2)		20.0	N/A	18.6	N/A	N/A	46.7	20.5	65

N/A : The data is not provided in the reference.

REFERENCES

[1] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (iot): A vision, architectural elements, and future directions," *Future generation computer systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[2] P. Rashidi and D. J. Cook, "Com: A method for mining and monitoring human activity patterns in home-based health monitoring systems," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 4, no. 4, pp. 1–20, 2013.

[3] A. Milenković, C. Otto, and E. Jovanov, "Wireless sensor networks for personal health monitoring: Issues and an implementation," *Computer communications*, vol. 29, no. 13-14, pp. 2521–2533, 2006.

[4] J. Scott, A. Bernheim Brush, J. Krumm, B. Meyers, M. Hazas, S. Hodges, and N. Villar, "Preheat: controlling home heating using occupancy prediction," in *Proceedings of the 13th international conference on Ubiquitous computing*, 2011, pp. 281–290.

[5] J. A. Oliveira-Lima, R. Morais, J. Martins, A. Florea, and C. Lima, "Load forecast on intelligent buildings based on temporary occupancy monitoring," *Energy and Buildings*, vol. 116, pp. 512–521, 2016.

[6] J. Ahmad, H. Larijani, R. Emmanuel, M. Mannion, and A. Javed, "Occupancy detection in non-residential buildings–a survey and novel privacy preserved occupancy monitoring solution," *Applied Computing and Informatics*, 2020.

[7] L. Rueda, K. Agbossou, A. Cardenas, N. Henao, and S. Kelouwani, "A comprehensive review of approaches to building occupancy detection," *Building and Environment*, p. 106966, 2020.

[8] D. Trivedi and V. Badarla, "Occupancy detection systems for indoor environments: A survey of approaches and methods," *Indoor and Built Environment*, vol. 29, no. 8, pp. 1053–1069, 2020.

[9] T. A. Nguyen and M. Aiello, "Energy intelligent buildings based on user activity: A survey," *Energy and buildings*, vol. 56, pp. 244–257, 2013.

[10] V. L. Erickson, S. Achleitner, and A. E. Cerpa, "Poem: Power-efficient occupancy-based energy management system," in *Proceedings of the 12th international conference on Information processing in sensor networks*, 2013, pp. 203–216.

[11] L. V. Thanayankizil, S. K. Ghai, D. Chakraborty, and D. P. Seetharam, "Softgreen: Towards energy management of green office buildings with soft sensors," in *2012 Fourth International Conference on Communication Systems and Networks (COMSNETS 2012)*. IEEE, 2012, pp. 1–6.

[12] M. Taddicken, "The 'privacy paradox'in the social web: The impact of privacy concerns, individual characteristics, and the perceived social relevance on different forms of self-disclosure," *Journal of Computer-Mediated Communication*, vol. 19, no. 2, pp. 248–273, 2014.

[13] D. Cali, P. Matthes, K. Huchtemann, R. Streblow, and D. Müller, "Co2 based occupancy detection algorithm: Experimental analysis and validation for office and residential buildings," *Building and Environment*, vol. 86, pp. 39–49, 2015.

[14] M. Zuraimi, A. Pantazaras, K. Chaturvedi, J. Yang, K. Tham, and S. Lee, "Predicting occupancy counts using physical and statistical co2-based modeling methodologies," *Building and Environment*, vol. 123, pp. 517–528, 2017.

[15] M. Naqi, S. Lee, H.-J. Kwon, M. G. Lee, M. Kim, T. W. Kim, H. K. Shin, S. Kang, S. Gandla, H.-S. Lee *et al.*, "A fully integrated flexible heterogeneous temperature and humidity sensor-based occupancy detection device for smart office applications," *Advanced Materials Technologies*, vol. 4, no. 12, p. 1900619, 2019.

[16] H. Hajj, W. El-Hajj, M. Dabbagh, and T. R. Arabi, "An algorithm-centric energy-aware design methodology," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 11, pp. 2431–2435, 2013.

[17] S. Mamidi, Y.-H. Chang, and R. Maheswaran, "Improving building energy efficiency with a network of sensing, learning and prediction agents," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, 2012, pp. 45–52.

[18] K. Weekly, N. Bekiaris-Liberis, M. Jin, and A. M. Bayen, "Modeling and estimation of the humans' effect on the co 2 dynamics inside a conference room," *IEEE Transactions on Control Systems Technology*, vol. 23, no. 5, pp. 1770–1781, 2015.

[19] T. H. Pedersen, K. U. Nielsen, and S. Petersen, "Method for room occupancy detection based on trajectory of indoor climate sensor data," *Building and Environment*, vol. 115, pp. 147–156, 2017.

[20] Z. Yang, N. Li, B. Becerik-Gerber, and M. Orosz, "A multi-sensor based occupancy estimation model for supporting demand driven hvac operations," in *Proceedings of the 2012 Symposium on Simulation for Architecture and Urban Design*, 2012, pp. 1–8.

[21] S. Petersen, T. H. Pedersen, K. U. Nielsen, and M. D. Knudsen, "Establishing an image-based ground truth for validation of sensor data-based room occupancy detection," *Energy and Buildings*, vol. 130, pp. 787–793, 2016.

[22] A. Alahi, L. Jacques, Y. Boursier, and P. Vandergheynst, "Sparsity driven people localization with a heterogeneous network of cameras," *Journal of Mathematical Imaging and Vision*, vol. 41, no. 1-2, pp. 39–58, 2011.

[23] G. Ansanay-Alex, "Estimating occupancy using indoor carbon dioxide concentrations only in an office building: a method and qualitative assessment," *11th REHVA World Congr. Energy Eff., Smart Healthy Build*, 2013.

[24] H. Zou, Y. Zhou, J. Yang, and C. J. Spanos, "Device-free occupancy detection and crowd counting in smart buildings with wifi-enabled iot," *Energy and Buildings*, vol. 174, pp. 309–322, 2018.

[25] Z. Yang, N. Li, B. Becerik-Gerber, and M. Orosz, "A systematic approach to occupancy modeling in ambient sensor-rich buildings," *Simulation*, vol. 90, no. 8, pp. 960–977, 2014.

[26] E. Hailemariam, R. Goldstein, R. Attar, and A. Khan, "Real-time occupancy detection using decision trees with multiple sensor types," in *Proceedings of the 2011 Symposium on Simulation for Architecture and Urban Design*, 2011, pp. 141–148.

[27] K. Tutuncu, O. Cataltas, and M. Koklu, "Occupancy detection through light, temperature, humidity and co2 sensors using ann," *International Journal of Industrial Electronics and Electrical Engineering*, vol. 5, 2016.

[28] B. Abade, D. Perez Abreu, and M. Curado, "A non-intrusive approach for indoor occupancy detection in smart environments," *Sensors*, vol. 18, no. 11, p. 3953, 2018.

[29] L. M. Candanedo and V. Feldheim, "Accurate occupancy detection of an office room from light, temperature, humidity and co2 measurements

using statistical learning models," *Energy and Buildings*, vol. 112, pp. 28–39, 2016.
[30] C. Wang, J. Jiang, T. Roth, C. Nguyen, Y. Liu, and H. Lee, "Integrated sensor data processing for occupancy detection in residential buildings," *Energy and Buildings*, p. 110810, 2021.
[31] A. Szczurek, M. Maciejewska, A. Wyłomańska, R. Zimroz, G. Żak, and A. Dolega, "Detection of occupancy profile based on carbon dioxide concentration pattern matching," *Measurement*, vol. 93, pp. 265–271, 2016.
[32] W. Fang, W. Zhang, Y. Yang, Y. Liu, and W. Chen, "A resilient trust management scheme for defending against reputation time-varying attacks based on beta distribution," *Science China Information Sciences*, vol. 60, no. 4, p. 040305, 2017.
[33] A. Josang and R. Ismail, "The beta reputation system," in *Proceedings of the 15th bled electronic commerce conference*, vol. 5, 2002, pp. 2502–2511.
[34] M. S. Christo, S. Meenakshi, and R. Subhashini, "An intelligent fuzzy beta reputation model for securing information in p2p health care applications," 2017.
[35] C. Zhang, L. Zhu, C. Xu, K. Sharif, K. Ding, X. Liu, X. Du, and M. Guizani, "Tppr: A trust-based and privacy-preserving platoon recommendation scheme in vanet," *IEEE Transactions on Services Computing*, 2019.
[36] K. K. Bharadwaj and M. Y. H. Al-Shamri, "Fuzzy computational models for trust and reputation systems," *Electronic commerce research and applications*, vol. 8, no. 1, pp. 37–47, 2009.
[37] K. Sentz, S. Ferson *et al.*, *Combination of evidence in Dempster-Shafer theory*. Sandia National Laboratories Albuquerque, 2002, vol. 4015.
[38] Y. Liu, F. Zhang, Y. Sun, and H. Huang, "Trust sensor interface for improving reliability of emg-based user intent recognition," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 7516–7520.
[39] H. Zhao, D. Sun, H. Yue, M. Zhao, and S. Cheng, "Dynamic trust model for vehicular cyber-physical systems." *IJ Network Security*, vol. 20, no. 1, pp. 157–167, 2018.
[40] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 267–282, 2007.
[41] J. Kim, K. Min, M. Jung, and S. Chi, "Occupant behavior monitoring and emergency event detection in single-person households using deep learning-based sound recognition," *Building and Environment*, vol. 181, p. 107092, 2020.
[42] C. Wang and H. Lee, "Economical and non-invasive residential human presence sensing via temperature measurement," in *ASME International Mechanical Engineering Congress and Exposition*, vol. 52071. American Society of Mechanical Engineers, 2018, p. V06AT08A051.
[43] R. M. Gray, *Entropy and information theory*. Springer Science & Business Media, 2011.
[44] A. K. Gupta and S. Nadarajah, *Handbook of beta distribution and its applications*. CRC press, 2004.
[45] A. P. Singh, V. Jain, S. Chaudhari, F. A. Kraemer, S. Werner, and V. Garg, "Machine learning-based occupancy estimation using multivariate sensor nodes," in *2018 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2018, pp. 1–6.
[46] R. Bitter, T. Mohiuddin, and M. Nawrocki, *LabVIEW: Advanced programming techniques*. CRC press, 2017.
[47] H. A. Imran, U. Mujahid, S. Wazir, U. Latif, and K. Mehmood, "Embedded development boards for edge-ai: A comprehensive report," *arXiv preprint arXiv:2009.00803*, 2020.
[48] T. Vafeiadis, S. Zikos, G. Stavropoulos, D. Ioannidis, S. Krinidis, D. Tzovaras, and K. Moustakas, "Machine learning based occupancy detection via the use of smart meters," in *2017 International Symposium on Computer Science and Intelligent Controls (ISCSIC)*. IEEE, 2017, pp. 6–12.
[49] N. Nnamoko and I. Korkontzelos, "Efficient treatment of outliers and class imbalance for diabetes prediction," *Artificial Intelligence in Medicine*, vol. 104, p. 101815, 2020.
[50] X. Shen, L. Niu, Z. Qi, and Y. Tian, "Support vector machine classifier with truncated pinball loss," *Pattern Recognition*, vol. 68, pp. 199–210, 2017.
[51] M. Dwarampudi and N. Reddy, "Effects of padding on lstms and cnns," *arXiv preprint arXiv:1903.07288*, 2019.