View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Institutional Knowledge at Singapore Management University

Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems

School of Computing and Information Systems

12-2020

Robust, fine-grained occupancy estimation via combined camera & WiFi indoor localization

Anuradha RAVI
Singapore Management University, anuradhar@smu.edu.sg

Archan MISRA
Singapore Management University, archanm@smu.edu.sg

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the Numerical Analysis and Scientific Computing Commons, and the Software Engineering Commons

Citation

RAVI, Anuradha and MISRA, Archan. Robust, fine-grained occupancy estimation via combined camera & WiFi indoor localization. (2020). *2020 IEEE 17th International Conference on Mobile Ad-Hoc and Smart Systems (MASS): Delhi, India, 10-13 December: Proceedings*. 558-566. Research Collection School Of Computing and Information Systems.
Available at: https://ink.library.smu.edu.sg/sis_research/5668

This Conference Proceeding Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

Robust, Fine-Grained Occupancy Estimation via Combined Camera & WiFi Indoor Localization

Anuradha Ravi, Archan Misra
School of Information Systems, Singapore Management University, Singapore
anuradhar@smu.edu.sg; archanm@smu.edu.sg

***Abstract*—We describe the development of a robust, accurate and practically-validated technique for estimating the occupancy count in indoor spaces, based on a combination of WiFi & video sensing. While fusing these two sensing-based inputs is conceptually straightforward, the paper demonstrates and tackles the complexity that arises from several practical artefacts, such as (i) over-counting when a single individual uses multiple WiFi devices and under-counting when the individual has no such device; (ii) corresponding errors in image analysis due to real-world artefacts, such as occlusion, and (iii) the variable errors in mapping image bounding boxes (which can include multiple possible types of human views: {head, torso, full-body}) to location coordinates. We develop statistical techniques to overcome these practical challenges, and finally propose a novel fusion algorithm, based on inexact bipartite matching of these two streams of independent estimates, to estimate the occupancy in complex, multi-inhabitant indoor spaces (such as university labs). We experimentally demonstrate that this estimation technique is robust and accurate, achieving less than 20% error, in an approx. $85m^2$ lab space (with the error staying below 30% in a smaller $25m^2$ area), across a wide variety of occupancy conditions.**

***Index Terms*—Indoor Localization, Occupancy Estimation, Camera Occupancy, RADAR Occupancy**

I. Introduction

Unobtrusive, cost-effective and fine-grained sensing of human occupancy in indoor spaces is a key enabler for many smart computing applications, such as occupancy-aware energy management [1] and dynamic recommendation of meeting spaces [2]. As an alternative to custom or purpose-built solutions (e.g., [3], [4]), there has been strong interest in the use of both WiFi-based [5] and vision-based [6] occupancy technologies, which piggyback on the existing commonly-deployed infrastructure (WiFi Access Points (APs) and security cameras).

Both WiFi and camera-based occupancy estimation techniques have their advantages, as well as unique limitations. For universal coverage, WiFi-based solutions cannot assume the active participation of (use of custom software on) all mobile devices, but must employ only *passive sensing-based server-side* localization techniques, which have been demonstrated to have median errors of ∼6-8 meters [7]. Practical WiFi-based techniques thus cannot support fine-grained localization (e.g., cannot accurately count people inside a 5x5m^2 meeting space), and fundamentally compute *device occupancy*, which can lead to both errors of *under-counting* (when individual occupants do not possess a WiFi-enabled device) and *over-counting* (when an individual carries multiple WiFi-enabled devices). On the other hand, deep neural network (DNN)-based vision sensing techniques (e.g., YoLoV3 [8], SSD [9]) have recently achieved impressive accuracy, with high object detection accuracy and localization errors of usually 2-3 meters under careful calibration. However, vision-based people counting techniques suffer especially from false negatives, due to artefacts such as full or partial occlusion, partial visual coverage and poor ambient lighting.

Our research is motivated by our ongoing work to support energy-efficient, smart usage of an operational, 5-story commercial Zero-Energy Building (ZEB). While previous work has focused primarily on optimizing HVAC settings (e.g., [10] achieved $\sim 17.8\%$ of energy savings without compromising occupant comfort), we shall also perform occupancy-aware optimization of LED lighting (using ideas discussed in [11], [12]). Such fine-grained adjustment, especially of lights that are typically separated by distances of 1-2 meters, requires accurate estimation of occupancy at similarly fine-grained spatial resolutions of 20-25m^2. To achieve this objective, we tackle the challenge: *How do we fuse the sensing modalities of both WiFi and cameras to achieve accurate, finer-resolution occupancy estimation in indoor shared spaces?* While the idea of fusing WiFi and camera sensing is not completely new, prior work (such as [13], [14]) has addressed this under fairly limited or specially-instrumented scenarios, without considering many impairments characteristic of real-world public spaces. In particular, our experience with real-world, shared spaces indicate the need to tackle the *different Localization Errors & Variance* encountered due to the problems of over/under-counting (for WiFi) and false negatives (for camera) mentioned previously.

Key Contributions: In this paper, we present the design of a practical WiFi+ camera-based occupancy estimation system, and show that it can overcome the challenges mentioned above to provide accurate, high-resolution estimates of human occupancy. The paper makes the following key contributions:

- *Improved Accuracy of Practical Camera-based Human Localization:* The ability to translate the image-specific bounding box coordinates of human objects to the physical-world space coordinates is an essential prerequisite for fine-grained, camera-based occupancy estimation. Prior approaches for such coordinate translation assume the use of homographic mappings, which require complex calibration; moreover, they implicitly assume a consistent view (e.g., full body) of all human subjects. We propose a simpler,

regression-based mapping technique, which combines separate DNN pipelines for *human* and *face* detection, to achieve an average localization error of 1.7 meters, irrespective of whether the human is close to or far from the camera and whether the human object is fully or partially visible.

- *Introduce a Robust Algorithm for Human-Device Matching:* To support high fidelity WiFi-and-camera based occupancy estimation, *while accommodating individuals with varying number of personal devices*, we propose a novel algorithm for matching the human objects (and their corresponding bounding boxes) identified in image frames to one or more WiFi-enabled devices. The algorithm treats the set of identified devices and human objects as elements of two bipartite graphs, and then employs *approximate b-matching* techniques to accurately match humans to devices, while accommodating the possibility of individuals having either no or multiple WiFi-enabled devices.
- *Demonstrate Practical Effectiveness at Multiple Spatial Granularity:* We implement and evaluate a real-world prototype, based on b-matching based fusion of RADAR-based passive WiFi localization and YoLoV3-based human object detection. Our experimental studies show that our system can estimate total human occupancy, with an estimation error of less than 20% over an $85m^2$ shared, lab space. The estimation error increases slightly to $\sim$ 30% when performed over a smaller $25m^2$ area. This represents a significant reduction, of $\sim 10-15\%$ over a competitive pure camera-based approach, and over 40% over a pure WiFi-based alternative.

II. Related Work

Occupancy estimates can be derived by a range of techniques, including the use of custom sensors, performing sensor fusion on muti-sensor data streams ([15], [16], [17] [18]), incorporating energy utilisation patterns along with sensor data [19], studying occupant behavioural patterns [20], or utilising WiFi to count devices and camera modules to count people. We discuss some of the key prior approaches that leverage on WiFi and camera sensing to estimate the people count.

A. WiFi and Camera Based Estimation

Camera-based head detection approaches (e.g., [21]) use overhead cameras to count and track people by employing image-processing based techniques for accurately detecting the head, and achieve people counting accuracy of 96%. POEM, a work by Erickson et al., [22] used a particle filtering approach to combine the occupancy estimates from a network of cameras with the occupancy estimates obtained from a network of motion detectors, using such occupancy estimates to infer the desired temperature set point and achieve $\sim$30% reduction in the energy consumption. Soltanaghaei et al., [23] characterize occupancy states as $\{$moving, stationary, unoccupied$\}$ by leveraging WiFi PHY layer Channel state information, with an accuracy of 96.7%. Wang et al., [24] deploy cameras at the entrance of workspaces to count the number of heads entering and leaving, using additional sensing of per-hour CO_2 levels, to alleviate false positive and false negative errors.

B. WiFi-Camera Fusion

The limited prior work on fusing camera-based human localization with WiFi-based sensing typically lacks the robustness to deal with various real-world artefacts and challenges. In particular, [13] presented an approach for opportunistic localization, where (a) vision-based object detection used background subtraction methods (unlike our focus on more modern DNN-techniques) and thus worked only for non-stationary humans, and (b) the fusion estimate, based on a linear weighted sum of WiFi & camera location estimates, implicitly assume the presence of a single user with a single device. Similarly, Domingo et al., [14] integrate WiFi and camera-based location estimates for a finer-grained tracking of trajectories of non-stationary humans, utilizing a highly-instrumented environment with a significantly higher density of WiFi APs and camera deployment (36 cameras over an $80m^2$ lab space, compared to our use of 2 cameras over an $85m^2$ area). This approach uses multiple camera views to perform camera localization, and performs client-side WiFi localization (using a custom Android App on each user's mobile device that provides RSSI measurements every 2 secs). It is thus not directly applicable to our goal of performing such localization universally, without explicit user participation.

Unlike these prior works, we consider a multi-person, partially occluded environment (with both stationary & moving subjects), tackle the *device-person association* challenge when individuals are free to carry zero or multiple WiFi-enabled devices and resolve the false-negatives of camera-based occupancy estimation by including additional *unmatched* devices.

III. System Overview

In our work, we set up a real-world testbed in an academic research workspace (layout illustrated in Figure 1, including camera locations denoted as C_i and C_F_i and representative WiFi landmarks denoted by L_i), instrumenting the area with **2 cameras** spaced 8m apart, sending video feeds to our server. We utilize the commercially deployed Aruba [25] WiFi AP infrastructure to perform server-side indoor localization. To establish an understanding of baseline performance, we first individually studied the accuracy of occupancy estimation based on either WiFi-based server-side localization (using the RADAR [26] algorithm) or camera-based human object detection (using the YoLoV3 object detector [27]).

To illustrate the key challenge, Figure 2 plots the performance of camera and WiFi system compared to the manually-recorded ground truth in our $\sim$ $85m^2$ testbed space over a 3 hour duration, estimated once every 2.5 minutes. We observed that WiFi-based estimation produces significantly higher counts (**140.1%** estimation error), most likely because human tend to carry multiple WiFi-enabled personal devices. Moreover, due to effects such as shadowing and multipath, WiFi devices located outside the region of interest occasionally

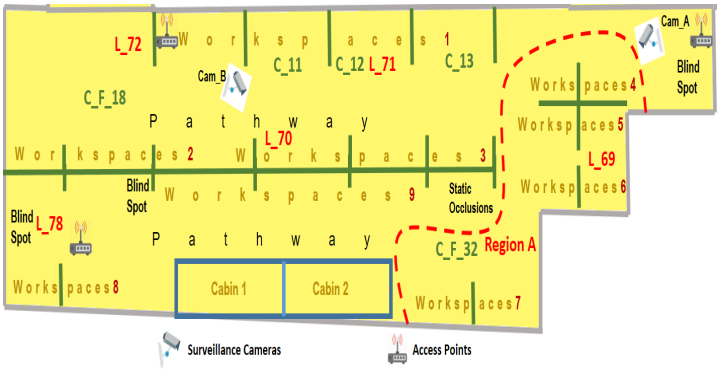


Fig. 1: Floor Plan for Observation Area (with Camera & WiFi Deployment)

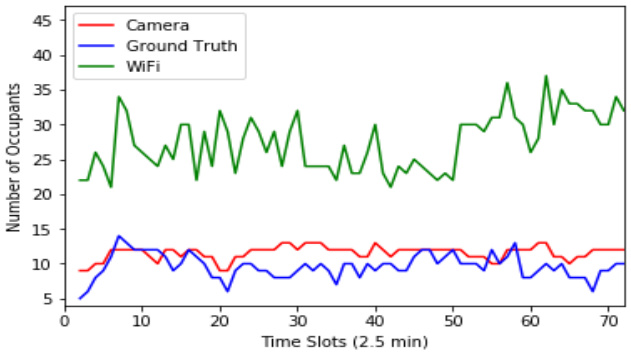


Fig. 2: WiFi & Camera vs. Ground Truth Occupancy

get mapped to fingerprinted landmarks inside, leading to a spurious increase in the device count. Camera-based analysis, on the other hand, provides better occupancy estimates (**18.9%** estimation error), but suffers from multiple false negatives and false positives (for example, human-like toy action figures detected as humans). Our target is to reduce the occupancy errors to approx. 15-20%, estimated over areas of roughly $25-40\ m^2$ (as desired by our candidate smart applications). In this section, we outline the basic steps of WiFi-based device localization and camera-based human object detection (both reusing state-of-the-art techniques), deferring our novel enhancements to the (a) camera-based estimation process and (b) the subsequent fusion pipeline to Sections IV-V.

A. WiFi based Location Estimation

We estimate the WiFi-based occupancy by first computing the location of each device based on the fingerprinting-based RADAR algorithm [26], as applied to server-side WiFi measurements collected passively by WiFi APs. The WiFi fingerprints are computed at designated landmarks, 4-5 meters apart. In the default version of RADAR, the mean RSSI measurements, obtained by multiple APs, are compared with the fingerprint readings at different landmarks, and the device is mapped to the landmark with the smallest Euclidean distance (in the RSSI space). Because of a variety of well-known artefacts (e.g., shadow fading, impact of humans, different antenna gains of mobile devices), such an approach exhibits a median localization error of $\sim 6-8$ meters [7]. To improve the localization accuracy, even if modestly, we utilize a modified "weighted centroid" version of the base RADAR algorithm, where (a) the localization procedure first finds the top-k ($k = 3$ in our implementation) closest landmark locations for each location estimation interval ($T_l = 5$ secs), and (b) then computes the device's location (X^r, Y^r) as the weighted mean of these top-3 locations, as follows:

$$X^r = \frac{(\frac{x_1}{RSSI_{Error_1}}) + ... + (\frac{x_k}{RSSI_{Error_k}})}{RSSI_{Error_1} + ... + RSSI_{Error_k}} \quad (1)$$

$$Y^r = \frac{(\frac{y_1}{RSSI_{Error_1}}) + ... + (\frac{y_k}{RSSI_{Error_k}})}{RSSI_{Error_1} + ... + RSSI_{Error_k}} \quad (2)$$

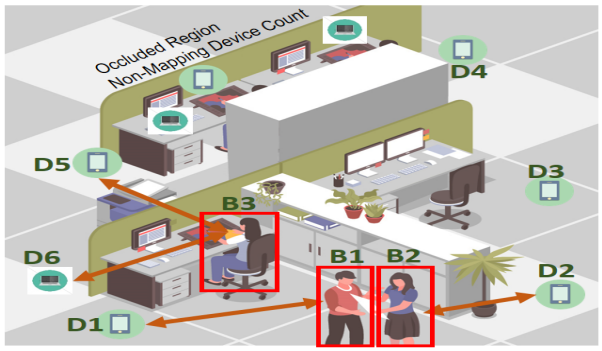


Fig. 3: WiFi - Camera Fusion: Bipartite Graph Matching

B. Camera-based Estimation and WiFi-Camera Fusion

We now provide an overview of the basic components (as part of the baseline approach) of the overall estimation process. **Camera based Occupancy Estimation:** To compute occupancy based on visual sensing, the individual frames obtained from the monitoring cameras are first processed using a state of the art object detection algorithm, such as YoLoV3 [27], to extract the bounding box coordinates of the detected humans. These boxes are subsequently converted from their on-screen (pixel) coordinates to the real-world (physical) coordinates via a process of *coordinate mapping*. Direct application of such DNN-based object detectors were seen to result in both high false positives and false negatives due to factors such as incomplete coverage (some portions of the physical space are not surveilled, sometimes due to privacy concerns) and multiple blind spots (occluded in the camera's field of view (FoV)). In particular, investigations with an off-the-shelf detector and images from our ceiling-mounted CCTV cameras were seen to result in 2-5 false negatives (FN) and ~ 2 false positives (FP) per frame.

WiFi-Camera Multimodal Fusion : The outputs from the WiFi and camera-based localization systems are subsequently fused together, based on the presumptive ability of each sensing mode to compensate for each other's errors, to develop an improved estimation of the number of distinct human objects within the region of interest. Central to such fusion and improved occupancy estimation is the idea of *human*$\leftarrow\rightarrow$*device*

matching, where one or more devices localized by the WiFi localization process is matched to one of the human beings captured by "bounding boxes" embedded within an image frame. This matching process is driven by the intuitive assumption that a user will be physically proximate to his or her personal devices. Figure 3 visually illustrates this core concept, which can be modeled as a form of inexact bipartite graph matching (*b-matching*) between a set of "device" ("D") nodes and a set of "human" (bounding box or "B") nodes. The inexact matching arises from the possibility that a single human may have (and thus be associated with) multiple devices or may not possess any WiFi-enabled device.

The *b-matching* step is further complicated by the presence of visually occluded regions. Users in such an occluded region would not be visible in the camera FoV, but may be represented by one or more localized WiFi devices. To incorporate such "invisible users" in the overall occupancy estimate, we increase the occupancy count by counting unmatched devices located in such occluded regions. The final total occupancy thus includes the (a) occupancy count of human objects (camera bounding boxes), matched to one or more devices via a b-matching algorithm; (b) unmatched devices located within visually occluded regions and (c) any unmatched humans not mapped to the devices due the distance constraints.

IV. Improved Camera-based Human Localization

In this section, we describe an enhancement to camera-based localization subsystem that makes it robust to the real-world artefact of partial & occluded views of humans. The enhanced approach improves the estimation of the physical world coordinates of a human, from the pixel-level bounding box coordinates extracted from individual image frames by a standard, state-of-the-art object detection DNN, YoLoV3 [27].

The classic technique for translating pixel coordinates to physical world coordinates involves the use of a homographic mapping function [28]. However, building such a homographic matrix requires precise knowledge of the camera's pose and the 3-D layout of the space, and is also subject to the intrinsic indeterminacy of projecting a 3-D space to 2-D coordinates: from pure geometry, a point in a 2-D space corresponds to a line in the 3-D space within the camera's *FoV* (field-of-view).

To overcome these limitations, we first trained a linear regressor (with annotated ground-truth data) that takes these features as covariates and outputs the corresponding x and y coordinates in the physical space ($z = 0$, implying that we attempt to map the human's coordinate on the floor of the lab). This simple regression model, however, turned out to have high error variance (as high as 17 meters). This problem persisted when we shifted to a multiclass logistic regressor, where a detected object was mapped to one of several location *classes* (each corresponding to a location grid). Each human object, identified by the object detector (YoLoV3) is associated with the following bounding box values: bounding box centers (X_{center}, Y_{center}), bounding box height and width (Bd_{ht} & Bd_{wd})). Closer inspection of the actual captured data revealed a major reason for this error: as illustrated in Figure 4, due to occlusion and partial views, the DNN human detector can provide a bounding box for three distinct *types* or classes of human views: (a) full-body (FB), (b) torso-only (TO) and (c) head only (HO), each of which may be present in one or more images. Each of these classes of views are associated with different distances of human objects from the camera. Accordingly, at the same physical location, the size of the bounding boxes, can vary depending on the class of view above.

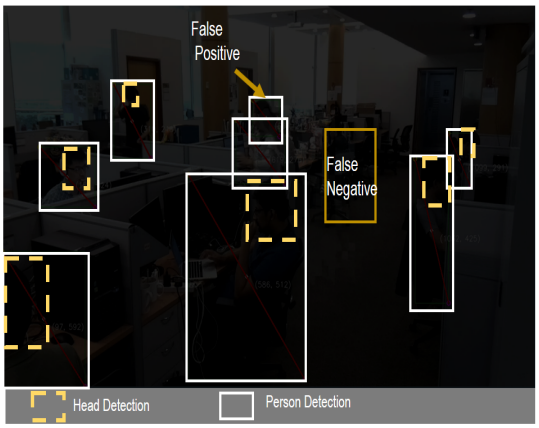


Fig. 4: Frame with Differently-Sized Bounding Boxes and Distinct Human Views

To tackle this variation, we incorporated a second DNN model to perform head/face detection–more specifically, given a set of pixels belonging to a bounding box, we apply a retrained YoLoV3-Head_Detection DNN model [29] on the extracted set of pixels to compute the bounding box coordinates of the 'head/face' object, which provides two additional covariates, Hd_{ht} and Hd_{wd}, corresponding to the height and width of the detected head/face. (Note that in cases where the YoLoV3 detector just isolated a human's torso, these variables would be null.) Our modified *human+ head* regressor now takes output of two cascaded object detectors (one identifying the *human* object with the other identifying an embedded *head* object) to derive the following 6 covariates: ($X_{Center}, Y_{Center}, BB_{ht}, BB_{wd}, Hd_{ht}, Hd_{wd}$), which are then used in the logistic regression model:

$$X^C, Y^C = \text{Logistic}(X_C, Y_C, BB_{ht}, BB_{Wd}, Hd_{ht}, Hd_{wd}) \tag{3}$$

The Hd_{ht} and Hd_{wd} parameters effectively help to distinguish between a location that is closer to a camera and one that's further away.

Performance Evaluation: Table I provides examples of the usefulness of these distinct covariates in the classification. As observed from the data, human objects located at landmark locations 117 and 122 have similar center coordinates, height and width for their YoLoV3-generated bounding boxes, but the significant difference in the size of the head object helps to disambiguate between these two candidate landmark locations (which are physically 8.5 meters apart, with landmark 122 being closer to the camera and consequently capturing a

larger-sized head). Similarly, landmark locations 33 and 41 are physically 6 meters apart: the width of the bounding box and the height and width of the embedded head objects help to discriminate between individuals located at these two landmarks.

TABLE I: Camera-based Localization: Classifier Covariates at Different Landmarks

X_Cen	Y_Cen	BB_ht	BB_Wd	Hd_ht	Hd_Wd	Loc.
95	43	190	583	21	25	**117**
95	44	187	581	135	148	**122**
16	33	115	95	74	73	**33**
16	28	103	170	31	28	**41**

Empirical evaluation showed that the initial 4-covariate classifier (mapping a bounding box to a landmark) achieved an accuracy of 90.2% with the images captured, with the accuracy increasing to 97.9% when the enhanced 6-covariate classifier (which included the head width and height) model was used. Figure 5 then plots the error in such human object localization (for both the baseline 4-covariate regressor and the refined 6-covariate model), as a function of the ground-truth distance of the human from the camera. We see that the enhanced regressor is able to localize humans across a range of camera-human distances, irrespective of whether the camera obtains a full or partial view, with an error $\leq 1.5-2$ meters, which is well within the spatial resolution desired for our final occupancy estimator.

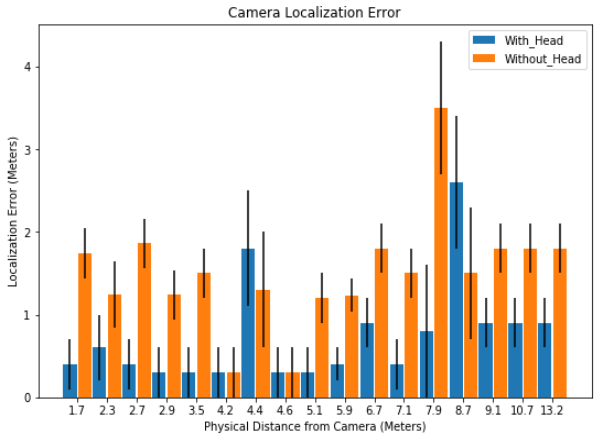


Fig. 5: Camera Localization Error

V. WiFi - Camera Fusion Algorithm

We now describe our fusion algorithm that computes overall occupancy from the (a) estimated set of device locations (obtained by WiFi) and (b) human locations (determined from camera data). The fusion algorithm first models the problem of establishing an association between the devices and the set of humans (bounding boxes) as one of inexact bipartite graph matching, between one set of nodes containing the WiFi-based device location coordinates and the camera-based human object coordinates. The weight of an edge between a pair of nodes in this graph represents the Euclidean distance between the corresponding physical coordinates. We specify a maximum permissible distance between any viable node pair, thereby defining the "admissible edges" in the graph. Our *inexact b-matching optimisation* algorithm then selects the set of edges that provide the best collective matching across all such viable node pairs, while accommodating the reality that one individual can be associated with multiple personal devices. Formally, we constrain a valid matching by a parameter "b" (set to $b = 2$ in our implementation), indicating one human object *may* be matched to a maximum of b devices (in our case, typically, a smartphone and a personal computer).

As $b-matching$ is known to be NP-Hard, we utilize a heuristic to compute the maximum number of such (device, bounding box) associations subject to the above constraints. As a pre-processing step, we first compute the bipartite graph, containing only the "admissible edges". Admissibility is defined based on a distance constraint: any feasible node pair should have a distance that lies in an interval $\{D_{min}$,$D_{max}\}$. These values are derived empirically. Figure 6 gives the CDF plot for the localization error, obtained by placing test devices at different landmarks, defined as the Euclidean distance between the ground-truth location and the RADAR-based estimate. Based on the plot, we set an admissible range $\{D_{min}$ =**3.5m**, D_{max} = **6m**$\}$, corresponding roughly to the 10^{th} and 70^{th} percentile of this estimation error. The fusion algorithm then picks the "optimal" set of associated (device, bounding box) pairs. Finally, we obtain the occupancy count based on a combination of such associated (device, bounding box) pairs as well as the remaining unassociated nodes.

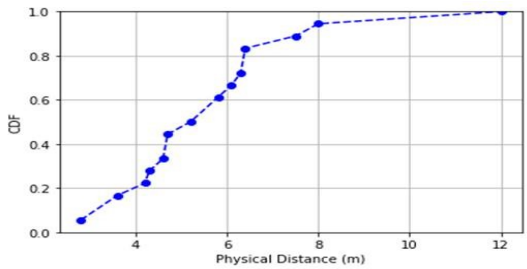


Fig. 6: CDF of Observed WiFi localization error (meters)

A. b-matching based Optimization

The primary goal of our b-matching algorithm is to find a *maximal matching*–i.e., one that finds the largest number of feasible (device, human) associations. The input to such an algorithm consists of a set of world coordinates of the bounding boxes (X^C, Y^C) of only *human objects*[1], computed as described in Section IV, estimated on a "per-frame" basis and a set of world coordinates of the devices (X^R, Y^R) as described in Section III, estimated from the real-time WiFi feeds from Aruba infrastructure at every "5sec" interval.

[1]While DNN object detectors can detect other object classes, we consider only identified "human objects".

Given the differences in sampling frequencies, we perform b-matching only on the frames (once 5 secs), where the camera and WiFi location timestamps coincide.

Mathematically, the maximal matching objective can be described as follows. Let E_{ij} denote an edge between the i^{th} bounding box and the j^{th} device, with $\mathcal{B}$ & $\mathcal{D}$ denoting the total number of bounding boxes and devices, respectively. Let $I(i,j)$ be an indicator function, such that $I(i,j) = 1$ if bounding box i is associated with device j. Moreover, a bounding box is said to be "covered" ($C_B(i) = 1$) if it is associated with at least one device (i.e., $\sum_{j=1}^{\mathcal{D}} I(i,j) \geq 1$), while a device is said to be "covered" ($C_D(j) = 1$) if it is associated with 1 bounding box (i.e., $\sum_{i=1}^{\mathcal{B}} I(i,j) = 1$). The b-matching objective is then as follows:

$$b_{match}(Opt) = \max_{\{I(i,j)\}} \{\sum_{i=1}^{\mathcal{B}} C_B(i) + \sum_{j=1}^{\mathcal{D}} C_D(j)\}$$

subject to the following constraints

$$\sum_{j=1}^{\mathcal{D}} I(i,j) \leq b; \text{ // bounding box matched to at most 2 devices}$$

$$\sum_{i=1}^{\mathcal{B}} I(i,j) <= 1; \text{ // each device matches to at most 1 human;}$$

$$I(i,j) = 0 \text{ iff; } D_{min} < E_{ij} < D_{max} \text{ // edge feasibility check} \quad (4)$$

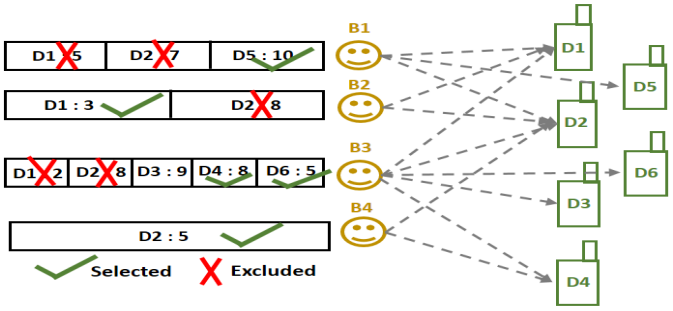


Fig. 7: Illustrating the Iterative *b-matching* Algorithm

We developed an improved b-matching heuristic that operates iteratively through a three-step matching process, with each iteration being applied on a reduced subset of nodes that excludes nodes that have already met their association upper bound (2 for bounding boxes, 1 for devices). Algorithm 1 details the pseudocode of the b-matching process. Each iteration starts off with a *device-centric* matching step, whereby we match devices that have only one feasible edge–i.e., can be mapped to only a single candidate human. Subsequently, we execute a *human-centric* matching step, where we match any available bounding boxes that have only a single feasible edge (i.e., can be mapped to only a single available device). In case of multiple such bounding boxes, the device is matched to the closest bounding box–i.e., the one with the least edge weight. If neither of the above steps are feasible, we execute a *greedy assignment* step, whereby we pick the edge (among the remaining unassigned (box, device) edges) with the lowest weight and match the corresponding bounding box and device.

Algorithm 1 WiFi - Camera Fusion Logic

{Input - (X^c, Y^c) : Per-Frame Camera World Coordinates, (X^r, Y^r) : Radar localization coordinates for each device at 5sec interval.}
if $i < j$ **then**
  {Match bounding box with device }
  b=2 {Each bounding box is mapped to multiple devices. Thus, b=2.}
else
  Match device with bounding box.
  b=1 {Each Device is mapped only to one bounding box. Thus, b=1.}
end if
G : Call Construct-Graph((X^C, Y^C),(X^R, Y^R))
O_{CR} : Call BFS(G) {O_{CR} is the fusion occupancy count }
O_F : Occupancy Final = O_{CR} + O_C + O_R {}O_C represents the number of unmatched bounding boxes, O_R is the count of devices identified across occluded region
procedure CONSTRUCT_GRAPH((X^C, Y^C),(X^R, Y^R))
G.Nodes($(X^C, Y^C)_{1...i}$,$(X^R, Y^R)_{1...j}$) {Add nodes to graph}
for Each $(X^C, Y^C)_1$ to $(X^C, Y^C)_i$ and $(X^R, Y^R)_1$ to $(X^R, Y^R)_j$ **do**
  E_{ij} = $((X^C - X^R) - Y^C - Y^R)^2$ {Here, E represents the distance between camera and RADAR coordinates.}
  if $D_{min} < E_{ij} < D_{max}$ **then**
    G.AddEdge(E_{ij}) {Add edge to graph}
  end if
end for
end procedure
procedure BFS(G)
$B_D = G_{B_D}$ {Create set of devices that are mapped to each bounding box}
for Each Edge in G_{Edges} **do**
  B_D = $Minimum(COUNT(B_D))$
  if $COUNT(B_D)$ = 1 **then**
    Selected-Edges.add($G.Edge(contains(B_D))$)
  else
    B_D = $Minimum(Distance(B_D)$
    Selected-Edges.add($G.Edge(contains(B_D))$)
  end if
  G.remove(Selected-Edges)
end for
end procedure=0

This sequence of steps is then repeated iteratively, until no further matching is possible.

Figure 7 provides an illustrative example of the *b-matching* algorithm. The figure illustrates the feasible set of edges, and their corresponding weights, between 4 bounding boxes (B1-B4) and 6 devices (D1-D6). In the first iteration, B3 is mapped with D6 & D4 and B1 is matched to D5, as these are the only possible matching for these devices. Subsequently, B4 is

matched with D2, as this is the only feasible matching for this bounding box. Based on these matchings, devices D2, D5 and D6 are excluded from further consideration. In the next round of iteration, B2 is matched with D1 as this is the only feasible remaining assignment for B2.

B. Computing the Final Occupancy Count

At the end of the matching process, we may still be left with a set of unmatched bounding boxes and a set of unmatched devices. The final occupancy count is then computed as:

1) We first compute, the sum of all the matched boxes (O_{CR}) and the unmatched boxes (O_C), representing the total number of observed human objects.
2) Subsequently, for the unmatched devices, we form and count O_R, which is likely associated with a human that was either located in an occluded region or was not detected by the DNN-based object detector.

The overall estimated occupancy for the given region is then computed as $O_{CR} + O_C + O_R$.

C. Aggregation

After computing occupancy at every time epoch (5secs), we *smoothen* the estimate over a longer time interval (5 mins), as our applications do not require more frequent estimates. Specifically, to eliminate outliers, we compute the median of the occupancy count over the $\frac{5*60}{5}$=60 individual estimates.

VI. Experimental Results

In this section, we empirically evaluate the efficacy of our proposed system, relative to alternative WiFi or camera-based approaches. We first present results over the approx. $85m^2$ research lab area, described in Figure 1 and subsequently study the accuracy of occupancy detection over a smaller $25m^2$ area (which helps us assess the feasibility of using the system for finer-grained occupancy estimation). To test the system under varying levels of occupancy (driven by the natural arrival & departure patterns of lab personnel or visitors), we conduct the experimental study over 4 different days, capturing the occupancy data and ground truth for approx. 1.5-2 hours/day.

A. Performance of Occupancy Estimation

We evaluate the performance of 3 distinct algorithms: (a) the unimodal WiFi-based and camera-based strategies, (b) our proposed b-matching based approach, with occupancy estimates once every 5 secs, and (c) our smoothed b-matching based estimate, *aggregated* once every 5 mins.

TABLE II: Average Occupancy Estimation Error

Component	Error(In Percentage)
WiFi	**62.9**
Camera	**25.3**
BMatch	**20.7**
BMatch (Aggregated)	**16.2**

We present detailed results on a variety of occupancy conditions, comparing the proposed techniques against the manually-annotated ground truth. Table II lists the average percentage estimation error (averaged over 3 different days under different occupancy levels) for the different techniques.

Based on the estimation error (and results obtained on different days), we derive the following observations:

- WiFi-based techniques perform very poorly, resulting in a significant over-estimate of the occupancy values. The average error accounted for WiFi-based technique is reported to be 65.5%, 54.7% and 69.9% on 3 different days, which is significantly higher than the joint Camera-Wifi b-match approach whose average error is reported to be 27.0%, 13.3% and 22% respectively.
- In contrast, Vision-based people counting is significantly more accurate in estimating the number of occupants across different crowd settings. Across the $\sim$ 6 hours of data collected over 3 days, the baseline Camera method offers an average estimation error of 36%, 21% and and 18.9%. However, the occupancy count error by itself does not provide a complete picture, as the false negatives (when human objects are missed by the object/people detector) and false positives can often cancel each other in terms of the overall occupancy estimate.
- Compared to the other baselines, our proposed approach, which fuses WiFi+ camera estimates and also includes additional unmatched devices in occluded regions, offers a lower average error of **20.7%**, and *performs consistently across a range of occupancy levels*. When smoothed by aggregating such estimates, the average error drops to **16.2%**, which represents a 9% and 48% reduction in error compared to the camera and WiFi-based estimators, respectively. Note that the current matching uses just a single camera frame (time-synced to the WiFi estimate), which is especially susceptible to false negatives. Aggregating human objects detected across multiple frames (in the 5sec interval) should further reduce the vision-based estimation errors, due to false positives and negatives of the object detectors–we plan to study this in future work.

Occupancy Count vs. Occupancy Level: We categorize the occupancy levels over the entire research lab area into 3 broad classes: *Low* (Ground-Truth Occupancy Count: 3-7), *Medium* (Ground-Truth Occupancy Count: 8-13) and *High* (Ground-Truth Occupancy Count: 14-17). Figure 8 plots the average occupancy count for all 4 approaches, for three different occupancy levels. Overall, we see that b-match is able to track the occupancy levels fairly well, and with lower std. deviation.

B. Finer-Grained Occupancy Estimates

One of our eventual goals is to obtain accurate estimates of occupancy at finer spatial granularity–i.e., over smaller-sized regions (as compared to the size of a $85m^2$ research lab)–so as to enable applications such as occupancy-driven smart lighting. To quantify the potential benefit of our approach, we compute and quantify the occupancy error for a smaller $25m^2$-sized region (Region A in Figure 1). Figure 9 plots the estimated

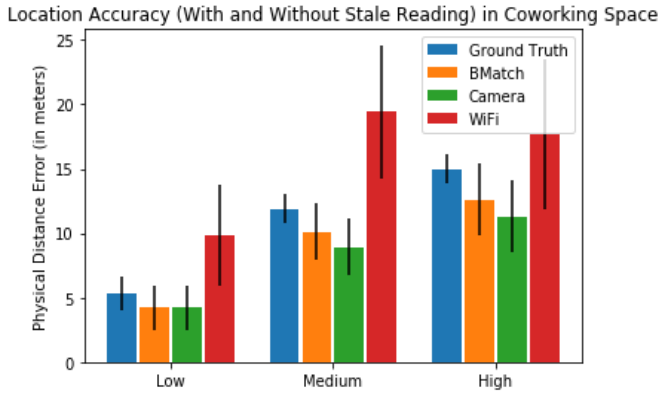


Fig. 8: Occupancy Estimation for Different Crowd Levels

and ground truth occupancy variation vs. time, under High occupancy. We make the following empirical observations:

- The camera-based method (Camera) achieves poor accuracy, underestimating the true occupancy primarily because the object detector DNN is unable to detect humans in the more distant areas under observation. In particular, the DNN detected 2 individuals present in farther views only 60% of time, in contrast to a detection rate of over 99% for individuals located closer to the camera.
- As before, the WiFi based occupancy estimator had the highest error. The overestimate in this case was not just due to the phenomena of individuals carrying multiple devices, but also due to the inherent 6-8 meter error of WiFi localization (as shown in Figure 6), which often leads to the inclusion of extraneous devices.
- The use of b-match (with Aggregation) resulted in an average estimation error of 29%, a significant improvement over both camera-based errors (43%) and WiFi-based errors (226%). In absolute terms, b-match ensured that its estimation error was less than 1-2 individuals 75% of the time.

Overall, we believe that our results demonstrate the superiority of b-match techniques for fusing WiFi and camera-based data. While camera-based approaches can sometimes be competitive, they require a far more extensive deployment to eliminate blind spots and to restrict each camera's monitoring zone to relatively short (3-4 meter) distances.

VII. Conclusion

In this paper, we have described and evaluated an approach for accurate human occupancy estimation in indoor public spaces that fuses together (a) location estimates of WiFi-enabled devices, obtained via passive infrastructure-side monitoring of WiFi transmissions, and (b) location coordinates of human subjects obtained via executing state-of-the-art DNN-based object detectors on camera-generated images. The main challenge is to overcome the non-negligible errors that arise from various real-world artefacts, such as the use of multiple personal devices by a single individual and the presence of partial/full views of humans in camera images. To robustly and accurately map DNN-extracted image human bounding boxes to physical coordinates, we proposed an enhanced regressor-based mechanism that works across a variety of {full body, torso, head} views of humans at different distances. Subsequently, we introduced our novel inexact bipartite graph-matching algorithm to match human objects to one or more WiFi devices, as part of an occupancy estimation strategy that is especially resilient to camera blind spots and occlusions. Experimental studies conducted over an approx. $85m^2$ indoor multi-occupant space shows that b-match (with Aggregation) is able to estimate the occupancy count with an error of $\sim 16\%$, across a wide variety of crowd levels, human movement behavior and ambient lighting conditions; over a smaller $25m^2$ region, the estimation error degrades only modestly to $\leq$30%.

Our ongoing work is investigating several possible directions of improvement, including: (a) the refinement of such estimation based on analysis of historical occupancy patterns, and (b) the use of more sophisticated "temporal-smoothing" mechanisms to both eliminate the inclusion of transient individuals and overcome the false negatives and positives of the visual object detector.

Acknowledgment

This research is supported by the National Research Foundation, Singapore under its International Research Centres in Singapore Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

[1] A. Ruzzelli, "BuildSys'10 - Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings: Message from the general chair," *BuildSys'10 - Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, pp. 1–6, 2010.

[2] U. Saralegui, M. Antón, O. Arbelaitz, and J. Muguerza, "Smart meeting room usage information and prediction by modelling occupancy profiles," *Sensors*, vol. 19, no. 2, p. 353, Jan 2019. [Online]. Available: http://dx.doi.org/10.3390/s19020353

[3] 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*, ser. SimAUD '11. San Diego, CA, USA: Society for Computer Simulation International, 2011, p. 141–148.

[4] S. Pan, A. Bonde, J. Jing, L. Zhang, P. Zhang, and H. Y. Noh, "Boes: Building occupancy estimation system using sparse ambient vibration monitoring," vol. 9061, 04 2014, p. 90611O.

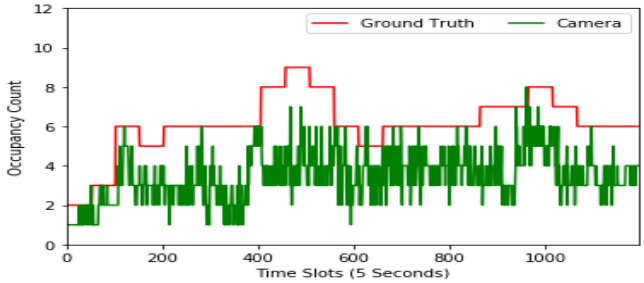
[5] Z. Wu, E. Jedari, R. Muscedere, and R. Rashidzadeh, "Improved particle filter based on WLAN RSSI fingerprinting and smart sensors for indoor localization," *Computer Communications*, vol. 83, pp. 64–71, 2016.

[6] J. Zou, Q. Zhao, W. Yang, and F. Wang, "Occupancy detection in the office by analyzing surveillance videos and its application to building energy conservation," *Energy and Buildings*, vol. 152, pp. 385–398, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.enbuild.2017.07.064

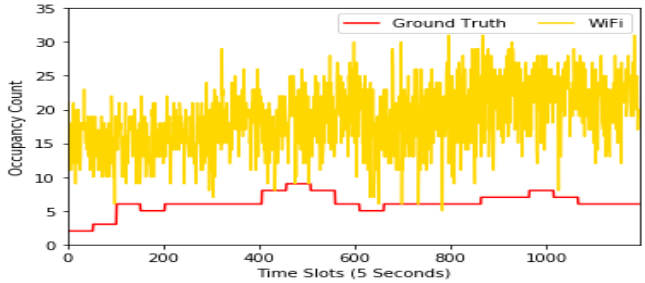
[7] R. K. Balan, A. Misra, and Y. Lee, "Livelabs: Building an in-situ real-time mobile experimentation testbed," in *Proceedings of the 15th Workshop on Mobile Computing Systems and Applications*, ser. HotMobile '14, 2014.

[8] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018. [Online]. Available: http://arxiv.org/abs/1804.02767

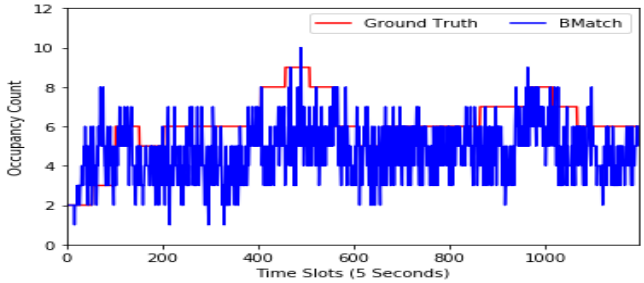
[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," *CoRR*, vol. abs/1512.02325, 2015. [Online]. Available: http://arxiv.org/abs/1512.02325



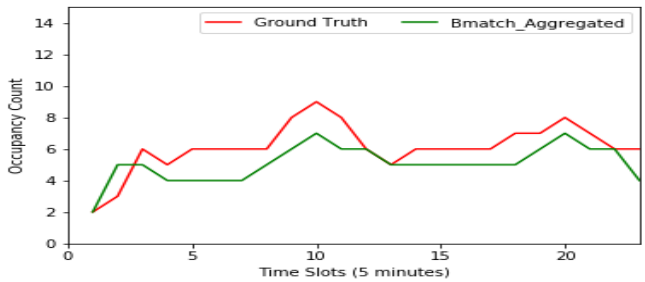
Ground Truth Vs Camera Occupancy



Ground Truth vs WiFi Occupancy



Ground Truth vs Fusion Occupancy



Ground Truth vs Fusion(Aggregated)

Fig. 9: Results for Small Region (High Occupancy)

[10] B. Balaji, J. Xu, A. Nwokafor, R. Gupta, and Y. Agarwal, "Sentinel: Occupancy based hvac actuation using existing wifi infrastructure within commercial buildings," in *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '13. New York, NY, USA: Association for Computing Machinery, 2013. [Online]. Available: https://doi.org/10.1145/2517351.2517370

[11] M. Milenkovic and O. Amft, "Recognizing energy-related activities using sensors commonly installed in office buildings," *Procedia Computer Science*, vol. 19, no. Seit, pp. 669–677, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.procs.2013.06.089

[12] H. Lee, C.-h. Choi, and M. Sung, "Development of a dimming lighting control system using general illumination and location-awareness technology," *Energies*, vol. 11, 11 2018.

[13] S. Van Den Berghe, M. Weyn, V. Spruyt, and A. Ledda, "Fusing camera and Wi-Fi sensors for opportunistic localization," *UBICOMM 2011 - 5th International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies; PECES 2011 - 3rd International Workshop on Pervasive Computing in Embedded Systems*, no. January, pp. 169–174, 2011.

[14] J. D. Domingo, J. Gómez-García-Bermejo, E. Zalama, C. Cerrada, and E. Valero, "Integration of computer vision and wireless networks to provide indoor positioning," *Sensors (Switzerland)*, vol. 19, no. 24, pp. 1–17, 2019.

[15] K. Padmanabh, A. Malikarjuna V, S. Sen, S. P. Katru, A. Kumar, C. Sai Pawankumar, S. K. Vuppala, and S. Paul, "ISense: A wireless sensor network based conference room management system," *BUILDSYS 2009 - Proceedings of the 1st ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings, Held in Conjunction with ACM SenSys 2009*, no. January, pp. 37–42, 2009.

[16] Y. Agarwal, B. Balaji, R. Gupta, J. Lyles, M. Wei, and T. Weng, "Occupancy-Driven Energy Management for Smart Building Automation," in *BuildSys'10 - Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, 2010, pp. 1–6.

[17] T. Weng and Y. Agarwal, "IEEE DESIGN AND TEST, SPECIAL ISSUE ON GREEN BUILDINGS 1 From Buildings to Smart Buildings-Sensing and Actuation to Improve Energy Efficiency," pp. 1–6, 2012. [Online]. Available: https://www.synergylabs.org/files/Weng_ieeeDT12_SmartBuildings.pdf

[18] Q. Zhu, Z. Chen, M. K. Masood, and Y. C. Soh, "Occupancy estimation with environmental sensing via non-iterative LRF feature learning in time and frequency domains," *Energy and Buildings*, vol. 141, pp. 125–133, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.enbuild.2017.01.057

[19] L. Yang, K. Ting, and M. B. Srivastava, "Inferring occupancy from opportunistically available sensor data," *2014 IEEE International Conference on Pervasive Computing and Communications, PerCom 2014*, pp. 60–68, 2014.

[20] M. V. Moreno, B. Úbeda, A. F. Skarmeta, and M. A. Zamora, "How can we tackle energy efficiency in iot based smart buildings?" *Sensors (Switzerland)*, vol. 14, no. 6, pp. 9582–9614, 2014.

[21] B. Li, J. Zhang, Z. Zhang, and Y. Xu, "A people counting method based on head detection and tracking," in *2014 International Conference on Smart Computing*, Nov 2014, pp. 136–141.

[22] V. L. Erickson, S. Achleitner, and A. E. Cerpa, "POEM: Power-efficient occupancy-based energy management system," *IPSN 2013 - Proceedings of the 12th International Conference on Information Processing in Sensor Networks, Part of CPSWeek 2013*, pp. 203–216, 2013.

[23] E. Soltanaghaei, A. Kalyanaraman, and K. Whitehouse, "Poster: Occupancy state detection using WiFi signals," *MobiSys 2017 - Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, no. June 2017, p. 161, 2017.

[24] F. Wang, Q. Feng, Z. Chen, Q. Zhao, Z. Cheng, J. Zou, Y. Zhang, J. Mai, Y. Li, and H. Reeve, "Predictive control of indoor environment using occupant number detected by video data and CO2 concentration," *Energy and Buildings*, vol. 145, pp. 155–162, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.enbuild.2017.04.014

[25] A. Networks, "Rtls - integrating with the rtls data feed," pp. 1–14, 2012. [Online]. Available: https://community.arubanetworks.com/aruba/attachments/aruba/unified-wired-wireless-access/23715/1/RTLS_integrationv6.docx

[26] P. Bahl and V. N. Padmanabhan, "Radar: an in-building rf-based user location and tracking system," in *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064)*, vol. 2, March 2000, pp. 775–784 vol.2.

[27] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018.

[28] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Deep image homography estimation," *CoRR*, vol. abs/1606.03798, 2016. [Online]. Available: http://arxiv.org/abs/1606.03798

[29] Pranoyr, "Head Detection using YOLO." [Online]. Available: https://github.com/biankatpas/head-detection-using-yolo