IEICE TRANSACTIONS

on Information and Systems

DOI:10.1587/transinf.2023EDP7106

Publicized:2024/04/26

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

A mmWave sensor and camera fusion system for indoor occupancy detection and tracking

Shenglei LI[†a)], **Haoran LUO**[†], **Tengfei SHAO**[†], ***Nonmembers***, ***and*** **Reiko HISHIYAMA**[†], ***Member***

SUMMARY Automatic detection and recognition systems have numerous applications in smart city implementation. Despite the accuracy and widespread use of device-based and optical methods, several issues remain. These include device limitations, environmental limitations, and privacy concerns. The FMWC sensor can overcome these issues to detect and track moving people accurately in commercial environments. However, single-chip mmWave sensor solutions might struggle to recognize standing and sitting people due to the necessary static removal module. To address these issues, we propose a real-time indoor people detection and tracking fusion system using mmWave radar and cameras. The proposed fusion system approaches an overall detection accuracy of 93.8 % with a median position error of 1.7 m in a commercial environment. Compared to our single-chip mmWave radar solution addressing an overall accuracy of 83.5 % for walking people, it performs better in detecting individual stillness, which may feed the security needs in retail. This system visualizes customer information, including trajectories and the number of people. It helps commercial environments prevent crowds during the COVID-19 pandemic and analyze customer visiting patterns for efficient management and marketing. Powered by an IoT platform, the system can be deployed in the cloud for easy large-scale implementation.

key words: *Sections; mmWave radar, camera, occupancy detection, tracking*

1. Introduction

With the implementation of automatic and smart space in Society 5.0, the automatic detection and recognition of people are becoming increasingly essential. The critical information needed for customized services and automation is to allow the space to recognize people and know where and how many of them are. Based on such information, better user analysis, as well as sustainable and emerging services, can be facilitated: these include services for Heating, Ventilation, Air Conditioning (HVAC), online-to-offline (O2O), visiting pattern analysis and social distance keeping [1]. A highly accurate people-recognizing system is urgently required to seamlessly integrate the above services without active human efforts and good user acceptance (non-intrusive), as shown as Figure 1.

Currently, the most commonly used method for indoor occupancy detection relies on devices. However, device-free methods are becoming increasingly attractive. However, the effectiveness of popular wearable devices, such as smartphones, smartwatches, and smart bands, is limited, as they depend on users constantly carrying them, which may not always be the case [2], [3]. The passive infra-red (PIR) sensor has been the most widely implemented device-free solution in past years. However, it may fail to recognize people moving slowly or crawling [4]. Advances in computational power, big data, machine learning, and deep learning have significantly enhanced the accuracy and adoption of optical methods [5]–[7]. Some of these methods, such as RGB-D cameras (stereo cameras) and Lidar, can provide accurate recognition and tracking but may be demanding to implement on a large scale due to their limited range, narrow tracking view, and high cost. Other camera solutions using web cameras or Closed-Circuit Television (CCTV) cameras may be more feasible for widespread implementation but share the drawbacks of optical methods. For example, they need a clear view relying on suitable lighting conditions and line-of-sight. These methods may fail in challenging scenarios like darkness, smoke, or obstruction. Moreover, they may raise privacy concerns and face user resistance due to their intrusive nature. Wireless sensings, such as Radiofrequency Identification (RFID) and Frequency Modulated Continuous Wave (FMCW), are state-of-the-art techniques for next-generation human detection and activity monitoring for multiple people. WiFi channel state information (WiFi CSI) is capable of subject count and activities performed by multiple people [8]. However, only the people walking between the separate transmitter (Tx) and receiver (Rx) could be detected, and it also needs some help in tracking multiple people in the same scene. mmWave sensor applications with excellent range detection performance used to be mainly geared toward the automotive market. Due to its wall-penetrated and unobtrusive nature, it is possible to place such sensors under thin walls or furniture, leading to less user resistance [9], [10]. Researchers documented that a single-chip solu-

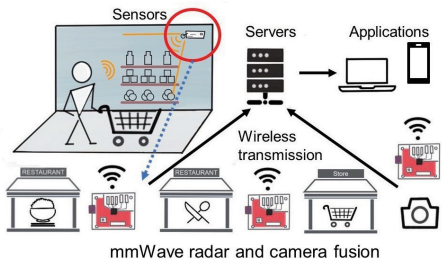


Fig. 1 Sensor-based occupancy detection IoT system with network structure for profiling analysis of customer visit patterns to provide O2O service.

[†]The author is with the Graduate School of Creative Science and Engineering, Waseda University, Tokyo, 169–8555 Japan.
a) E-mail: shenglei.lee@toki.waseda.jp

(a) Object tracking using mmWave tracker

(b) Object detection using Improved YOLOv3

Fig. 2 An example of complementary properties of mmWave radar and vision-based detection systems are illustrated in their respective capabilities and limitations. While mmWave radar is effective in challenging scenarios such as darkness, it struggles to differentiate between individuals walking or standing closely due to sparse data and noise. On the other hand, vision-based detectors accurately estimate objects in suitable-light conditions but fail in low-light and other challenging environments.

tion could provide detection accuracy from 50 % to 90 % for one to a dozen people [2], [3], [11]–[14]. However, the accuracy falls substantially when the number of people in the same scene increases. Additionally, the performance in detecting stationary people still is unfavorable. Figure 2 illustrates the complementary properties of mmWave radar and camera, based on our previous works. Figure 2(a) shows people detection and tracking using a single-chip mmWave radar solution, while Figure 2(b) shows camera-based people recognition using Convolutional Neural Networks (CNNs)) and You Only Look Once (YOLO). As can be seen, the radar sensor detects and localizes targets, but the vision-based detector fails due to insufficient illumination. Besides, the vision-based detector successfully recognizes people walking and standing closely, but the radar sensor fails to separate them. Neither a single-chip radar sensor solution nor an optical method satisfactorily meets the need for automatic detection in smart spaces. Therefore, a fusion system of two sensing modalities should be considered and motivated.

Conventional fusion methods of radar and camera, predominantly applied in the automotive industry, rely on the standard Kalman Filter (KF) [15], [16] and its variants, such as the Interval Kalman Filter (IKF) [17] and the Two-Stage Kalman Filter [18]. Despite their effectiveness, these methods often simplify radar detections to point detections and share common drawbacks like the need for accurate object modeling and extensive calibration. Alternatively, neural network models, with their increasing layers, show promise in complex problem-solving and classification tasks, prompting researchers to explore their use in enhancing the accuracy and generalizability of mmWave radar and vision fusion systems. Notably, initiatives, like Millieye[22], have adopted models, originally successful in other domains, such as CNNs, Long Short-Term Memory (LSTM) networks, and variants of YOLO. These models are applied to the integrative processing of radar and vision data, as documented in [19]–[22]. However, the high computational resource demands of these deep learning-based methods pose limitations for deployment in small or portable devices, potentially restricting their application in popular smart space devices like Amazon Echo, Google Nest, and Mi Home.

In this case, we propose an IoT platform-based fusion system using mmWave radar and vision to detect and track people indoors. Based on both the vision-based detector and mmWave radar tracker, it could be functional under challenging scenarios, such as darkness, smoke, and individual remaining stillness, where either the image-only methods [5], [6], [23] or the single-chip mmWave radar solution [2], [3], [11], [13] may fail. Since we apply an IoT platform and use a result-level fusion strategy, it needs less local calculation overhead than the conventional radar and camera fusion system, mainly for the automotive market. [19]–[22], [24]–[26]. By visualizing the real-time positions and the number of individuals within a space, the system allows users to observe visiting patterns, crowd dynamics, and customer preferences in smart environments. This capability underscores the system's utility in facilitating crowd control, enhancing security, and optimizing seating or business operations for restaurants and retail sectors amidst the COVID-19 pandemic. Leveraging a 60-64 GHz radar, Raspberry Pi, and the IoT platform, the system offers a significant reduction in cost and size compared to traditional setups that employ 70-74 GHz radars with PC or laptop backends [14]. These improvements not only lower the overall expense and footprint of the system but also increase its adaptability across various applications, thereby supporting extensive deployment. The system design adheres to the philosophy of privacy protection by minimizing the collection and processing of personal data [27]. In commercial environment tests, user acceptance has been improved by reducing the intrusive nature of the vision detector. For example, it excludes individuals' facial features and blends into the environment, thereby reducing the sensation of being surveilled. In brief, the main contributions of this work are concluded as follows:

- We proposed a mmWave radar sensor and camera fusion system for indoor occupancy detection and tracking in smart space. It helps to address the drawbacks of the single-chip mmWave radar system that may fail to detect stationary individuals and overcome some environmental limitations of the vision-based detector.
- We conducted extensive experiments, including in actual commercial environments, to evaluate the performance and adaptability of the proposed system. Compared to the single-chip mmWave radar solution with an overall accuracy of 70 $\sim$ 84% [2], [11], [13], [14] and our first version using cameras as sub-sensors with 93 % accuracy under suitable lighting conditions [28], the proposed system achieved 93.8% in a commercial environment including challenging scenarios. It also

demonstrated the same level of tracking accuracy with a median error of 0.17 m and better adaptability than other fusion systems that require heavy calibration since the computation could be carried out on the IoT platform.

The remainder of this paper is organized as follows: Section 2 introduces the background of this category; Section 3 describes the detailed architecture of the proposed systems; Section 4 shows the experimental configurations; Section 5 provides evaluations of the system and discussions; Finally, Section 6 concludes the paper.

2. Background

2.1 mmWave radar sensor detection and tracking

Millimeter-wave (mmWave) radars, operating in the 30-300 GHz range, use an active transceiver for object detection and tracking by measuring signal time delays [29]. As shown in Figure 3, our Texas Instruments mmWave radar transmits a chirp, a sinusoidal signal with a linearly increasing frequency over time characterized by its start frequency, bandwidth, and duration, to extract objects' range, radial velocity, and angle relative to the radar receiver through standard post-processing steps, including range, Doppler, and Angle of Arrival (AOA) estimation. The radar's Intermediate Frequency (IF) signals, obtained by calculating the frequency differences between transmitted and received signals, are analyzed with the Fast Fourier Transform (FFT) to generate a frequency spectrum, where each peak indicates an obstacle's range. Utilizing a 3×4 Multiple-Input-Multiple-Output (MIMO) array from three transmitters (TX) and four receivers (RX), it estimates the Angle of Arrival (AOA) through phase differences, achieving angular resolutions of approximately 14° azimuth and 57° elevation. This system generates a point cloud, with each point representing an object's (x,y,z) coordinate and radial velocity regarding the radar, although with potential noise, offering valuable occupancy tracking data in smart spaces.

2.2 Vision-based detection and classification

Vision detection methods are broadly categorized into one-stage and two-stage detectors. One-stage detectors, such as Single Shot Detector (SSD) [30], YOLO [5], [16],, and its variants, directly perform regression and classification on predefined anchor boxes, offering a low computational overhead suitable for edge and embedded devices, real-time services, and large-scale implementations. Conversely, two-stage detectors like Faster Region-CNN (Faster RCNN) [7], [31] and Region-based Fully CN (R-FCN) [32] incorporate a distinct module for generating proposal regions before conducting separate object classification. Unlike one-stage detectors that employ a direct learning approach for classification and localization, two-stage detectors initiate with a Region Proposal Network (RPN) to generate object proposals, each with an objectness score. These proposals are then classified in a separate stage, utilizing a RoI pooling layer to crop and resize regions from convolutional feature maps. Cropped regions are processed through a fully connected layer to yield fixed-length feature vectors, which are subsequently input to two sibling output layers: one for softmax probability estimates across object classes and another for bounding-box refinements of each proposal. This method achieves greater object detection accuracy but is more computationally demanding than one-stage detectors.

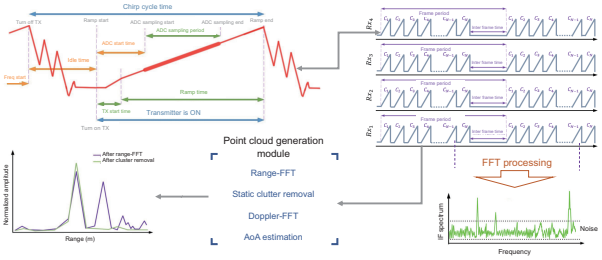


Fig. 3 Flow of radar data through FFT processing.

3. System Design

3.1 Overview

The proposed system comprises three modules: a mmWave radar-based object tracker, a vision-based detector, and a refinement head. The separated weight pre-training of the vision-based detector reduces the reliance on extensive labeled radar-vision data. Compared to the one-stage detector, such as the Tiny YOLOv3 [33] used in our previous work [28], the two-stage detectors employed in this system include an additional refinement step to increase accuracy. As described in Section 2.1, the radar detector provides information about indoor occupancy and object trajectories. It can assist the refinement module in better distinguishing objects in the smart space and help with challenging scenarios where the vision-based detector may fail, such as darkness or exposure that may cause a dim view. Additionally, the vision-based detector can help handle challenging scenarios where the single-chip mmWave radar detector may fail, such as when individuals remain still or when multiple occupants are close to each other or at the boundary of the indoor smart space.

An overview of the proposed system is presented in Figure 4. The system follows a two-stage pipeline from a system-level perspective. In the first stage, aggregation of Box Proposals $N = \{N_c, N_r\} = \{n_k\}_{k=1}^{K}$, where $N = N_c, N_r$ are the box proposals from the camera detector and radar tracker, respectively, and K is the total number of RoIs. Then the Local Feature Extraction per RoI for camera and radar, L_c,L_r can be obtained by feature extraction $L_c, L_r = \text{Cropping}(G_c, G_r; N)$. from the global multi-modality feature maps for camera detector (G_c) and radar

tracker(G_r). After removing redundant bounding boxes, local features are obtained by cropping the global features based on their positions using a Region of Interest (RoI) layer in the second stage. A refinement head estimates a new location for each box within the frame and assigns it a confidence score. Individuals who are walking undergo the two stages mentioned above. However, stationary individuals may be detected only by the vision-based detector, as the mmWave radar tracker might fail to consistently detect them.

3.2 mmWave radar tracker

Due to the potential for the vision-based detector to fail in generating confident object detection under harsh conditions such as darkness, exposure, and non-line-of-sight in indoor environments, a radar tracker is proposed to generate desirable detection based on point cloud data. In the radar tracker module, a mmWave radar sensor is employed, as introduced in Section 2.1, to obtain estimated features based on their unique properties measured by the time delay between the transmission and reception of the pulse. The linearly increased frequency ramp of the periodically transmitted mmWave signal $T_r(t)$ is commonly recognized as a chirp. Multiple chips are usually emitted to measure the range and velocity information of the target. The transmitted signal $T_r(t)$ and receiver-captured signal $R_r(t)$ could be donated as:

$$T_r(t) = A_1 e^{j(2\pi(f_c)t+\pi\frac{B}{T}t^2)} \tag{1}$$

$$R_r(t) = A_2 e^{j\left(2\pi f_c(t-t_d)+\pi\frac{B}{T}(t-t_d)^2\right)} \tag{2}$$

where f_c, B, T_c are the start frequency, bandwidth, duration and t_d is the time delay of the corresponding signal reflected off the human body. A_1 and A_2 represent the amplitudes before and after propagation and circuit losses, respectively, which are approximately equivalent due to the short detection range within 10 meters. This radar tracker module is based on previous work [14] with similar approaches adopted in [2], [13]. The data processing of mmWave radar includes point cloud generation, clustering, and tracking. These steps can be summarized as follows.

3.2.1 Point cloud generation

The point cloud generation module is based on range-FFT, clutter removal, Doppler-FFT, and angle of arrival (AoA) estimation. Clutter removal filters out stationary obstacles from the scene using range information processed by range-FFT. In each frame, radar data comprise a set of points that include (x, y, z) coordinates and radial velocity.

$$p_i := [x_i, y_i, z_i, v_i] \in R^4 \tag{3}$$

However, reflections from occluded areas change over time as people move, resulting in noise. The contaminated radar point could contain clutter and noise signals that trigger failure and confusion in object detection.

3.2.2 Clustering

To distinguish the points of the foreground targets from the clutter and noise, we employ the clustering module from our previous work [14]. The primary procedures involve randomly selecting a point that does not belong to a cluster or is an outlier. The point is then classified as a core point or not based on its distance and radial velocity to its neighbors. The mean of the cluster is then recalculated as a new centroid or marked as noise. The cluster is expanded by adding reachable points until an outlier is added. The added outlier is marked as a boundary point, and the above steps are repeated. Outliers (noise) are filtered out, and the clusters are passed to the next module. This module could perform better on the varying density point cloud and less processing time than the DBSCAN and DBmeans used in the related scholars [2], [11], [13], [22]. Unlike K-means [34], which requires prior information about the number of clusters, this method can detect an arbitrary number of targets without such information. The distance between point i and point j is given by

$$d_{ij} = W \times [p_i - p_j]^\intercal, W \in R^{1\times 4} \tag{4}$$

where p_i is the (x, y, z) coordinate and radial velocity of point i, W is the weight vector to optimize the contribution of each parameter.

3.2.3 Box proposal and Tracking

Each point passing through the clustering module is labeled by either an index of a cluster or a flag of noise. After the noise is removed, the position of each cluster's centroid can be estimated. Based on the distance from the boundary points of each cluster to its centroid, an approximated box representing each cluster can be proposed.

In the tracking module, a Kalman Filter [35] calculates the prior state and covariance estimation for track prediction. We employ the Hungarian method [36] to associate multiple clusters generated in each frame and across frames with multiple tracks. A matrix is constructed using the Euclidean distances between the centroids of tracks and detected objects in the current frame. The cost matrix needs to add dummy rows/columns because the number of detections and tracks is only sometimes matched. This approach shows good performance in multi-detections and reduces the effect of flicker in frames for temporally consistent tracks. If detection is outside all gates, it is sent back to initialization for a new track iteratively. After the number of points associated with a track exceeds the threshold, it is transmitted to the updating step, where we apply an Extended Kalman Filter (EKF) to smooth the locations and sizes of clusters. The state of a track that has gone through the above period is changed or deleted based on inactivity.

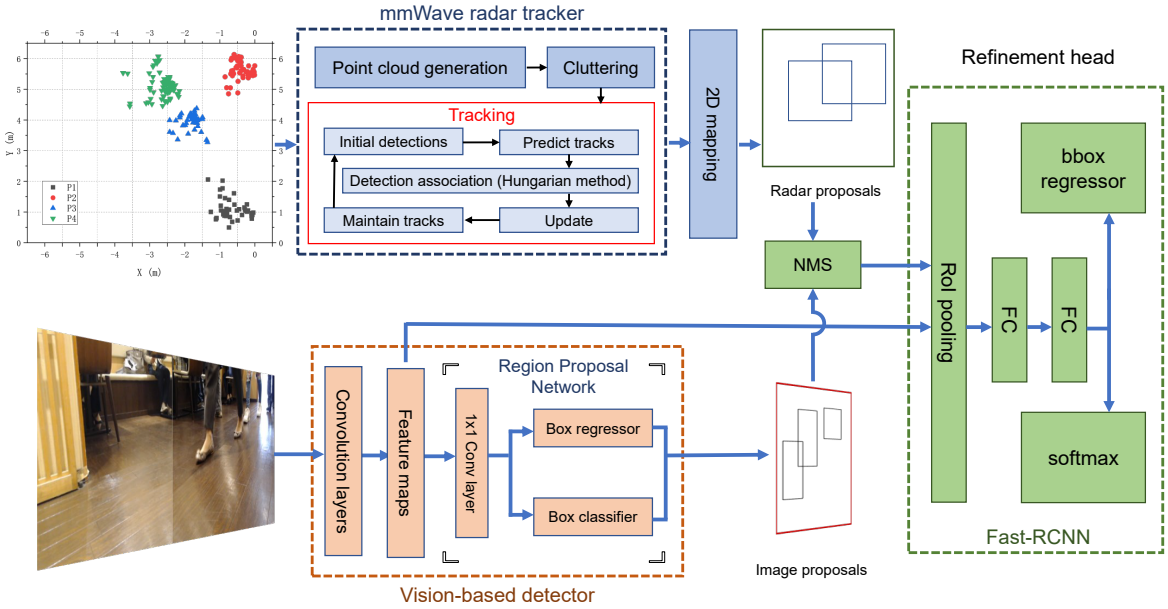


Fig. 4 The proposed network architecture. Inputs to the network are radar point cloud and camera image. The proposals generated from radar point cloud are fused with image features to improve box localization.

The proposed boxes are shadowed on the 2D coordinate plane with boundaries having the same color as their respective trajectories. Additionally, the 2D images of the boxes serve for the radar tracker to fuse with the vision-based detector at the same timestamp.

3.3 Vision-based detector

The mmWave radar tracker may experience difficulties in generating accurate estimation boxes, particularly when its performance is impaired by the state of targets. This impairment is often associated with the clutter removal feature in the point cloud generation module (refer to Section 3.2.1 and Figure 3 for details). Drawing on related studies[2], [11], [13], as well as our previous work[14], it has been observed that this feature might inadvertently filter out stationary individuals, especially in settings like restaurants or shops where people tend to sit or stand for extended periods.

In this case, a two-stage vision-based detector based on CNN is utilized for addressing these scenarios in the smart space. As described in Section 2.2, it detects targets and proposes bounding boxes with category and confidence scores if providing a clear view of the objects. Similar to the one-stage vision-based detector used in our previous work [28], the two-stage detector follows the archetype consisting of a feature extractor, feature maps, and a head network. A feature extractor, which typically consists of convolutional, activation, and pooling layers, takes an image as input and outputs a set of feature maps. Such feature maps are then processed for the head network to generate a set of boxes representing the location of objects in the image. The difference is that the two-stage detector conducts a Region Proposal Network (RPN) to efficiently scan the image to assess whether further processing needs to occur in a given region. In contrast, the one-stage detector does not need such RP. Outputs from this two-stage detector, after being filtered by a confidence threshold, are merged with the 2D estimations from the radar tracker to form comprehensive estimations. These are then refined through a result-level strategy within the refinement head module

3.4 Refinement head

In the refinement head module, the proposals from both the radar tracker and vision-based detector are merged for the second stage of detection. NonMaximum Suppression (NMS) is employed to merge highly overlapped redundant proposals before moving on to the next stage. Note that vision-based proposals are less reliable in ranging distance compared to radar-based proposals. It is due to the fact that radar detection is based on signal transmission and reception time delay, while vision-based detection typically relies on 2D images without depth information. However, NMS typically removes overlapping proposals without discriminating based on such characteristics. In this case, matching proposals are first identified using an Intersection over Union (IoU) threshold. Then, range information measured by the radar is used for these matching proposals. The bounding box offset is learned as a regression, and the Euclidean loss

is employed for each candidate. Then, the remaining proposals, regardless of their origin, are fed to the second stage of detection network.

The second stage of detection network is based on Fast R-CNN [7]. Based on the inputted feature map from the remaining proposals, every single object proposal would be cropped from the feature map. The feature vector of the same size Input the feature map from the remaining proposals, and crop every object proposal from the feature map. Then, a RoI pooling layer extracts a feature vector with the same size for each object proposal from the feature map. Process the feature vectors in a sequence of fully connected layers and pass to the softmax bounding box regression layers. Out the category classification and bounding box regression for each proposal. Note that, in this work, only the people would be classified and detected due to the commercial environments required, and the trajectory of the estimations would be shown on the manual inputted 2D map of the smart space. The loss function follows the real-time Fast R-CNN [31] using a multi-task loss as objective function.

$$L(p_i, t_i) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \tag{5}$$

where i represents the anchor index, while p_i denotes the predicted probability of the i^{th} anchor. The value of p_i^* is determined by whether the anchor is positive (1) or negative (0). The vector ti represents the four parameterized coordinates of the estimated bounding box, with t_i^* representing the ground-truth box. **Classification Loss** (L_{cls}): This component of the loss function is computed using a log loss over two classes (object vs. not object). It's calculated for each anchor and is used to classify whether an anchor is an object or not. $L_{\text{reg}}(t_i, t_i^*) = R(t_i - t_i^*)$ where R is the robust loss function smooth $L_1(x) = 0.5x^2$ if $|x| < 1$, otherwise $|x| - 0.5$. **Regression Loss** (L_{reg}): This part of the loss function involves bounding box regression, where a robust loss function (smooth L_1) is used. It's activated only for positive anchors and is disabled otherwise. This component computes the difference between the predicted bounding box and the ground-truth box. N_{cls} and N_{reg} are normalization parameters, and λ is a balance parameter weight the two terms.

4. Implementation

4.1 Experiment settings

In this paper, a data capture site includes a backend connecting to a mmWave radar , which is elevated by arms, and cameras hidden in the box to reduce customer concerns as lower as possible. Since the user's resistance to camera-involved systems is significant, and privacy concerns are sincerely respected in this paper. The camera in our system is oriented towards the legs of the users. It ensures that only

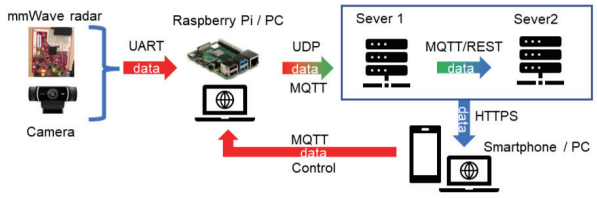


Fig. 5 IoT platform designed for the proposed system. It allows remote controlling and combining multiple data-capture sites.

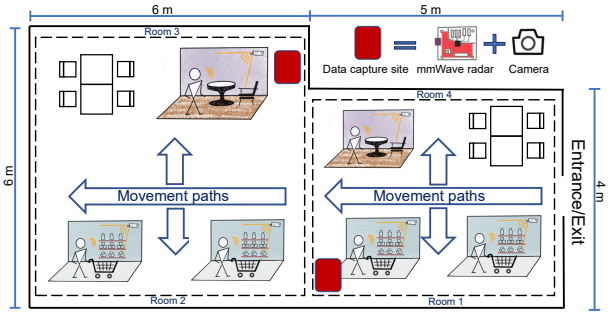


Fig. 6 The layout of the commercial spaces for experiments.

the lower body part is captured by our cameras, excluding any facial features used in other ordinary optical recognition methods. Due to the physical nature of mmWave radar and camera, each room needs one data capture site. For the mmWave radar sensor, we employ a low-cost COTS one (IWR6843ISK) from Texas Instruments. Both Raspberry Pi and laptop could be used as backends. Here we use a Raspberry Pi 3 (1.4GHz, 32GB RAM) to transmit data to the IoT platform built on AWS. The IoT platform is the same one from our previous works [14], [28], as shown in Figure 5. In Figure 6, the layout of the commercial environments utilized in the experiment is depicted, demonstrating the strategic positioning of radar and camera systems alongside the corridor, facing the main pathways of each room. The overhead group is situated at the entrance/exit points. The evaluation areas, measuring approximately $6m \times 6m$ and $4m \times 5m$, encompass the main routes. Individuals within these areas engage in various activities, including walking, standing, and sitting, along the main pathways where the data capture sites are located. Such smart spaces accommodate up to 40 customers and five staff. Customers predominantly occupy positions along both sides of a central passageway, extending linearly from the entrance to the far end. Due to the physical limitation of our mmWave radar, the maximum number of occupancy per scene is set to 10. The real-time visualization of the number and position of users in the smart space is valuable for analyzing the customer visiting pattern, crowd management, and other business impacts. Moreover, in light of significant user resistance to camera-based

systems, this work diligently addresses privacy concerns by implementing a system design that deliberately avoids collecting facial data and incorporates unobtrusive elements to minimize user resistance. Utilizing our indoor smart space occupancy detection and tracking systems, including our previous works [14], [28], the commercial space effectively reorganized its business schedule and limited customer capacity during the COVID-19 to mitigate crowds and maintain social distances.

4.2 Datasets

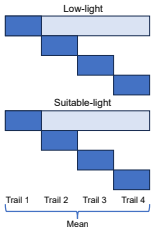
Three datasets are used in this work.

Microsoft COCO [37] is a comprehensive resource for object detection, segmentation, and captioning, featuring over 200,000 labeled images, which include 250,000 instances of people with keypoints. As a significant benchmark for object detection tasks, the Microsoft COCO dataset sees widespread use. However, it includes only 565 images captured in low-light conditions, constituting a mere 0.23% of its entire collection. To augment the representation of low-light data, we utilize a 6-class sub-dataset from COCO to cooperate with the low-light dataset, encompassing more than 40,000 images in total.

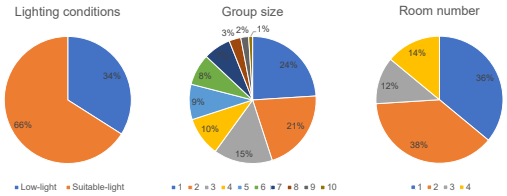
ExDark [38] is employed as the low-light dataset, which contains 7363 images to complement the dark scenarios. Same as the sub-dataset of COCO, a 6-class sub-dataset of Exdark, contained the exact same categories, is employed to enhance the performance of people detection in the commercial smart space. Each categories account for 14 % to 20 % of the total images to relatively distribute the sub-dataset.

Self-Collected Data is a single-class dataset of occupancy detection with 1400 frames captured in the pre-experiment and field experiments in indoor commercial environments. Figure 7(a) shows each data capture site has two cameras, one mmWave radar, setting to a sampling frequency of 30Hz, and key-frames at 4Hz. As shown in Figure 7(b), a 4-fold cross-validation paradigm is employed to obtain the average of four trails, since it is a small-scale dataset compared to the other two. A fundamental principle guiding our division into four folds is to ensure that the data in each fold are collected from distinct locations, thereby showcasing the model's ability to generalize. The average of four trails would be taken as results. As shown in Figure 7(c), the maximum number of people in the same scene can reach up to 10. Customers and staff are randomly walking, standing, and sitting in front of our data capture sites in four kinds of rooms in the restaurant and shop, as shown in Figure 6. Note that the private information of the customer and staff, such as the face, gender, and age, are not included in our dataset due to users' resistance to a camera-involved system in a commercial environment. The illumination level has an unbalanced share due to the necessary lighting conditions in business time. As a preprocessing step of the mmWave tracker involved system, the annotations and point cloud data first need to be transformed to the users' coordinates and then converted to their equivalent bounding boxes, as described in Section 3.4.

(a) Data capture site



(b) Four-fold sub-datasets



(c) The distribution of our dataset

Fig. 7 (a) 1080p cameras and 60-64GHz mmWave radars (TI IWR6843) are used for data collection. (b) Training data is represented by gray parts, and test data by blue parts. Each illumination level's dataset is divided into four parts. The final results are obtained by averaging four trials during evaluation. (c) The distribution of our dataset on lighting conditions, group size, and data collection locations.

Implementation The first two training stages are conducted on the mixed dataset of COCO [37] and ExDark [38], and the third training stage is conducted on our dataset. As recommended in [31], the classification loss is normalized by the mini-batch size N_{cls} and the bounding box regression loss by the number of anchor locations N_{reg}. This approach ensures losses are proportionally scaled based on the processed samples and anchor positions for box predictions. To ensure that both loss components contribute equally to the model's learning process, we use a balance parameter, λ, set to 10. This setting aims to equalize the influence of both the classification accuracy and the precision of the bounding boxes on the overall training performance. The NMS threshold has been set as 0.5 for all experiments.

4.3 mmWave radar configuration

The available bandwidth was 4 GHz, ranging from 60 to 64 GHz. The chirp cycle time (T_c) was 58.23 µs, and the frequency slope (S) was 77.73 GHz/ms. The maximum detection range was 6.0 m with a range resolution of 0.045 m and a velocity revolution of 0.25 m/s.

5. Evaluation and Discussion

5.1 Evaluation Metrics

5.1.1 Precision, Recall, F1 score , and Overall accuarcy

The precision and recall are given by:

$$Precision = \frac{TP}{TP + FP} \quad , \quad Recall = \frac{TP}{TP + FN} \tag{6}$$

where TP, FP, and FN are true positives, false positives, and false negatives, respectively. Precision measures the percentage of all instances where the positive prediction is actually positive. Recall measures the percentage of actual positive instances that are correctly classified. The sample's true or false attribute is determined by whether its IoU ($\frac{Intersectionarea}{Unionarea}$) between the ground truth boxes exceeds the threshold. The F1 score measures the trade-off between precision and recall, given by their harmonic mean ($\frac{2\times Precision\times Recall}{Precision+Recall}$). Maximize the F1 score implying maximizing both precision and recall simultaneously. The overall accuracy ($\frac{TP+FP}{Total}$) is employed to assess the performance of the proposed system and other related systems during field tests, as described in Section 5.3.1.

5.1.2 mean Average Precision (mAP) and confusion matrix

The mean Average Precision (mAP) serves as a widely used metric for object detection and classification, quantifying model accuracy by comparing predicted bounding boxes to ground truths. It is derived from the area under the precision-recall (P-R) curve, which is constructed from detections ranked by confidence. While lower confidence thresholds (e.g., 0.001) can theoretically maximize mAP by extending the P-R curve, they often lead to excessive false positives, making such thresholds impractical. Consequently, a moderate threshold is preferred to strike a balance between precision and recall, allowing for a more realistic evaluation of model performance through mAP and F1 scores across varied thresholds.

The confusion matrix is a commonly used tool for evaluating binary and multi-class classification. We employ it on our real-time data captured in the actual commercial environment, which consists of a single class of people but varies in the number of individuals present in each scene.

5.2 Performance on datasets

5.2.1 Overall performance

This section evaluates the performance of the proposed system on the COCO, ExDark, and field-captured datasets, using the improved Tiny-YOLOv3 and the Refinement head based on Faster R-CNN [7] as reference models.

Improved YOLOv3 is an enhanced version of Tiny-YOLOv3 serving as a vision-based detector in our previous work. It is lightweight and well-suited for use on micro-PCs, mobile devices, and distributed architectures, performing well in real-time IoT people detection services under limited internet conditions.

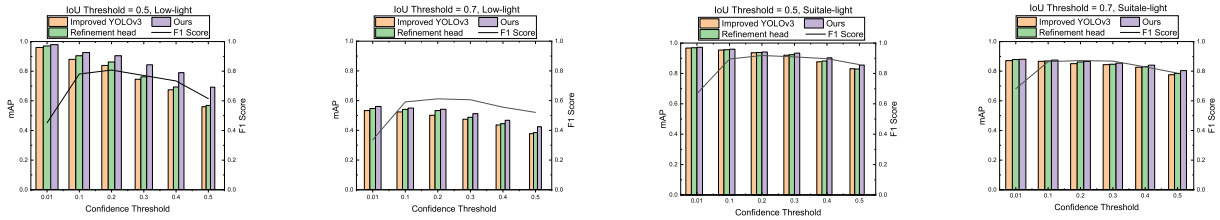
Refinement head based on Faster R-CNN, serves as the second stage of detection in this work. As a comparison, it replaces the decision-level fusion module (CNN) connected with the improved YOLOv3. In other words, it can be viewed as an enhancement of our first version [28], but without a mmWave radar tracker to demonstrate the effectiveness of the radar proposals in our fusion system. All models are pre-trained on COCO and ExDark before being tested on field-captured datasets. Both references are without a radar tracker and rely solely on images.

Figure 8 presents the comparative performance of various models on our field-captured datasets[†] focusing on the detection of individuals within groups of varying sizes across different environments.. The one-stage detector (Improved YOLOv3) and two-stage detector (Refinement head) exhibit performance on par with the mmWave-camera fusion system under suitable lighting conditions but are somewhat inferior in low-light scenarios. In suitable lighting, both the Improved YOLOv3 and Refinement head perform satisfactorily due to the simplicity of ourour people detection datasets in terms of density and diversity, thereby limiting the advantages of radar sensors in these conditions. Conversely, under low-light conditions, the fusion system leveraging radar notably surpasses both Improved YOLOv3 and Refinement head, which are trained on COCO and ExDark, particularly at an IoU threshold of 0.5. This underscores the capability of radar tracker to detect element that optical methods may overlook due to inherent physical constraints.. The differences between each method are minimal when the confidence threshold is as low as 0.1 but become more pronounced as the confidence threshold increases. When the IoU threshold is set as low as 0.5, the main contribution of improvement belongs to the radar-based tracker. Meanwhile, the refinement head module with NMS improves significantly compared to the one-stage detector, validating its importance in filtering and adjusting bounding box positions. As a result, the differences between each method are generally more minor than at higher IoU thresholds but follow a similar trend of increasing differences with raised confidence thresholds.

5.2.2 Performance on COCO and Exdark

This section evaluates the efficacy of the vision-based detector and refinement head on image-only datasets, including COCO and ExDark. For this purpose, the mmWave radar tracker was excluded, and the fusion system's remaining components were retrained, since there was no radar data in COCO and Exdark datasets. The diverse nature and minimal inter-category bias of these datasets enable a focused analysis of model performance while reducing the influence of dataset variations. Millieye[22], recognized for its effective CNN-based camera detection and DBSCAN-based radar tracking, serves as a baseline in our fusion system assessment. Other includes Improved YOLOv3 for consistency with prior comparison. Distinct training datasets delineate the reference models: Imp.YOLO_COCO and Imp.YOLO_ExDark were trained on COCO and ExDark datasets, respectively. Mean-

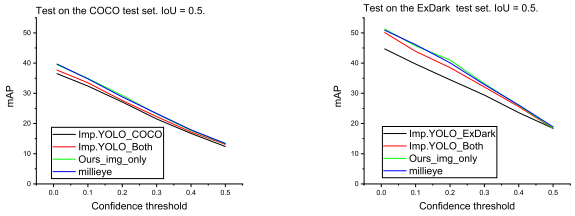
[†]The training set and the testing set are collected in different places to ensure variability and generalizability.



(a) IoU threshold = 0.5 and 0.7 in low-light condition.

(b) IoU threshold = 0.5 and 0.7 under suitable lighting.

Fig. 8 The comparison of mAP of the proposed system and reference group. The F1 score curve is derived from the vision-based detector. The horizontal coordinate is the confidence threshold of the vision-based detector, and the vertical coordinates are the mean Average Precision and F1 score, respectively.



(a) Test on the COCO test set.

(b) Test on the ExDark test set.

Fig. 9 mAP of different approaches on COCO and ExDark datasets under 0.5 IoU threshold.

while, millieye, Imp.YOLOv3_Both and Ours_img_only received training on both datasets. Figure 8 summarizes the mAP across 12 categories COCO and ExDark. Consistent with existing literature, models trained on more extensive datasets exhibited superior performance. . Notably, when the confidence threshold was set below 0.2, models trained on both COCO and ExDark demonstrated marked improvement compared to those trained on a single dataset. However, this enhancement was not as significant as anticipated for confidence thresholds above 0.2.

5.3 Performance on people detection and tracking task in field experiments

Different from the metrics discussed in previous sections, the overall performance highlighted in this section specifically pertains to the detection and tracking of individuals through field experiments conducted in real commercial environments near Waseda Campus, emphasizing random customer information. The radar tracker, unaffected by lighting conditions, achieves up to 99% accuracy for single targets. However, this accuracy drops to below 80% in scenarios involving approximately a dozen individuals. Conversely, the vision-based detector, trained on low-light datasets such as ExDark, enhances the detection of multiple targets under low-light conditions but may falter in minimal lighting. Consequently, the integration of these systems improves the overall performance, offering broader scenario coverage than either the radar tracker or vision-based detector alone. However, the accuracy under conditions below the specified light threshold remains unchanged, without enhancement.

5.3.1 Overall performance

As shown in the confusion matrix (Figure 10), the proposed system achieves an overall accuracy of 93.8% in scenarios involving up to 10 people in the same scene. The group size significantly affects the performance of radar detection, leading to a variance in overall accuracy in field tests compared to the results obtained from our dataset in the previous section. Notably, when customers are mindful of maintaining social distancing, the prevalence of smaller groups increases. The confusion matrix is calculated by comparing the predictions of different group sizes against actual observations, and the precision and accuracy are given by $Precision = \frac{TP}{TP+FP}$, $Accuracy = (\frac{TP}{Total})$. Where TP is true positive, FP is false positive, and Total is the total number of the predictions.

In comparison, single-chip mmWave radar solutions show accuracies of 71.1 % ([11]†) and 83.5 % [14] . Vision-based detectors have an accuracy of 91.4 % [7], [31]while our previous system integrating mmWave radar with a sub-sensor camera achieved 92.1% accuracy [28]. A commercial solution using a stereo camera for overhead counting reaches 94.0% accuracy under suitable lighting conditions for up to five individuals. However, the system may encounter difficulties in challenging conditions, such as overexposed environments. These limitations can be attributed to the physical constraints of the commercial cameras utilized in this work, as well as to the inadequacy of datasets in environmental dynamics. For a more detailed analysis of environmental dynamics and their impact, refer to Section 5.4.

5.3.2 Impact of number of people

In the mmWave radar people detection category, the number of people and their status substantially affect performance due to the sparse point cloud. Even when using the Hungarian method [36] to improve multi-target assignment in

†In this work, the single-chip mmWave radar system from Taxes Instrument demonstrates an overall accuracy of 71.1%, and 73.0% when the group size is smaller than four. Please refer to https://www.ti.com/sensors/mmwave-radar/overview.html for more information.

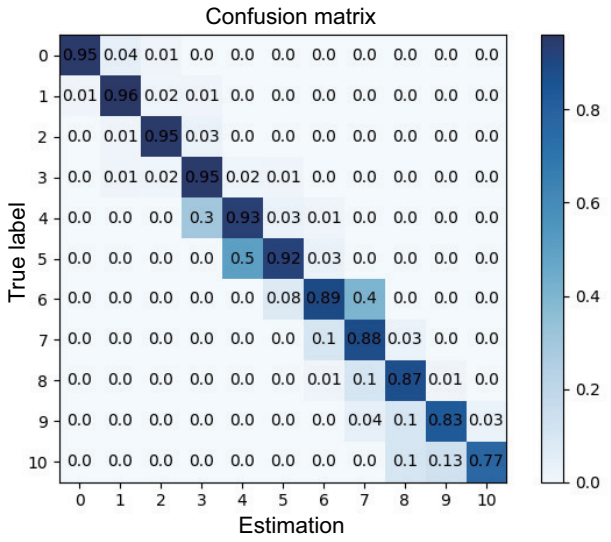


Fig. 10 Confusion matrix of 10 insiders.

single-chip solutions [14], the results are still unsatisfactory. In contrast, the vision-based detector can recognize dozens of objects but typically require line-of-sight and lighting conditions. As shown in Figure 11, the proposed fusion system performs well in detecting multiple people. Meanwhile, the single-chip mmWave radar systems [2], [11], [13], [14] experience a substantial decrease in accuracy when detecting more than four individuals in a smart space. mID [2] shows a solid performance but requires prior knowledge of insiders' walking information, while others do not. Vision-based detectors based on the front of view perform well and stably but fail to recognize individuals correctly when some parts of their bodies overlap. Such issues could be resolved by raising cameras like CCTV, but we found that this may raise more vital privacy concerns than placing cameras at lower heights with a limited view of individuals' legs and feet. One of our future goals is to find a balance between performance and user acceptance. Overhead counting systems also perform well in this task but cannot provide real-time information about users in a smart space. Additionally, the maximum number of people detected by this commercial system is limited to five, which is unsatisfactory for commercial environments with large numbers of simultaneous customers, such as restaurants and stations.

5.3.3 Tracking

With the aid of the mmWave radar tracker, the proposed system can provide accurate tracking and real-time user positioning. As the coordinates of users are solely derived from radar sensor data, our proposed system exhibits comparable accuracy and detection range to that of a single-chip mmWave radar solution [14], achieving a median error of 0.17 m in a detection range of 6 m. Similar to reference methods like TI and mID, false detection is excluded in tracking performance evaluation but analyzed in the impact of group size. In comparison, one of the state-of-the-art optical tracking methods, Kinect v2[†], achieves a median error of 0.88 m within a detection range of 4.5 m. Additionally, Kinect v2's error fluctuates between 0.7 m and 1.0 m, while the mmWave radar tracker-involved system provides much more stable and low-level tracking errors., as shown in Figure 12.

5.3.4 Standing and sitting individuals

Single-chip mmWave radar sensor solutions [2], [13], [14] primarily focus on detecting walking individuals due to the nature of radar sensors requiring static removal, which can significantly hinder the performance of detecting stationary people. To address such issues, the use of a camera-involved radar fusion system, commonly used in automotive applications, has gained attention in the indoor smart space category and provides users with more options. Table 1 summarizes the merits and demerits of the mmWave radar-camera fusion system compared to the single-chip mmWave radar systems and vision-based technologies. In comparison to single-chip solutions, the mmWave-camera fusion system shows significant improvement in accurately detecting both walking and stationary individuals. This function is essential for commercial environments such as restaurants, theaters, and waiting halls. However, it also shares the drawbacks of optical methods that require a clear view of the body, a minimum lighting level, and line-of-sight. These limitations will be discussed in the next section.

5.4 Robustness to Environmental Dynamic

5.4.1 Lighting

Darkness scenarios: Considering the substantial impact of lighting conditions on the efficacy of vision-based detectors, our evaluation includes scenarios of darkness within our mixed datasets derived from COCO and ExDark. For example, we consider environments like a restaurant at night, lit only by emergency exit lights. Table 2 presents the mAP results under such darkness scenarios, also comparing a reference group relying solely on a camera sensor. The proposed fusion system performs better than the vision-based detectors based on improved YOLOv3 and the refinement head using Fast R-CNN in such challenging scenarios. Furthermore, as detailed in Section 5.2.2 and depicted in Figure 9, vision-based detectors trained on datasets incorporating darkness scenarios show a marked improvement over those lacking exposure to such environments.

Unfit lighting: In the field tests, unfit lighting conditions (Unfit LT), including overexposure and low light, significantly impacted the performance of vision-based detectors (both front-view and overhead) but had minimal effect on the radar tracker. Table 3 summarizes the comparisons under

[†]According to `https://github.com/mcgi5sr2/kinect2_tracker`.

Table 1 Comparison of different people detection methods.

	single-chip mmWave solutions (TI, mID, cTracker, Li et al.)	Optical methods (Front-view, overhead)	Our radar-camera fusion system (No facial features)
Overall Accuracy	Moderate (70% ~ 90%)	Very Precision (90% ~ 96%)	Precision (93%)
Walking people	Moderate	Very Precision	Precision
Stationary people	Need improvement	Very Precision	Precision
Tracking	Precision	Moderate	Precision
Scale of crowd	Moderate	Very large (overhead: small)	Large
Env. limitations	Few	Many	Few
Privacy Concerns	None	High	Medium
User Acceptance	High	Low	Medium

*The improved privacy concern of our radar-camera fusion system is primarily due to its orientation towards users' legs. It ensured that only the lower body part was captured by the camera, excluding any facial features. The privacy concerns could reach a similar level as the refere -nced optical methods if facial features are involved.

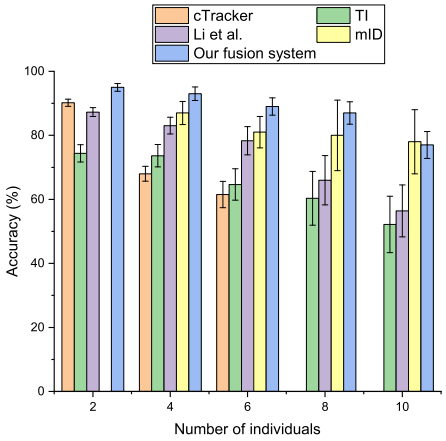


Fig. 11 Impact of the number of insiders between the single-chip mmWave radar system and the proposed camera-radar fusion system.

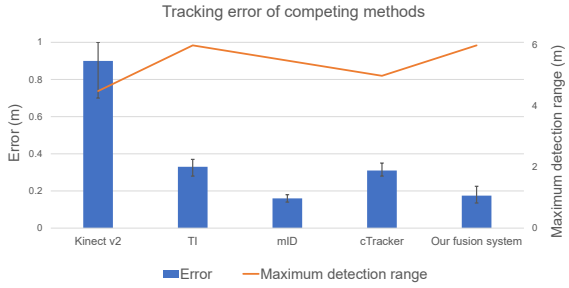


Fig. 12 Tracking performance of competing methods, including our proposed fusion system.

various conditions: The field test was conducted in a commercial indoor setting, where bright natural light was present at the entrance and windows during specific times, typically at noon and sunset. "Good light" denotes suitable indoor lighting conditions minimally influenced by natural light. "Unfit light" includes environments with high-intensity light exposure, as illustrated in Figure 2(b). "Darkness" pertains to scenarios where business operations have ceased, leaving only emergency or exit lighting active, as depicted in

Table 2 mAP of our proposed system and competing methods (IoU threshold: 0.5) under darkness scenarios.

Conf. Threshold	0.01	0.1	0.2	0.3	0.4	0.5
Improved YOLOv3	78.7	67.8	60.9	47.6	39.3	30.8
Refinement head	80.5	77.1	71.3	62.5	51.4	40.2
Ours	83.1	78.4	74.5	68.1	62.0	57.1

Table 3 Performance under different lighting conditions. It contains the overall accuracy of people detection from competing single-chip mmWave radar solutions, image-only detectors including Kinect v2, stereo camera (overhead), and our mmWave-radar fusion system. Unfit lighting (Unfit LT) includes natural environment dynamics such as overexposure and low-light during the business time.

	mmWaves (TI,mID,etc.)	RGBs (Front view)	RGB (Overhead)	Fusion (Ours)
Good LT	70% ~ 90%	90% ~ 96%	94%	93%
Unfit LT	Robust	80% ~ 90%	70% ~ 80%	93%
Darkness	Robust	Need help	Need help	90%

Table 4 Impact of different materials on the non-line-of-sight performance of mmWave radar. It is evaluated by the point cloud density change when different materials cover the radar. All included materials are commonly used in Chinese and Japanese architecture.

MATL.	Plastic	Xuan paper	Chipboard	Wood	ALUM
Diff.(%)	0.11	0.26	0.33	0.38	0.70

Figure 2. The proposed system Given that the overhead people counting system is ideally positioned at the entrance/exit of the smart space, the impact of natural lighting is more significant than that experienced by front-of-view optical methods and indoor radar setups. It has a range of results to enhance such influence. In summary, the vision-based detector's proficiency in accurately detecting stationary individuals affords the proposed system a performance edge over single-chip mmWave radar systems, as elaborated in Section 5.3.4.

5.4.2 Non-line-of-sight

Previous research [2], [11], [14], [28] has shown that mmWave radar has good material penetration capabilities. As demonstrated in Table 4, we found that there is less than

a 1 % difference in point-cloud density change when the radar sensor is covered by sheets (approximately $0.3cm \times 10cm^2$) made of different materials. However, unlike the mmWave radar-based system using a camera as the sub-sensor where the radar tracker can work independently [28], this fusion system cannot penetrate such sheets due to the requirement for line-of-sight from the vision-based proposal. On the other hand, our proposed system still outperforms vision-based detectors (cameras) in many scenarios in actual commercial environments during experiments. For example, when individuals are partially obscured by each other or furniture. Furthermore, this penetration capability has considerable potential, particularly in traditional Chinese and Japanese architecture, where many tables, chairs, doors, windows, and room dividers are made of wood and paper.

5.5 Privacy concerns

In this work, experiments were conducted in real commercial environments within a shopping district near Waseda University, Japan. Both customers and business owners expressed significant privacy concerns regarding the use of camera-involved systems. Such concerns primarily arise from technologies requiring a clear view of the user's face, as this is often perceived as intrusive. mmWave radar sensor-based systems appear more user-friendly and raise fewer privacy concerns due to the non-intrusive nature of mmWave radar and the spread of 5G. Notably, the proposed system has shown reduced privacy concerns and improved user acceptance compared to traditional optical methods in commercial settings, primarily due to the omission of facial features. This has been evidenced by the positive response from business owners who participated in our experiments. However, further research is required to comprehensively understand customer/user acceptance. Detailed comparisons and findings are presented in Table 1.

6. Related works

mmWave radar-based people detection. Many works have been on using mmWave radar for device-free people detection as an alternative to device-based technology like PIR sensors. Wei et al. documented a new passive tracking method (mTrack) that can pinpoint the target's initial position and track its trajectory with high precision, at a cost of limited in short range detection like touch events and writing [12]. Huang et al. proposed indoor people detection and tracking method based on DBmeans+RKF to improve the clustering performance on mmWave data and achieved 84 % accuracy, but over on only five insiders [13]. Zhao et al. built a people tracking and identifying method using Softmax modified network with Bi-LSTM layers analyzing people's gait. They provided an accuracy of 87 % for four insiders and 73 % for 12 insiders [2], at high deployment cost of users' information before they visit. Our previous work use a clustering method based on k-means and a assignment module based on Hungarian method presented an overall accuracy of 83.7 % for ten insiders but loss more than excepted in detecting stationary individuals [14]. A comparison of different mmWave radar detection methods could be found in Table 2.

Radar-camera fusion system people detection. Thanks to the great success of the supervised object detection methods like YOLO [5], Faster R-CNN [7] in recent years, many fusion approaches are proposed but mainly in the automotive market. The conventional radar-camera fusion approaches, like [15], are basically feed Kalman filter and its variants, and simplified radar detection to a point leading more potential failures. Recently, the fusion based on deep learning have undergone significant developments [19]–[26]. . Some of these cross-domain object detection approaches typically involve the use of end-to-end CNN architectures utilizing raw data captured by radar and cameras. This has led to a substantial demand for multi-modal datasets and labeled data. Shuai et al. described Millieye, a radar-camera fusion system featuring a replaceable CNN-based camera detector, but a need for enhancements in radar component. Bijilic et al. produced a multi-sensors fusion to see through the fog on the road [24] at high deployment and training cost. Chadwick et al. proposed a process for automatically labeling a new dataset by combining detection from multiple sensors, but limited to only simple image-like radar representation [25]. However, this category is mainly focused on automotive market and only limited research narrow such technologies to the indoor scenarios [23].

7. Conclusion

In this paper, we present a real-time, and robust people detection system that uses mmWave radar and camera fusion. Through multi-module cooperation, our system improves the performance of people detection using COTS. sensors. Notably, it performs well in challenging environments that the single-chip mmWave radar trackers and image-only vision-based detector may fail, including low-light conditions, stationary individuals, and when people are close to each other. Our evaluations illustrate that even including the above challenging scenarios, the proposed system achieves an mAP of 74.4 % and an overall people detection accuracy of 93.8 % with a median position error of 1.7 m in an actual commercial environment. As a comparison, the improved-tiny-YOLOv3 used in our previous work [28] achieves an mAP of 66.9 %, and the refinement head based on Faster R-CNN gives an mAP of 72.6 %. Besides, the single-chip mmWave radar systems, including TI [11], ctracker [13], and one of our previous work [14], have a people detection accuracy of 74.1 % (group size: 1-4), 84.7 % (group size: 1-5), 83.5 % (group size: 1-10), respectively; One current commercial overhead counting system gains an accuracy of 93.7% under suitable conditions. However, it cannot provide user information inside the smart space and fails to detect people in challenging scenarios. With our IoT platform based on AWS, our proposed system provides real-time services for commercial environments. It also demonstrates the poten-

tial of artificial intelligence IoT (AIoT) technologies to detect and recognize users automatically in smart city implementations. Despite its strengths, the proposed system has some limitations that warrant future study. In this paper, the point cloud from mmWave radar does not directly participate in object classification but provides proposals with coordinates. It leads to an improvement in accuracy at the cost of losing material penetration. Furthermore, at the current stage, the sparse point cloud generated by FFT suffers from low angular resolution, making the mmWave radar system unsuitable for classification tasks. In the near future, radar technology could provide a viable alternative to image-only methods for classification tasks as it provides more detailed information.

Acknowledgments

The acknowledgment belongs to Mr.Naotaka Saito and the Hyakunincho Oumiya, Shinjuku, Tokyo. This work was supported by JST SPRING, Grant Number JPMJSP2128.

References

[1] M.A. Ramírez-Moreno, S. Keshtkar, D.A. Padilla-Reyes, E. Ramos-López, M. García-Martínez, M.C. Hernández-Luna, A.E. Mogro, J. Mahlknecht, J.I. Huertas, R.E. Peimbert-García, *et al.*, "Sensors for sustainable smart cities: A review," Applied Sciences, vol.11, no.17, p.8198, 2021.

[2] P. Zhao, C.X. Lu, J. Wang, C. Chen, W. Wang, N. Trigoni, and A. Markham, "mid: Tracking and identifying people with millimeter wave radar," 2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS), pp.33–40, IEEE, 2019.

[3] S. Li and R. Hishiyama, "A field people counting test using millimeter wave radar in the restaurant," 2021 20th forum on information technology (FIT), pp.53–56, 2021.

[4] M. Kastek, H. Madura, and T. Sosnowski, "Passive infrared detector for security systems design, algorithm of people detection and field tests result," International Journal of Safety and Security Engineering, vol.3, no.1, pp.10–23, 2013.

[5] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[6] M. Simon, S. Milz, K. Amende, and H.M. Gross, "Complex-yolo: Real-time 3d object detection on point clouds," arXiv preprint arXiv:1803.06199, 2018.

[7] R. Girshick, "Fast r-cnn," Proceedings of the IEEE international conference on computer vision, pp.1440–1448, 2015.

[8] A.M. Ashleibta, A. Taha, M.A. Khan, W. Taylor, A. Tahir, A. Zoha, Q.H. Abbasi, and M.A. Imran, "5g-enabled contactless multi-user presence and activity detection for independent assisted living," Scientific Reports, vol.11, no.1, p.17590, 2021.

[9] A. Banerjee, K. Vaesen, A. Visweswaran, K. Khalaf, Q. Shi, S. Brebels, D. Guermandi, C.H. Tsai, J. Nguyen, A. Medra, *et al.*, "Millimeter-wave transceivers for wireless communication, radar, and sensing," 2019 IEEE Custom Integrated Circuits Conference (CICC), pp.1–11, IEEE, 2019.

[10] D.D. Ferris Jr and N.C. Currie, "Microwave and millimeter-wave systems for wall penetration," Targets and Backgrounds: Characterization and Representation IV, pp.269–279, SPIE, 1998.

[11] "People counting demonstration using ti mmwave sensors." Online, 2017. Accessed: Aug.1.2022.

[12] T. Wei and X. Zhang, "Mtrack: High-precision passive tracking using millimeter wave radios," Proceedings of the 21st Annual International Conference on Mobile Computing and Networking, MobiCom '15, New York, NY, USA, p.117–129, Association for Computing Machinery, 2015.

[13] X. Huang, H. Cheena, A. Thomas, and J.K. Tsoi, "Indoor detection and tracking of people using mmwave sensor," Journal of Sensors, vol.2021, 2021.

[14] S. Li and R. Hishiyama, "Counting and tracking people to avoid from crowded in a restaurant using mmwave radar," IEICE TRANSACTIONS on Information and Systems, vol.106, no.6, pp.1142–1154, 2023.

[15] H. Cho, Y.W. Seo, B.V. Kumar, and R.R. Rajkumar, "A multi-sensor fusion system for moving object detection and tracking in urban driving environments," 2014 IEEE International Conference on Robotics and Automation (ICRA), pp.1836–1843, IEEE, 2014.

[16] C. Yi, K. Zhang, and N. Peng, "A multi-sensor fusion and object tracking algorithm for self-driving vehicles," Proceedings of the Institution of Mechanical Engineers, Part D: Journal of automobile engineering, vol.233, no.9, pp.2293–2300, 2019.

[17] A. Motwani, S. Sharma, R. Sutton, and P. Culverhouse, "Interval kalman filtering in navigation system design for an uninhabited surface vehicle," The Journal of Navigation, vol.66, no.5, pp.639–652, 2013.

[18] J. Zhang, G. Welch, G. Bishop, and Z. Huang, "A two-stage kalman filter approach for robust and real-time power system state estimation," IEEE Transactions on Sustainable Energy, vol.5, no.2, pp.629–636, 2013.

[19] J. Wu, K. Hu, Y. Cheng, H. Zhu, X. Shao, and Y. Wang, "Data-driven remaining useful life prediction via multiple sensor signals and deep long short-term memory neural network," ISA transactions, vol.97, pp.241–250, 2020.

[20] V. Lekic and Z. Babic, "Automotive radar and camera fusion using generative adversarial networks," Computer Vision and Image Understanding, vol.184, pp.1–8, 2019.

[21] J. Kim, Y. Kim, and D. Kum, "Low-level sensor fusion network for 3d vehicle detection using radar range-azimuth heatmap and monocular image," Proceedings of the Asian Conference on Computer Vision, 2020.

[22] X. Shuai, Y. Shen, Y. Tang, S. Shi, L. Ji, and G. Xing, "millieye: A lightweight mmwave radar and camera fusion system for robust object detection," Proceedings of the International Conference on Internet-of-Things Design and Implementation, pp.145–157, 2021.

[23] D. Liu, X. Guan, Y. Du, and Q. Zhao, "Measuring indoor occupancy in intelligent buildings using the fusion of vision sensors," Measurement Science and Technology, vol.24, no.7, p.074023, 2013.

[24] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.11682–11692, 2020.

[25] S. Chadwick, W. Maddern, and P. Newman, "Distant vehicle detection using radar and vision," 2019 International Conference on Robotics and Automation (ICRA), pp.8311–8317, IEEE, 2019.

[26] R. Nabati and H. Qi, "Rrpn: Radar region proposal network for object detection in autonomous vehicles," 2019 IEEE International Conference on Image Processing (ICIP), pp.3093–3097, IEEE, 2019.

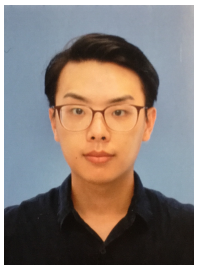
[27] R. Beringer, A. Sixsmith, M. Campo, J. Brown, and R. McCloskey, "The "acceptance" of ambient assisted living: Developing an alternate methodology to this limited research lens," International Conference on Smart Homes and Health Telematics, pp.161–167, Springer, 2011.

[28] S. Li and R. Hishiyama, "An indoor people counting and tracking system using mmwave sensor and sub-sensors," IFAC-PapersOnLine, vol.56, no.2, pp.7096–7101, 2023.

[29] C. Iovescu and S. Rao, "The fundamentals of millimeter wave sensors," Texas Instruments, pp.1–8, 2017.

[30] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, and A.C. Berg, "Ssd: Single shot multibox detector," Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pp.21–37,

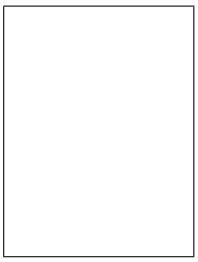
Springer, 2016.
[31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, vol.28, 2015.
[32] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," Advances in neural information processing systems, vol.29, 2016.
[33] Z. Yi, S. Yongliang, and Z. Jun, "An improved tiny-yolov3 pedestrian detection algorithm," Optik, vol.183, pp.17–23, 2019.
[34] J. MacQueen, "Classification and analysis of multivariate observations," 5th Berkeley Symp. Math. Statist. Probability, pp.281–297, 1967.
[35] R.E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," Journal of Basic Engineering, vol.82, no.1, pp.35–45, 03 1960.
[36] H.W. Kuhn, "The hungarian method for the assignment problem," Naval research logistics quarterly, vol.2, no.1-2, pp.83–97, 1955.
[37] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, "Microsoft coco: Common objects in context," Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pp.740–755, Springer, 2014.
[38] Y.P. Loh and C.S. Chan, "Getting to know low-light images with the exclusively dark dataset," Computer Vision and Image Understanding, vol.178, pp.30–42, 2019.



Shenglei Li is a Ph.D. student at Waseda University's Graduate School of Creative Science and Engineering, sponsored by JST Spring. He earned his Bachelor's degree in Civil Engineering from Southwest Jiaotong University and his Master's degree in System and Information Engineering from Tsukuba University. His research interests lie in Artificial Intelligence, Smart Cities, and automation.

Haoran Luo is a Ph.D. student at the Graduate School is currently a Ph.D. student at the Graduate School of Creative Science and Engineering, Waseda University. He is now sponsored by JST for research. He has a dual master's degree in Computer Science from Central China Normal University and University of Wollongong. His research interests include: sentiment analysis, semantic extraction, image style transfer and smart city construction, etc..

Tengfei Shao is a Ph.D. student at the Graduate School completed his master's degree in engineering at Waseda University. He is currently a Ph.D. student at the Department of Industrial and Management Systems Engineering, Waseda University, and sponsored by JST SPRING for research. His research interests include artificial intelligence, knowledge graphs, and complex networks.



Reiko Hishiyama is a professor at the Graduate School of Creative Science and Engineering of Waseda University, where she directs the Intelligent Information System Laboratory. She received her Doctor of Informatics degree in 2005 from Kyoto University, Japan. Her current research interests include artificial intelligence, autonomous multi-agent systems, knowledge representation, autonomy-oriented computing, and related areas.