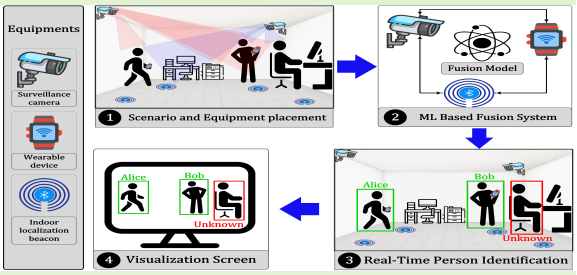
Enhancing Person Identification for Smart Cities: Fusion of Video Surveillance and Wearable Device Data based on Machine Learning

Jia-Ming Liang, *Member, IEEE*, Shashank Mishra, Chun-Che Wu

***Abstract*— For smart cities, video surveillance has been widely used for security and management purposes. In video surveillance, a fundamental challenge is person identification (PID), which involves promptly tagging individuals in videos with their IDs. Using RFID and fingerprint/iris/face recognition is a possible solution. However, the identification results are highly related to environmental factors, such as line of sight, lighting conditions, and distance. Fingerprint/face recognition also has privacy concerns. In this work, we show how to achieve immediate PID through two sensor data sources: (i) human objects and their pixel locations retrieved from videos and (ii) user trajectory data retrieved from wearable devices through indoor localization. By fusing these pixel trajectories and indoor trajectories, we demonstrate an enhancing-vision capability in the sense that PID can be achieved on surveillance videos even when no clear human biological features are seen. Two types of fusion are proposed: (i) similarity-based and (ii) machine learning-based. We have developed lightweight prototyping with off-the-shelf equipment and validated our results through extensive experiments. The performance evaluation showed that our system has an accuracy of up to 92% for person identification.**



***Index Terms*— Internet of Things (IoT), localization, machine learning, sensor fusion, video surveillance.**

I. Introduction

FOR smart cities, surveillance systems have been widely used in streets, factories, and public areas [1], [2]. Tracking particular persons in a surveillance video usually takes a lot of manpower if objects in the video are not properly tagged. We call this problem *person identification (PID)*. PID has long been achieved using RFID and fingerprint/iris/face recognition. However, RFID is limited by antenna coverage[1] and has little sense of source direction; fingerprint/iris recognition [4], [5] requires close contact with special devices; face recognition [6], [7] requires high-resolution facial images. To continuously identify a person in a surveillance region, the above solutions are not feasible. In addition, a lot of environmental factors, such as line of sight, view angle, lighting, and distance, need to be considered.

In this work, we assume that users are tagged by wearable devices, which are quite popular these days [8], [9]. Wearables are usually equipped with communication modules and lots of sensors, such as Bluetooth, accelerometer, and gyroscope. This allows us to capture user IDs, behaviors, and conduct indoor positioning at the same time. Based on this observation, we propose a fusion technique that combines sensors and video information to achieve PID for surveillance purposes. It integrates surveillance cameras and users' wearable devices, as shown in Fig. 1. An indoor localization system consisting of beacons is deployed in the surveillance region (e.g., lobby, living room, or interaction zone) with some cameras. The employees wear provided wearable devices, such as badges, for personal identification. Each wearable has a unique ID. Through image recognition, it is possible to capture human objects in surveillance videos and their pixel locations. On the other hand, the locations of wearables can also be tracked

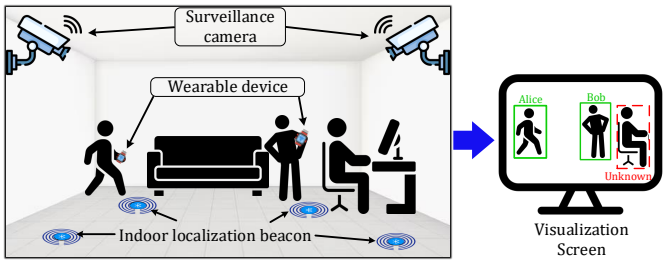


Fig. 1. Our PID scenario.

Jia-Ming Liang and Shashank Mishra are with National University of Tainan, Tainan 700301 Taiwan (e-mail: jmliang@mail.nutn.edu.tw, d10982003@stumail.nutn.edu.tw)

Chun-Che Wu is with Taiwan Semiconductor Manufacturing Company, Hsinchu City, 300093 Taiwan (e-mail: sampooh1@gmail.com). Corresponding author: Jia-Ming Liang.

[1]Generally, the coverage of passive RFID with common frequency is ranged from 10 cm to 1.5 m [3].

through the localization system. After some transformation, this location information may present some correlation. Hence, we can tag the IDs of wearable devices to human objects in videos, thus achieving PID.

In order to correlate video and wearable sensor data, we propose two types of fusion algorithms: (i) similarity-based and (ii) machine learning-based. For the first type, three schemes of information are proposed: Euclidean distance, non-linear distance, and non-linear distance with angular similarity. For the second type, the following solutions are explored: Support Vector Machine (SVM), Random Forest, and XGBoost. By quantifying correlations as similarity scores, we have developed a pairing mechanism to associate human objects with their respective IDs. Once paired, the IDs (PID) can be displayed on the screen. A practical system has been implemented, and comprehensive experiments have been conducted to evaluate the accuracy of this tagging process. When there are two people walking in the environment, the accuracy is up to 92%. With more complex scenarios and interleaving trajectories, the accuracy can still be maintained up to 82% $\sim$ 85%. As a result, we demonstrate an enhancing-vision capability in the sense that PID can be achieved on surveillance videos even when no clear human biological features are seen. Note that our scheme can co-work with other recognition solutions, such as face recognition, to improve accuracy, but this is out of the scope of this paper.

The rest of this paper is organized as follows. Section II reviews related works. Section III introduces our system model. Section IV describes our fusion and pairing algorithms. The evaluation results are presented in Section V. Section VI draws some conclusions.

II. Related Work

Wearable devices have become personalized and these devices can be considered as unique IDs of users. Several studies have focused on indoor positioning and tracking [10]–[16]. The works [10], [11] use inertial sensors to detect user paces. The studies [12]–[16] integrate wireless signal strength and a variety of sensors to conduct indoor positioning. The works [17]–[20] consider RFID signals integrated with a variety of sensors. Image information is studied in [21], [22]; they combine wireless signal strength and image information to improve positioning accuracy and conduct person identification. However, the signal strength may drift over time and objects in images obtained by background subtraction will suffer from the shadowing effect. The work [23] combines RFID and image information to instantly track mice and their IDs in a test box. It is also based on background subtraction, and its accuracy can be improved by increasing the density of the antenna, which incurs much cost. The work [24] proposes a WiFi RTT based indoor positioning system that relies on inertial sensor readings for walk motion tracking. However, WiFi signals would be more prone to drift than Bluetooth over time, which may accumulate and reduce location accuracy. The work [25] utilizes optical camera communication with LED arrays for multi-person monitoring. This approach is susceptible to limitations in range. The work [26] uses particle swarm optimization with signal propagation for mobile device positioning. While this scheme can improve accuracy, it can be computationally expensive, especially for real-time applications.

Information fusion has also been intensively studied recently. The study [27] determines the moving directions of two people through RFID and cameras. When the surveillance environment enlarges, more RFID antennas are needed. In [28], an RFID tag is attached to a target object and a camera is used to detect any moving object. The object can be recognized when it passes through a pre-defined RFID reader. The accuracy highly depends on antenna density, which also requires a higher cost. In the study [29], an object identification system is proposed leveraging surveillance camera data to identify objects. Work [30] presents a fusion approach designed for IoT devices and video objects, utilizing data captured from cameras. Reference [31] proposes a novel approach for video prediction based on data gathered from wearable inertial devices. Reference [32] proposes a fusion system consisting of an RFID reader and cameras on a robot platform to track people in the scene. A hybrid system with a Kinect camera and RFID for determining the standing area of a person is proposed in [33]. Reference [34] applies the Synthetic Aperture Radar technology to detect people. It then achieves PID by matching with Kinect skeleton information. The work [35] also achieves PID through Kinect skeleton and inertial information. Limited by Kinect, these works can only track people within the limited range (i.e., 4.5 meters in maximum) [36]. The work [37] proposes an indoor positioning system using visible light beacons and a device's camera for pose reconstruction. However, this method relies on clear visibility between the camera and beacons. The work [38] tackles inaccurate tracking data in multi-camera people tracking systems by using clustering. While this approach improves data quality, it might not be suitable for real-time applications with high processing demands. Additionally, clustering algorithm struggles with situations where people are very close together. Our goal is to break the above limitation. Table I summarizes the comparison of existing works.

III. System Model

Our system architecture is shown in Fig. 2. Each wearable device has inertial sensors and a Bluetooth interface[2]. Each wearable device runs a positioning algorithm, which collects nearby Bluetooth beacons' signals and combines inertial sensor data to compute its location. The trajectory information is then sent back to the fusion server. Each camera also continuously captures videos and transmits them, through RTSP streaming, to the fusion server for retrieving human bounding boxes. These bounding boxes also form trajectories. Our fusion server combines these data and finds their correlation. Finally, bounding boxes are tagged with user IDs, and the results can be visualized on a screen. We use YOLO[3] (You only

[2]The Bluetooth technology, also known for its lower cost [39] and extended range [40], is prevalent in smart wearables [41].

[3]YOLO is well-known for its speed, simplicity, and real-time effectiveness [42]. Its single-stage detection directly predicts bounding boxes and class probabilities, simplifying implementation [43]. Thus, it can balance between accuracy and speed makes it ideal for real-time applications [44].

TABLE I
COMPARISON OF EXISTING SYSTEMS.

Ref. & Year	Techniques/Methods	Equipment Cost / Computational Cost	Features					Data Types		
			SOD	MOD	PID	OMD	IL	Camera	Wearable Device	Beacon
[23] 2023	Computer Vision	High / High	✓	×	×	✓	✓	✓	✓	×
[27] 2021	Deep Learning	High / High	✓	×	×	✓	×	✓	✓	×
[32] 2021	Waving Action	Moderate / High	✓	×	×	×	✓	✓	×	×
[33] 2019	Deep Learning	High / High	✓	×	×	×	✓	×	✓	×
[34] 2021	Multi-Angle Synthetic Aperture	High / High	✓	×	×	×	✓	×	×	×
[35] 2018	Data Fusion	Moderate / Moderate	✓	✓	✓	×	✓	✓	✓	×
Proposed Scheme	**ML and Similarity Based Data Fusion**	Moderate / Moderate	✓	✓	✓	✓	✓	✓	✓	✓

SOD: Single object detection; **MOD:** Multi object detection; **PID:** Person identification; **OMD:** Object movement detection; **IL:** Indoor localization.
Note: Systems with 'High' equipment and computational costs, requiring significant hardware resources and complex processing capabilities. Systems with 'Moderate' costs generally use less resource-intensive methods, integrating multiple data sources without relying on high-end computational power.

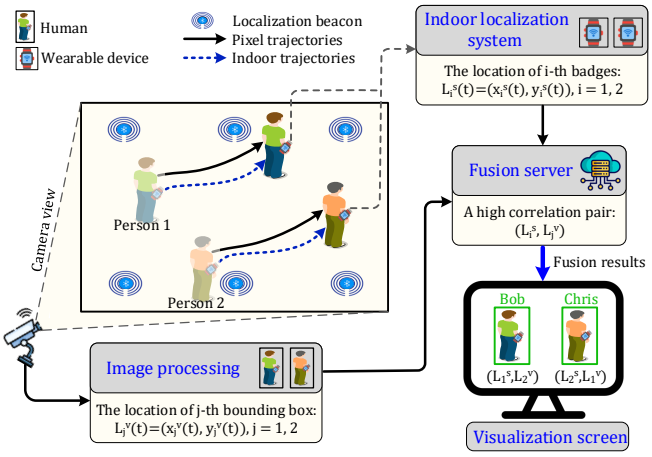


Fig. 2. System Architecture.

look once) [45] to retrieve human objects from each image[4]. Bounding boxes mark human objects. Each bounding box is first converted to a pixel coordinate. For example, in a top-view shot[5] as shown in Fig. 3, we can use the head position in a bounding box as a user's location. From continuous images, a user's walking trajectory can be combined with continuous bounding boxes. SORT (Simple, online, and real-time tracking of multiple objects in a video sequence) [46] is used here to connect bounding boxes.

To convert a pixel coordinate to the surveillance region, we take advantage of the fact that the positions of our beacons can also be seen in the videos, as shown in Fig. 3. We can use them as reference points for the conversion. At time point t, the location of the j-th bounding box (head position) captured by the camera is denoted as $L_j^v(t) = (x_j^v(t), y_j^v(t))$, $j = 1...n$.

There may be multiple cameras in the same surveillance region. The coordinates of the bounding boxes of an individual person captured by these cameras may not be consistent. We first group bounding boxes from cameras with a distance threshold. Bounding boxes classified as the same person are then averaged to get their physical coordinates.

[4]It can mitigate the shadowing effect caused by background subtraction.
[5]If the camera is not positioned directly above (i.e., if it is tilted), the pixel coordinates of the image can still be obtained through simple conversion.

TABLE II
NOTATIONS AND DESCRIPTION

Notations	Description
Θ_{low}	Lowest normalized distance
Θ_{high}	Highest normalized distance
$Dist_{ED}$	Euclidean distance of two locations
f_i	The i-th feature of trajectory
$L_j^v(t)$	Location of the j-th personnel bounding box at time t
$L_i^s(t)$	Location of the i-th wearable device at time t
$x_i^v(t)$	The x-th coordinate of location at time t of the j-th bounding box
$y_i^v(t)$	The y-th coordinate of location at time t of the j-th bounding box
k	The same coordinates for person and wearable device
$Sim_x(L_i^s, L_j^v)$	Similarity score between wearable device location L_i^s and person location L_j^v for any fusion algorithm x
$d(L_i^s(k), L_j^v(k))$	The distance similarity of wearable device location L_i^s and bounding box location L_j^v at coordinate k
$a(L_i^s(k), L_j^v(k))$	The angular similarity of wearable device location L_i^s and bounding box location L_j^v at coordinate k
$US_x(L_i^s)$	Uniqueness score of wearable device location L_i^s for any fusion algorithm x
$Var(\cdot)$	The variation of a set of values
W_s	Sliding window
D	Number of times
P	Set of outcome pairs of schemes
ED	Euclidean Distance
ND	Non-linear distance
NA	Non-linear distance with Angle
ML	Machine learning
$MOTA$	Multi-Object Tracking Accuracy

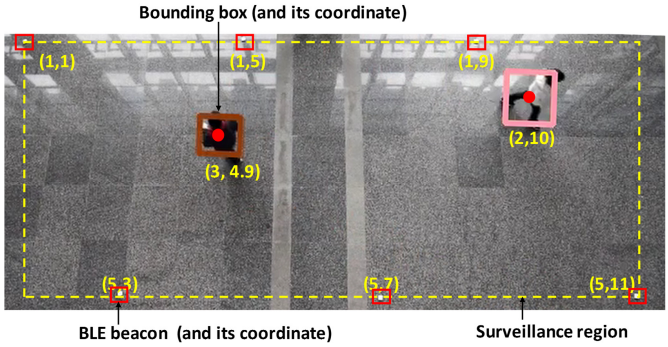
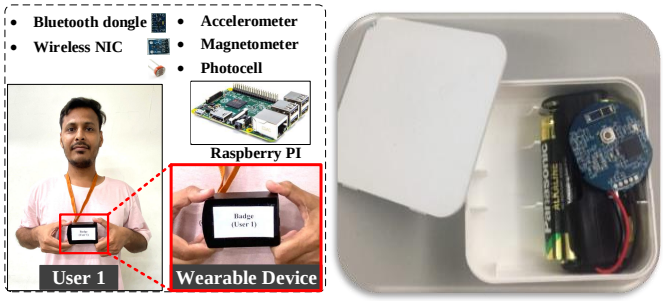


Fig. 3. Bounding boxes and coordinate conversion.

Wearable devices are carried by users. We design a wearable



(a) (b)

Fig. 4. (a) Wearable device and (b) BLE beacon.

device in a badge form, which has a Raspberry PI, inertial measurement components, and a Bluetooth module, as shown in Fig. 4(a). We apply the indoor positioning algorithm [47] to get badges' locations. BLE beacons, as shown in Fig. 4(b), are placed in the surveillance region. When a user moves, his walking patterns and displacements are tracked. Then, the positioning algorithm combines step information and beacon signal strengths by a particle filter algorithm to predict the user's trajectory (refer to [47] for details). At time point t, the location of the i-th badge is denoted as $L_i^s(t) = (x_i^s(t), y_i^s(t))$, $i = 1...m$.

Legal persons who walk around the environment will carry our badges. These badges contribute trajectories: $L_1^s, L_2^s, ..., L_m^s$. On the other hand, cameras also generate user trajectories: $L_1^v, L_2^v, ..., L_n^v$. Note that m may not be equal to n. The data fusion server calculates the correlation between these two sets. When a high correlation is found between L_i^s and L_j^v, a pair (L_i^s, L_j^v) is coupled, which means that the person corresponding to the bounding box of L_j^v is the person corresponding to the wearable device L_i^s. Then we can synchronously display the ID on the person's bounding box in the visualization screen. For a user not carrying a badge, no correlation may be found, which will be marked as 'unknown'. Table II summarized the notations used in this paper.

IV. Fusion and Pairing Algorithms

In this section, we derive several fusion algorithms. We show how to derive the similarity score $Sim_x(L_i^s(t - W_s : t), L_j^v(t - W_s : t))$ between $L_i^s(t - W_s : t)$ and $L_j^v(t - W_s : t)$, where x is any fusion algorithm and W_s is a sliding window. Whenever clear, time series $L_i^s(t - W_s : t)$ and $L_i^v(t - W_s : t)$ may be denoted as L_i^s and L_j^v, respectively. The sampling rate of wearable devices is 1 Hz [48], while the sampling rate of videos is 15 Hz [49]. In order to align these data, we down-sample the video data to 1 Hz whenever necessary[6]. Two types of fusion algorithms are presented, namely *similarity-based* and *machine learning-based*. Then, for all combinations of L_i^s and L_j^v, we present a pairing algorithm to determine their binding relations.

[6]Note that the synchronization can be solved by the existing work [50].

A. Similarity-based Algorithms

We present three algorithms based on Euclidean distance, non-linear distance, and non-linear distance with angular similarity. The goal is to compute score $Sim_x(L_i^s, L_j^v)$ for all i and j.

The first one is called *Euclidean Distance (ED)* scheme. We simply sum up the distance of all pairs of data points in the two trajectories. We use W_s to calculate the similarity score of L_i^s and L_j^v:

$$Sim_{ED}(L_i^s, L_j^v) = \frac{\sum_{k=(t-W_s)...t} \left\| L_i^s(k), L_j^v(k) \right\|}{W_s}, \quad (1)$$

where $\left\| L_i^s(k), L_j^v(k) \right\|$ is the Euclidean distance of two locations. Note that a larger value implies a lower similarity.

The second one is called *Non-linear Distance (ND)* scheme. In the ED scheme, there is a linear relation between distance and similarity. In the ND scheme, we normalize distance between 0 and 1 and set a distance range (Θ_{low}, Θ_{high}). When the distance is within the range, a reverse weight is given. When the distance is above Θ_{high}, a 0 weight is given. When the distance is below Θ_{low}, a weight of 1 is given. Here a larger value implies a higher similarity. This is defined as follows:

$$d(L_i^s(k), L_j^v(k)) = \begin{cases} 1 - \frac{Dist_{ED} - \Theta_{low}}{\Theta_{high} - \Theta_{low}} & , \text{if } \Theta_{low} \leq Dist_{ED} \leq \Theta_{high} \\ 1 & , \text{if } Dist_{ED} < \Theta_{low} \\ 0 & , \text{if } Dist_{ED} > \Theta_{high}, \end{cases} \quad (2)$$

where $Dist_{ED} = \left\| L_i^s(k), L_j^v(k) \right\|$. Then, we apply the same slide window again to calculate their similarity score:

$$Sim_{ND}(L_i^s, L_j^v) = \frac{\sum_{k=(t-W_s)...t} d(L_i^s(k), L_j^v(k))}{W_s}. \quad (3)$$

The third one is called *Non-linear distance with Angle (NA)* scheme. It considers the walking angles in a trajectory. For each pair $L_i^s(k)$ and $L_j^v(k)$, we can calculate their orientations and the angle between them from previous time point, denoted as $\angle(L_i^s(k), L_j^v(k))$. The angular similarity of them is defined as:

$$a(L_i^s(k), L_j^v(k)) = \frac{1 + Cos(\angle(L_i^s(k), L_j^v(k)))}{2}. \quad (4)$$

Then we integrate the previous distance similarity and this angular similarity by taking a multiplication and define the similarity score as:

$$Sim_{NA}(L_i^s, L_j^v) = \frac{\sum_{k=(t-W_s)...t} d(L_i^s(k), L_j^v(k)) \times a(L_i^s(k), L_j^v(k))}{W_s}. \quad (5)$$

B. Machine Learning-based Algorithms

Instead of using designed rules, here we try to use machine learning to learn important features for judging the similarity of L_i^s and L_j^v. Feature information extracted from videos and wearable devices are stored in a trajectory pool. Through machine learning models, the probability of two trajectories

belonging to the same user is computed. There are three phases: (a) feature extraction, (b) model training, and (c) similarity scoring.

For feature extraction, we design a variety of walking paths in a lobby and record the trajectories of bounding boxes and wearable devices. Because the recorded lengths are different and most machine learning algorithms take fixed length inputs, we have to cut data in the trajectory pool in the same length (in terms of the number of points). Then, we design three features. The first one is f_1 = "total length difference". Let $\|L_i^s\|$ and $\left\|L_j^v\right\|$ be the trajectory lengths of L_i^s and L_j^v in a window W_s, where the length of a trajectory is defined by connecting consecutive points by straight lines. We define $f_1 = |\,\|L_i^s\| - \left\|L_j^v\right\|\,|$. The second one is f_2 = "speed difference". The speed of a trajectory is defined as its length divided by its number of data points:

$$f_2 = |\frac{\|L_i^s\|}{t} - \frac{\left\|L_j^v\right\|}{t}|. \tag{6}$$

The third one is f_3 = "average distance difference". We calculate the distance between each pair of points and then average them:

$$f_3 = \frac{\sum_{k=t-W_s \ldots t} |L_i^s(k) - L_j^v(k)|}{W_s}. \tag{7}$$

The next phase is model training. We use the well-known classifiers such as SVM (Support Vector Machine) [51], Random Forest [52], and XGBoost [53] to train models from the above features[7]. Each pair is labeled as $+1$ or 0. When two trajectories are from the same person, the label is $+1$; otherwise, the label is 0. We randomly pick some sample pairs from the trajectory pool as our dataset and control the number of positive outcomes to be the same as the number of negative outcomes. SVM transforms the raw data into higher dimensions, constructs a hyperplane from the data, divides the data into two categories, and finally makes classifications. Random Forest integrates multiple decision trees by ensemble learning. Each tree in the last round will give its own category selection and vote accordingly. The output will be the category with the most votes. XGBoost is based on ensemble learning. Finally, voting options are used to select the category with the highest number of votes as the final result. The main difference is that Random Forest is an integrated algorithm that uses bagging to improve classification by combining randomly generated training sets, while XGBoost uses boosting to combine weak learning algorithms into strong learning algorithms. The selection of Boosting training sets is not independent. Each selected training set is sampled based on the error rate and depends on the result of the previous learning.

The third phase is similarity scoring. We use a window of size $W_s + D$ to score the similarity of two sequences. Let ML be one of SVM, Random Forest, and XGBoost. We define:

$$Sim_{ML}(L_i^s, L_j^v) = \frac{\sum_{k=1 \ldots D} ML\Big(L_i^s(\tau_k), L_j^v(\tau_k)\Big)}{D},$$

where $\tau_k = t - W_s - k : t - (k-1)$ and $ML(L_i^s, L_j^v)$ is the result of a model running on two sub-sequences. So the average number of 1s after running a model for D times is its score.

C. Pairing Algorithm

Based on the similarity scores of all combinations of L_j^v and L_i^s, we next design a strategy to determine the pairing result. Note that the sequence length for the similarity-based scheme is W_s, while that for the ML-based schemes is $W_s + D$. We need to extend the window size for the former to $W_s + D$ in order to make a uniform solution. Also, note that a user's trajectory from bounding boxes in videos may be broken due to reasons such as occultation and user leaving the camera view area. That is, some video trajectories may be shorter than $W_s + D$. In this case, we will not make any rush pairing decision until its trajectory is long enough. To avoid a sensor trajectory being similar to multiple video trajectories, we propose to use deviation to quantify the statistical dispersion of the combinations of each L_i^s and all possible candidates of L_j^v. Specifically, we define the Uniqueness Score of L_i^s as follows:

$$US_x(L_i^s) = Var\left(Sim_x(L_i^s, L_j^v) | \|L_j^v\| \geq W_s + D, \text{for all } j\right), \tag{8}$$

where x is any scheme and $Var(\cdot)$ is the variation of a set of values.

Intuitively, the higher the uniqueness score of a sequence L_i^s, the higher the priority that L_i^s will be paired first. Based on this concept, the L_i^s with the highest US will be paired with the most similar L_j^v depending on the value of $Sim_x(L_i^s, L_j^v)$. Note that all our schemes favor higher values, except the ED scheme. Then the one with the second highest US will be paired next. This is repeated until no more pairing is possible. The outcome is a set of pairs $P = \left\{(L_{i_1}^s, L_{j_1}^v), (L_{i_2}^s, L_{j_2}^v), \ldots\right\}$.

V. Performance Evaluation and Prototyping

For the experiments, we test our result in a lobby area with high ceiling, as shown in Fig. 3. We design a variety of walking paths, including eight designated trajectories and three random movements, as shown in Fig. 5. We deploy six BLE beacons for indoor localization with a single camera. Some users wear our badges, which have Raspberry PI, accelerometer, gyroscope, compass, WiFi, and Bluetooth modules.

Our fusion server has an Intel i7-6700 (3.4GHz) with GeForce GTX1080 GDDR5 (8GB) and 16GB DDR4 SDRAM. It has the following software components:

- Operating system: Ubuntu 16.04 64bit
- Eclipse version: Oxygen.2 (4.7.2)
- HTTP Web server version: Tomcat 7.0
- Nvidia driver version: 384.90
- CUDA version: 8.0
- cuDNN version: 5.1

[7]SVM, Random Forest, and XGBoost are well-established and highly efficient models that can effectively classify discrete data [54]. Their ensemble nature ensures accurate identification by adeptly capturing non-linear relationships and feature interactions, thus enhancing the overall performance [55], [56]

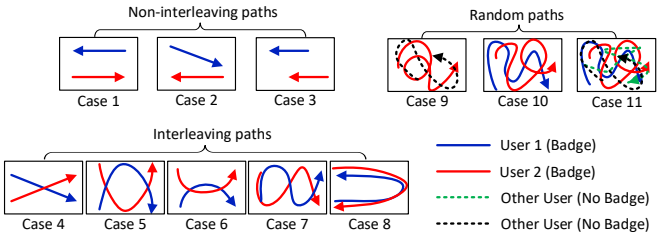


Fig. 5. Experimental walking paths.

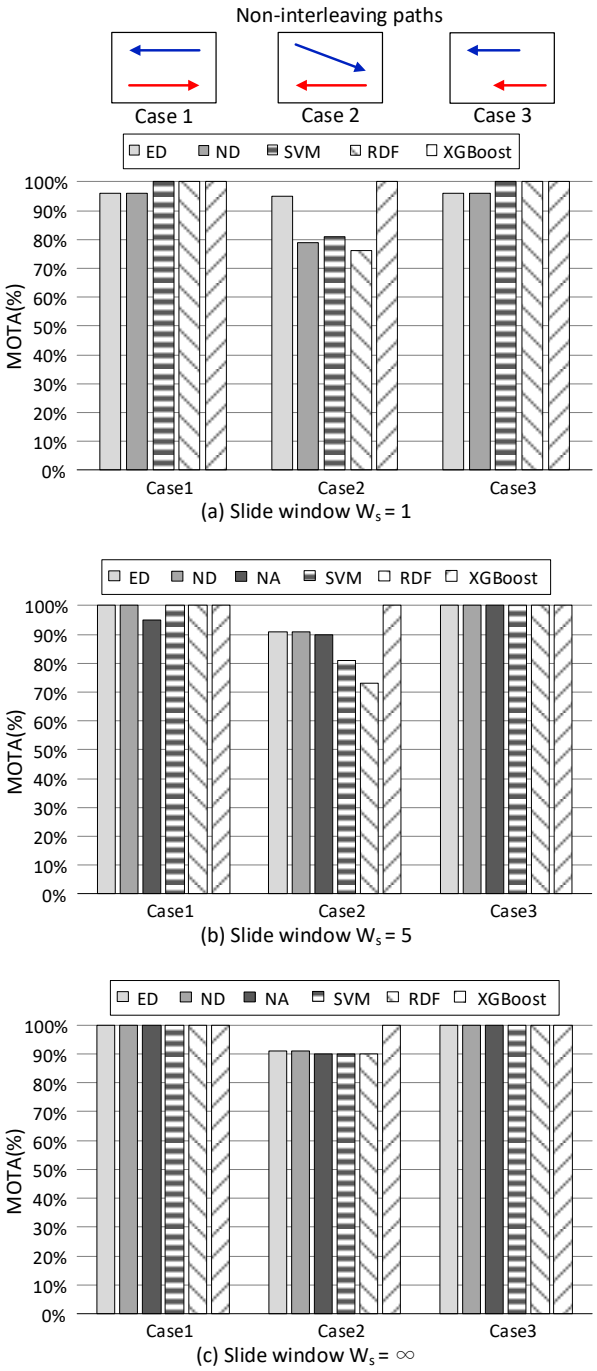


Fig. 6. Non-interleaving paths with various slide windows.

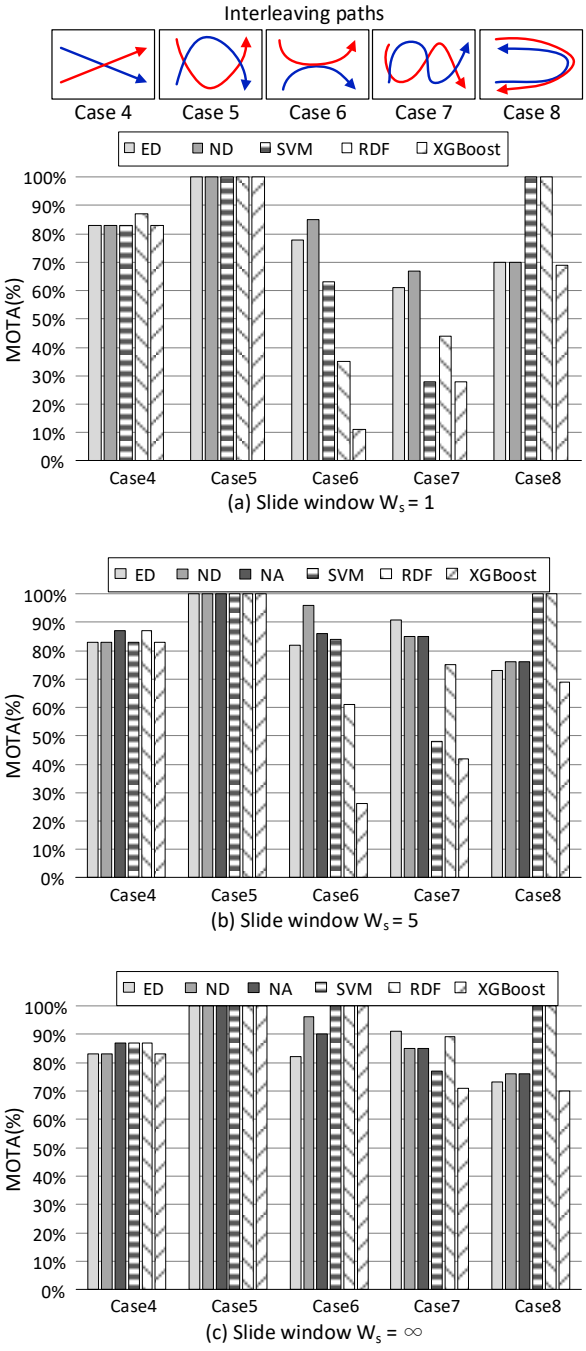


Fig. 7. Interleaving paths with various slide windows.

- opencv version: 3.1.1

Each beacon has a transmission range of 6 $\sim$ 15 m and a transmission interval of 100 ms. We use Compro IP570 PTZ camera which has a resolution of 1280 x 1024, rotation range of 340-degree, tilt range of 100-degree, and optical zoom ratio of 12x. Our Raspberry PI 3 Model B has a 1.2 GHz 64-bit quad-core ARM Cortex-A53 CPU with 1GB LPDDR2 RAM, MPU-6050 accelerometer, and GY-271 compass.

The experiments are conducted at the building in Micro-electronics and Information Research Center (MIRC). The environment size is 12 m $\times$ 5 m. There are 4 participants, where 1$\sim$3 users are with the wearable badge and 1$\sim$2 users

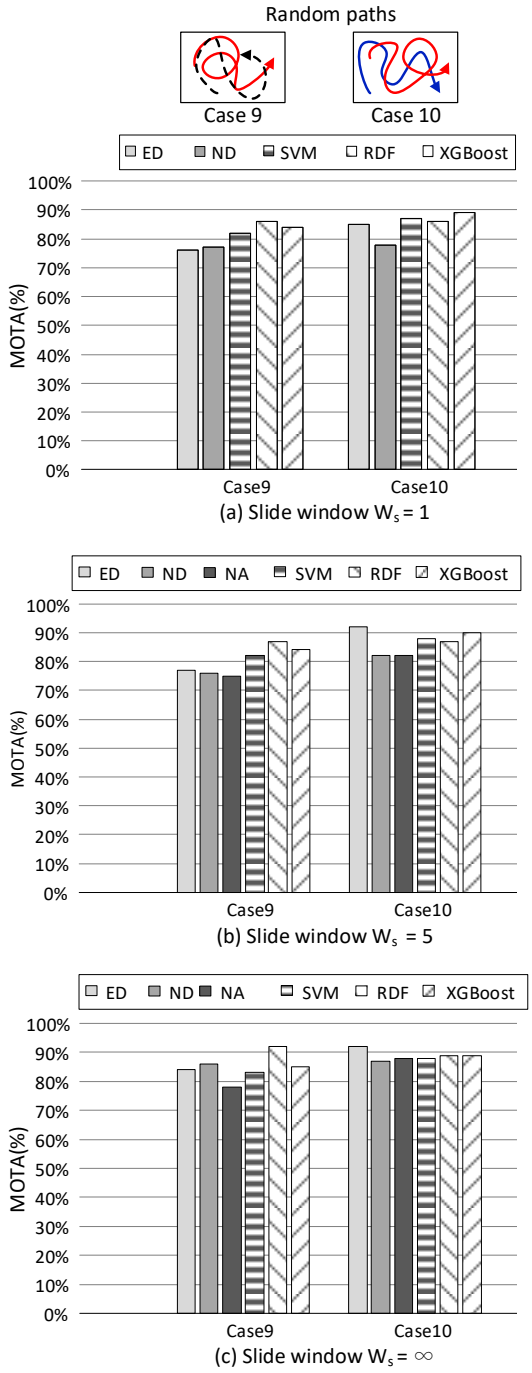


Fig. 8. Random paths with various slide windows.

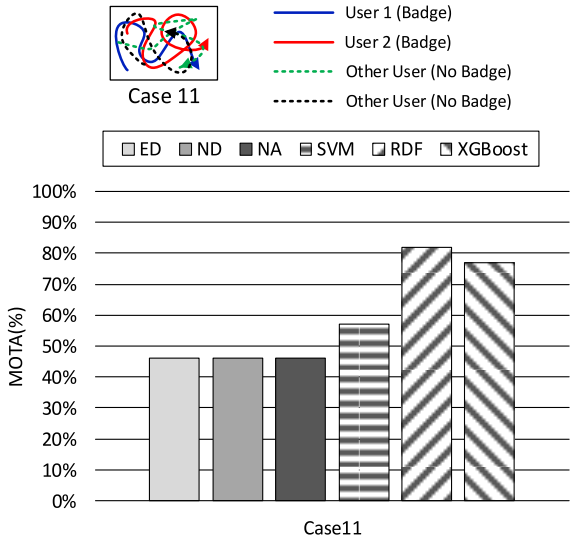


Fig. 9. Special Complex Scenario (Case 11) with slide window $W_s = \infty$.

do not carry a badge.

A. Evaluation Results

In order to examine the performance of the proposed schemes, we adopt Multi-Object Tracking Accuracy (MOTA) from [57], [58] to calculate the accuracy of identification in videos:

$$MOTA = \frac{\sum_{\forall t} correct\ identification\ in\ t}{\sum_{\forall t} all\ identification\ in\ t}. \quad (9)$$

In our designed walking paths, some may take about 20 seconds to complete, while some may take longer (such as case 9, 10, and 11, for which we take 3 minutes of data). Note that the results based on machine learning methods are evaluated with ten-fold cross-validation and leave-one-subject-out-validation methods. Fig. 6 shows the accuracy of our schemes by varying the sliding window with non-interleaving paths (case 1-3). We can see that most schemes reach 90% MOTA. When we extend the slide window, MOTA is improved, which is reasonable with the complex scenario of interleaving paths (case 4-8). Fig. 7 shows that the overall MOTA decreases. This is due to many broken video trajectories. This is difficult for pure image-based solutions, while our schemes can still achieve 85% MOTA on average, showing an enhancing-vision capability. Again, using a longer sliding window would help [8].

For random paths, case 9 simulates one authorized staff and one guest; case 10 simulates two authorized staffs; case 11 simulates a special complex scenario. The evaluation results are shown in Fig. 8. Note that case 9 has a user without a badge; the MOTA falls in 75% to 85% due to the presence of unknown users in the environment. In case 10, the MOTA is improved (80 $\sim$ 92%) due to the equal number of pairs. For the special complex scenario (Case 11), it involves four persons engaging in random walks within the lobby area: two staffs with badges and two guests without badges, as shown in Fig. 9. The experiment results show greater matching difficulty as the number of pairing trajectories increases. Specifically, most similarity-based methods show a decrease in performance, machine learning-based algorithms still achieve up to 82% MOTA. This is because the susceptibility of Bluetooth wireless signals to environmental effects, causing signal drift and interference [59]. Such instability in location trajectories is a longstanding challenge in wireless localization [60] and is difficult to overcome. Fortunately, as wireless localization technology advances, our algorithms can still be applied to new wearable localization technologies.

[8]Note that NA requires at least two data to calculate angular similarity. Thus, it is omitted when the sliding window=1.

B. Intrusion and Asset Protection Application

Here, we introduce several application scenarios for our system. These scenarios are: (i) two legal employees, (ii) one legal and one illegal employee, and (iii) asset tracking. Specifically, in the first scenario, two legal employees entered the scene. The fusion system swiftly identified their IDs individually and showcased them on the visualization screen. In the second scenario, it unfolded with the entry of a legal employee, followed by an authorized individual. As the legal employee's ID was swiftly identified and displayed on the visualization screen, the fusion system captured the presence of the illegal person and labeled them as 'Unknown' for identification purposes. Finally, the third scenario commenced with the placement of an asset bearing a badge within the room. As the asset moved, its location trajectory was tracked through BLE signals, and the image tracking system (YOLO) captured images of the person in the area, including the one carrying the asset. Thus, if an unauthorized person entered and absconded with the asset, this suspicious activity would be detected by the fusion system, flagging the person as carrying a 'Valuable Object' for further monitoring and tracking.

VI. Conclusion

We have introduced a surveillance system in smart cities that tackles the crucial problem of person identification, a significant research challenge across various fields. Unlike prior studies, our approach integrates trajectory data from both wearable devices and surveillance cameras. To correctly pair these trajectories data, we propose a fusion architecture that is based on the similarity scores obtained by different schemes. The result can greatly reduce manual efforts in finding out proper information in surveillance videos. For future directions, we will continue to explore applications in larger and more complex environments with a wider field of view, aiming to achieve broader camera coverage while maintaining efficiency with a minimal number of cameras. In addition, we will investigate the system's capabilities to identify a wider range of objects, including personnel and diverse assets, beyond the current scope to enhance its versatility. Finally, we plan to evaluate the proposed system's performance in real-world scenarios with higher personnel flow, such as in transportation hubs or city streets, to provide valuable insights into scalability and robustness. For the wearable components, we will continue exploring new wireless localization technologies, based on reasonable implementation costs and computational complexity, to enhance the positioning trajectories.

Acknowledgement

This research is co-sponsored by NSTC 109-2221-E-024-012, NSTC 109-2221-E-024-013 and NSTC 113-2221-E-024-011. Thanks to the students for their participation in the experiment and thanks to K.-R. Wu and Prof. Y.-C. Tseng for the technical consultations and valuable experience.

References

[1] Y. Pang, Z. Ni, and X. Zhong, "Federated Learning for Crowd Counting in Smart Surveillance Systems," *IEEE Internet of Things Journal*, vol. 11, no. 3, pp. 5200–5209, 2024.

[2] M. A. Rahman, M. S. Hossain, A. J. Showail, N. A. Alrajeh, and A. Ghoneim, "AI-Enabled IIoT for Live Smart City Event Monitoring," *IEEE Internet of Things Journal*, vol. 10, no. 4, pp. 2872–2880, 2023.

[3] "RFID Range Overview." [Online]. Available: https://skyrfid.com/RFID_Range.php

[4] Y. Duan, J. Feng, J. Lu, and J. Zhou, "Estimating Fingerprint Pose via Dense Voting," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2493–2507, 2023.

[5] P. Das, R. Plesh, V. Talreja, N. A. Schmid, M. Valenti, J. Skufca, and S. Schuckers, "Empirical Assessment of End-to-End Iris Recognition System Capacity," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 5, no. 2, pp. 154–169, 2023.

[6] H.-B. Kim, N. Choi, H.-J. Kwon, and H. Kim, "Surveillance System for Real-Time High-Precision Recognition of Criminal Faces From Wild Videos," *IEEE Access*, vol. 11, pp. 56 066–56 082, 2023.

[7] K. Yan, H. Shan, T. Sun, H. Hu, Y. Wu, L. Yu, Z. Zhang, and T. Q. S. Quek, "Reinforcement Learning-Based Mobile Edge Computing and Transmission Scheduling for Video Surveillance," *IEEE Transactions on Emerging Topics in Computing*, vol. 10, no. 2, pp. 1142–1156, 2022.

[8] F. John Dian, R. Vahidnia, and A. Rahmati, "Wearables and the Internet of Things (IoT), Applications, Opportunities, and Challenges: A Survey," *IEEE Access*, vol. 8, pp. 69 200–69 211, 2020.

[9] A. M. Rahmani, W. Szu-Han, K. Yu-Hsuan, and M. Haghparast, "The Internet of Things for Applications in Wearable Technology," *IEEE Access*, vol. 10, pp. 123 579–123 594, 2022.

[10] J. Kuang, D. Xia, T. Liu, Q. Chen, and X. Niu, "Shin-INS: A Shin-Mounted IMU-Based Inertial Navigation System for Pedestrian," *IEEE Sensors Journal*, vol. 23, no. 21, pp. 25 760–25 769, 2023.

[11] W. Pi, P. Yang, D. Duan, C. Chen, X. Cheng, L. Yang, and H. Li, "Malicious User Detection for Cooperative Mobility Tracking in Autonomous Driving," *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 4922–4936, 2020.

[12] A. Juárez, S. Fortes, E. Colin, C. Baena, E. Baena, and R. Barco, "UWB-based Positioning System for Indoor Sports," in *13th International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2023, pp. 1–6.

[13] X. Ma and S. Särkkä, "Indoor Positioning Methods Based on Dual Feet-Mounted IMUs With Distance Constraints," in *13th International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2023, pp. 1–6.

[14] J. Lu, C. Shan, K. Jin, X. Deng, S. Wang, Y. Wu, J. Li, and Y. Guo, "ONavi: Data-driven based Multi-sensor Fusion Positioning System in Indoor Environments," in *IEEE 12th International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2022, pp. 1–8.

[15] S. Jia, L. Ma, S. Yang, and D. Qin, "Semantic and Context Based Image Retrieval Method Using a Single Image Sensor for Visual Indoor Positioning," *IEEE Sensors Journal*, vol. 21, no. 16, pp. 18 020–18 032, 2021.

[16] I. Silva, C. Pendão, and A. Moreira, "Real-World Deployment of Low-Cost Indoor Positioning Systems for Industrial Applications," *IEEE Sensors Journal*, vol. 22, no. 6, pp. 5386–5397, 2022.

[17] T.-M. T. Dinh, N.-S. Duong, and Q.-T. Nguyen, "Developing a Novel Real-Time Indoor Positioning System Based on BLE Beacons and Smartphone Sensors," *IEEE Sensors Journal*, vol. 21, no. 20, pp. 23 055–23 068, 2021.

[18] A. Poulose, O. S. Eyobu, and D. S. Han, "An Indoor Position-Estimation Algorithm Using Smartphone IMU Sensor Data," *IEEE Access*, vol. 7, pp. 11 165–11 177, 2019.

[19] H. Xia, J. Zuo, S. Liu, and Y. Qiao, "Indoor Localization on Smartphones Using Built-In Sensors and Map Constraints," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 4, pp. 1189–1198, 2019.

[20] M. Merenda, L. Catarinucci, R. Colella, D. Iero, F. G. D. Corte, and R. Carotenuto, "RFID-Based Indoor Positioning Using Edge Machine Learning," *IEEE Journal of Radio Frequency Identification*, vol. 6, pp. 573–582, 2022.

[21] S. Jia, L. Ma, S. Yang, and D. Qin, "A Novel Visual Indoor Positioning Method With Efficient Image Deblurring," *IEEE Transactions on Mobile Computing*, vol. 22, no. 7, pp. 3757–3773, 2023.

[22] X. Zhang, J. Lin, Q. Li, T. Liu, and Z. Fang, "Continuous Indoor Visual Localization Using a Spatial Model and Constraint," *IEEE Access*, vol. 8, pp. 69 800–69 815, 2020.

[23] T. Fong, H. Hu, P. Gupta, B. Jury, and T. H. Murphy, "PyMouseTracks: Flexible Computer Vision and RFID-Based System for Multiple Mouse Tracking and Behavioral Assessment," *eneuro*, vol. 10, no. 5, 2023.

[24] R. Jurdi, H. Chen, Y. Zhu, B. Loong Ng, N. Dawar, C. Zhang, and J. K.-H. Han, "WhereArtThou: A WiFi-RTT-Based Indoor Positioning System," *IEEE Access*, vol. 12, pp. 41 084–41 101, 2024.

[25] H. Herfandi, O. S. Sitanggang, M. R. A. Nasution, H. Nguyen, and Y. M. Jang, "Real-Time Patient Indoor Health Monitoring and Location Tracking with Optical Camera Communications on the Internet of Medical Things," *Applied Sciences*, vol. 14, no. 3, 2024.

[26] Y. Assayag, H. Oliveira, E. Souto, R. Barreto, and R. Pazzi, "A Model-Based BLE Indoor Positioning System Using Particle Swarm Optimization," *IEEE Sensors Journal*, vol. 24, no. 5, pp. 6898–6908, 2024.

[27] X. Liu, D. Liu, J. Zhang, T. Gu, and K. Li, "RFID and camera fusion for recognition of human-object interactions," in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, 2021, pp. 296–308.

[28] C. Duan, W. Shi, F. Dang, and X. Ding, "Enabling RFID-Based Tracking for Multi-Objects with Visual Aids: A Calibration-Free Solution," in *IEEE INFOCOM - IEEE Conference on Computer Communications*, 2020, pp. 1281–1290.

[29] L.-Y. Zhang, H.-C. Lin, K.-R. Wu, Y.-B. Lin, and Y.-C. Tseng, "FusionTalk: An IoT-Based Reconfigurable Object Identification System," *IEEE Internet of Things Journal*, vol. 8, no. 9, pp. 7333–7345, 2021.

[30] K.-L. Tong, K.-R. Wu, and Y.-C. Tseng, "The Device–Object Pairing Problem: Matching IoT Devices with Video Objects in a Multi-Camera Environment," *Sensors*, vol. 21, no. 16, 2021.

[31] J.-Y. Li, J. C.-H. Lin, K.-R. Wu, and Y.-C. Tseng, "SensePred: Guiding Video Prediction by Wearable Sensors," *IEEE Internet of Things Journal*, vol. 10, no. 6, pp. 4698–4707, 2023.

[32] Y. Tian, S. Chen, J. Zhang, Z. Liu, X. Liu, and K. Li, "Localization of tagged objects on shelf via a portable camera-augmented rfid reader," in *International Conference on Computer Communications and Networks (ICCCN)*. IEEE, 2021, pp. 1–9.

[33] C. N. Phyo, T. T. Zin, and P. Tin, "Deep Learning for Recognizing Human Activities Using Motions of Skeletal Joints," *IEEE Transactions on Consumer Electronics*, vol. 65, no. 2, pp. 243–252, 2019.

[34] S. Gui, Y. Yang, J. Li, F. Zuo, and Y. Pi, "THz Radar Security Screening Method for Walking Human Torso With Multi-Angle Synthetic Aperture," *IEEE Sensors Journal*, vol. 21, no. 16, pp. 17 962–17 972, 2021.

[35] W.-C. Chang, C.-W. Wu, R. Y.-C. Tsai, K. C.-J. Lin, and Y.-C. Tseng, "Eye on You: Fusing Gesture Data from Depth Camera and Inertial Sensors for Person Identification," *in Proc. IEEE Int'l Conference on Robotics and Automation (IRCA)*, pp. 2021–2026, 2018.

[36] Z. Cai, J. Han, L. Liu, and L. Shao, "RGB-D Datasets Using Microsoft Kinect or Similar Sensors: A Survey," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 4313–4355, 2017.

[37] J. D. Gutiérrez, T. Aguilera, F. J. Álvarez, J. Morera, and F. J. Aranda, "Precise Local Positioning of a Mobile Device Based on Pose Reconstruction From a Visible Light Beacon," *IEEE Access*, vol. 12, pp. 20 517–20 529, 2024.

[38] J. Kim, W. Shin, H. Park, and D. Choi, "Cluster Self-Refinement for Enhanced Online Multi-Camera People Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024, pp. 7190–7197.

[39] L. Bai, F. Ciravegna, R. Bond, and M. Mulvenna, "A low cost indoor positioning system using bluetooth low energy," *IEEE Access*, vol. 8, pp. 136 858–136 871, 2020.

[40] P. Zand, J. Romme, J. Govers, F. Pasveer, and G. Dolmans, "A high-accuracy phase-based ranging solution with Bluetooth Low Energy (BLE)," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2019, pp. 1–8.

[41] S. Akiyama, R. Morimoto, and Y. Taniguchi, "A study on device identification from ble advertising packets with randomized mac addresses," in *IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, 2021, pp. 1–4.

[42] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "YOLO-World: Real-Time Open-Vocabulary Object Detection," *arXiv preprint arXiv:2401.17270*, 2024.

[43] K. Tong and Y. Wu, "I-YOLO: a novel single-stage framework for small object detection," *The Visual Computer*, pp. 1–18, 2024.

[44] Y. Zhou, "A yolo-nl object detector for real-time detection," *Expert Systems with Applications*, vol. 238, 2024.

[45] G. Li, Z. Ji, X. Qu, R. Zhou, and D. Cao, "Cross-Domain Object Detection for Autonomous Driving: A Stepwise Domain Adaptative YOLO Approach," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 3, pp. 603–615, 2022.

[46] X. Liu, L. Lin, S. Yan, H. Jin, and W. Jiang, "Adaptive Object Tracking by Learning Hybrid Template Online," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 11, pp. 1588–1599, 2011.

[47] C.-C. Lo, T.-H. Chiang, T.-K. Lee, L.-J. Chen, and Y.-C. Tseng, "Wireless Location Tracking by a Sensor-assisted Particle Filter and Floor Plans in a 2.5-D Space," *in Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, 2018.

[48] H. Tan, A. M. Wilson, and J. Lowe, "Measurement of stride parameters using a wearable GPS and inertial measurement unit," *Journal of biomechanics*, vol. 41, no. 7, pp. 1398–1406, 2008.

[49] N. Wagle, J. Morkos, J. Liu, H. Reith, J. Greenstein, K. Gong, I. Gangan, D. Pakhomov, S. Hira, O. V. Komogortsev *et al.*, "aEYE: a deep learning system for video nystagmus detection," *Frontiers in Neurology*, vol. 13, p. 963968, 2022.

[50] H. Zhang, J. Guo, X. Liu, D. Zhou, and Y. Hou, "A Time Synchronization Algorithm Based on Correlation Analysis in GNSS/INS Integrated Navigation," in *IEEE International Conference on Unmanned Systems (ICUS)*, 2022, pp. 1327–1332.

[51] M. Tanveer, M. Tabish, and J. Jangir, "Pinball Twin Bounded Support Vector Clustering," in *IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2021, pp. 1–4.

[52] C. Galen and R. Steele, "Performance Maintenance Over Time of Random Forest-based Malware Detection Models," in *11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2020, pp. 0536–0541.

[53] G. Minghui, Z. Hang, M. Li, Z. Zhijun, L. Kai, C. Xudong, H. Jicheng, and N. Zhiyan, "Research on Intrusion Detection Model Based on DAE- XGBoost," in *IEEE 10th International Conference on Information, Communication and Networks (ICICN)*, 2022, pp. 57–62.

[54] S. Demir and E. K. Şahin, "Liquefaction prediction with robust machine learning algorithms (SVM, RF, and XGBoost) supported by genetic algorithm-based feature selection and parameter optimization from the perspective of data processing," *Environmental Earth Sciences*, vol. 81, no. 18, 2022.

[55] S. Fatima, A. Hussain, S. B. Amir, S. H. Ahmed, and S. M. H. Aslam, "XGBoost and Random Forest Algorithms: An in Depth Analysis," *Pakistan Journal of Scientific Research*, vol. 3, no. 1, pp. 26–31, 2023.

[56] M. Mallik, A. K. Panja, and C. Chowdhury, "Paving the way with machine learning for seamless indoor–outdoor positioning: A survey," *Information Fusion*, vol. 94, pp. 126–151, 2023.

[57] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *European conference on computer vision*. Springer, 2020, pp. 107–122.

[58] A. Kim, A. Ošep, and L. Leal-Taixé, "EagerMOT: 3D Multi-Object Tracking via Sensor Fusion," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 11 315–11 321.

[59] M. O. Al Kalaa, W. Balid, N. Bitar, and H. H. Refai, "Evaluating bluetooth low energy in realistic wireless environments," in *IEEE Wireless Communications and Networking Conference*, 2016, pp. 1–6.

[60] Y. Yang, M. Chen, Y. Blankenship, J. Lee, Z. Ghassemlooy, J. Cheng, and S. Mao, "Positioning Using Wireless Networks: Applications, Recent Progress and Future Challenges," *arXiv preprint arXiv:2403.11417*, 2024.