



Things in the air: tagging wearable IoT information on drone videos

Lan-Da Van¹ · Ling-Yan Zhang³ · Chun-Hao Chang¹ · Kit-Lun Tong¹ · Kun-Ru Wu¹ · Yu-Chee Tseng²

Received: 20 November 2020 / Accepted: 21 January 2021

© The Author(s) 2021  

Abstract

Drones have been applied to a wide range of security and surveillance applications recently. With drones, Internet of Things are extending to 3D space. An interesting question is: Can we conduct person identification (PID) in a drone view? Traditional PID technologies such as RFID and fingerprint/iris/face recognition have their limitations or require close contact to specific devices. Hence, these traditional technologies can not be easily deployed to drones due to dynamic change of view angle and height. In this work, we demonstrate how to retrieve IoT data from users' wearables and correctly tag them on the human objects captured by a drone camera to identify and track ground human objects. First, we retrieve human objects from videos and conduct coordination transformation to handle the change of drone positions. Second, a fusion algorithm is applied to measure the correlation of video data and inertial data based on the extracted human motion features. Finally, we can couple human objects with their wearable IoT devices, achieving our goal of tagging wearable device data (such as personal profiles) on human objects in a drone view. Our experimental evaluation shows a recognition rate of 99.5% for varying walking paths, and 98.6% when the drone's camera angle is within 37°. To the best of our knowledge, this is the first work integrating videos from drone cameras and IoT data from inertial sensors.

Keywords Computer vision · Data fusion · Drone · IoT · Person identification

1 Introduction

Recently, drones (or Unmanned Aerial Vehicles, UAVs) have been applied to a lot of fields, such as smart agriculture [2], military [3], construction site monitoring [4], and environmental monitoring [5]. On the other hand, IoT has been proved powerful in location tracking [6], physiology monitoring [7], and behavior analysis [8]. This work investigates what future surveillance system would look like. We observe that both cameras and IoT devices have their particular capabilities in identifying and tracking moving objects and their behaviors. Their correlations are, however, unclear. We intend to tackle the challenge of integrating visual data and IoT data for person identification (PID), where human objects captured by a drone camera can be correctly tagged by their wearable IoT information (such as personal profiles).

PID can be achieved by technologies such as RFID and fingerprint/iris/face recognitions. Face recognition solutions [9, 10] are sensitive to lighting condition, occlusion, and viewing direction. Moreover, face recognition requires large labeled datasets to train classifiers, limiting its scalability. Fingerprint [11] and iris recognition [12] need users' biology

A preliminary version of this paper was presented in ACM MobiCom 2019 as a demo [1].

✉ Ling-Yan Zhang, lingyan_zhang@csu.edu.cn; Lan-Da Van, ldvan@cs.nctu.edu.tw; Yu-Chee Tseng, yctsen@cs.nctu.edu.tw | ¹Department of Computer Science, National Chiao Tung University, Hsinchu 300, Taiwan. ²Department of Computer Science, College of Artificial Intelligence, National Chiao Tung University, College of Health Sciences Kaohsiung Medical University, Hsinchu 300, Taiwan. ³Department of Computer Science and Engineering, Central South University, Changsha, Hunan 410083, People's Republic of China.



features, and require close contact to specific devices. Hence, these solutions are not applicable to drones with changing height and view angle.

In this work, we present an approach to identifying and tracking human objects in the videos taken from a drone and correctly tagging their personal profiles retrieved, through wireless communications, from their wearable IoT devices. Figure 1 shows what our system can achieve by integrating a drone camera and wearable devices. To the best of our knowledge, this is the first work correlating IoT data and computer vision from a drone camera. Through combining IoT and computer vision, the future aerial surveillance systems could be even smarter and more user-friendly. For example, in Fig. 1, the visualized information can even include users' personal profiles and past activities before they actually entered the camera view. Potential applications also include military training and construction site monitoring where people are required to wear badges. Whenever needed, the fusion server can also send alert messages to specific persons (for reasons such as the persons enter a dangerous region found by the drone).

We propose a data fusion approach that combines videos with inertial sensor data. Our system has an aerial camera, some wearable devices, and some ground markers. A drone carries a camera to continuously record videos. A crowd of people are in the drone view and some of them may put on their wearable devices, each of which is an IoT device with some inertial sensors. This allows us to collect user profiles and motion data. Both video data and IoT data are transmitted to a fusion server for correlation analysis. The correlation procedure consists of four steps. First, we retrieve human objects from videos by using a deep learning network. Second, we transform these human objects from a pixel space to a ground coordinate to handle the change of drone positions. We use some ground ArUco markers [13] for the transformation. Third, we extract human motion features from both video data and inertial sensor data. Finally, a fusion algorithm is applied to measure the similarity of each pair of normalized motion feature series, one from videos and the other from wearable devices. By quantifying all-pair similarity scores, we are able to couple human objects with their IoT devices, achieving our goal of tagging wearable IoT data on drone-recorded videos.

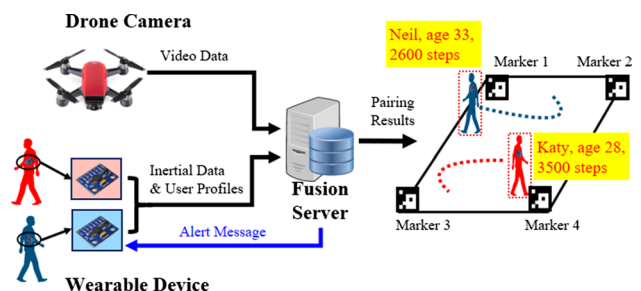
To validate our idea, we conduct a number of experiments by changing peoples' walking paths, the number of people, and the drone's view angle. Experimental results show over 99.5% in accuracy for two people walking in a variety of walking paths, including circular, non-interleaving, interleaving, and random. The accuracy is 98.6% when the drone's camera angle is within 37° , showing that our system can handle the change of drone positions during PID. Our work makes the following contributions:

- The proposed surveillance system works for drones by firstly and novelly integrating with wearable IoT devices.
- Our fusion approach correlates two independent sources, video data and inertial data, in an interesting way, thus achieving PID and tracking with less effort and even without seeing human biological features.
- Our system tackles the changing height and view angle issues of drones, and can work without any data labeling process. However, if certain training effort (such as face labeling) is available, our system can leverage it and further improve its accuracy.

2 Related work

The PID issue has been extensively studied in the computer vision and IoT fields. It plays a critical role in many security [14], surveillance [15], and business-intelligence [16] applications. The work [14] presents a smart homecare system by utilizing PID and behavior identification methods. The work [15] improves the PID performance in surveillance system by combining deep learning with multiple metric ensembles. Reference [16] proposes a PID method for human-robot interaction by using a simplified fast region-based convolutional network. In this section, we briefly review the state-of-the-art

Fig. 1 Person identification and tracking scenario



vision-based PID approaches, biometric-based PID approaches, multi-sensory fusion solutions, and drone-based PID solutions.

In the area of computer vision, PID has been widely studied in pedestrian tracking [17–19], and face recognition [20, 21] recently. The work [17] realizes PID by video ranking using discriminative space-time and appearance feature selection. In [18], the detected pedestrians across a whole video are linked by fusing human pose information. Reference [19] addresses the problem of pedestrian misalignment in a video sequence by combining appearance features and pose features. How to conduct face recognition using a deep convolutional neural networks (CNN) is discussed in [22]. Din and Ta [20] proposed a comprehensive framework based on CNN for video-based face recognition. Reference [21] addresses low-resolution face recognition in the wild via selective knowledge distillation. However, vision-based PID solutions are sensitive to lighting condition, occlusion, and viewing direction. The work [23] investigates the face recognition performance on drones, including Face++ [24], Rekognition [25], and OpenCV [26]. It points out three obstacles for face recognition: (i) the small-sized facial images taken by drones from long distances, (ii) the pose variances introduced by large angle depression, and (iii) the large labeled datasets required for model training.

Biometric-based PID approaches identify human objects by measuring their unique physical or behavioral characteristics, mainly including hand geometry, fingerprint, and iris. Reference [11] studies curvature features for 3D fingerprint recognition, while [27] proposes a finger-vein-based biometric identification system under different image quality conditions. Iris recognition based on human-interpretable features is proposed in [12]. Information fusion by combining multiple sources (e.g., face plus fingerprint, and face plus iris) is studied in [28]. However, fingerprint and iris recognitions need users' biology features, and require close contact to specific devices. Hence, biometric-based solutions are more suitable for indoor applications such as smart home [6], physiology monitoring [7], and behavior analysis [8]. Applying them to drones would be difficult.

Multi-sensory fusion solutions identify and track human objects by fusing data from multiple cameras, inertial measurement units (IMUs), and IoT devices. In general, when partially independent information is available from multiple sources, their combination may improve accuracy over a single-source solution. Reference [29] uses the features from static cameras and IMU sensors to identify and track multiple persons. References [30, 31] propose a deep learning multi-modal framework to match silhouette video clips and accelerometer signals to identify subjects. A through-wall PID system based on gait features from WiFi signal and video-based simulations is proposed in [32]. However, these solutions require static cameras with their lens paralleling to the subjects, making it difficult to apply to highly dynamic drone views and angles.

Drone-based PID solutions aim to identify and track human objects from drone videos. Reference [33] studies multi-view human motion capturing with an autonomous swarm of drones. The work [34] presents a large-scale visual object detection and tracking benchmark, named VisDrone2018. The work [35] develops fully CNNs for human crowd detection for drone-captured images. However, detecting human objects and tracking them do not imply getting their IDs. Reference [36] presents an introductory study pushing PID on mobile platforms, such as drone, and introduces some differences and challenges compared to standard fixed cameras. One major challenge is drone's continually varying view angle and position. Reference [37] uses thermal cameras on UAVs for human tracking. Reference [38] proposes a visual representation for drones to detect persons. The study [39] combines color feature descriptors and frontal face perception for UAVs to track walking persons. Reference [40] proposes multiple person detection and tracking method using CNN and Hungarian algorithm in drone images. However, these solutions may suffer from accuracy due to changing environment factors and target appearances.

3 System model

We see a new opportunity from the widely accepted IoT devices, such as smart watches and smart phones, which have become virtually the unique identity of a person. Most wearable devices are equipped with some inertial sensors and, more importantly, have access to the owners' personal profiles that may reveal (under the owners' consent) a lot of critical information (such as identity, age, sex, citizenship, social networks, past activities, purchase records, etc.). Our goal is to study PID in an aerial platform, such as a drone, by integrating an aerial camera and wearable IoT devices. Identifying a specified person from a down view and knowing his personal information become not so difficult with the help of wearable devices. We can visualize not only people's IDs, but also their personal profiles, even under a lot of occlusions

Fig. 2 A smart surveillance scenario in a square

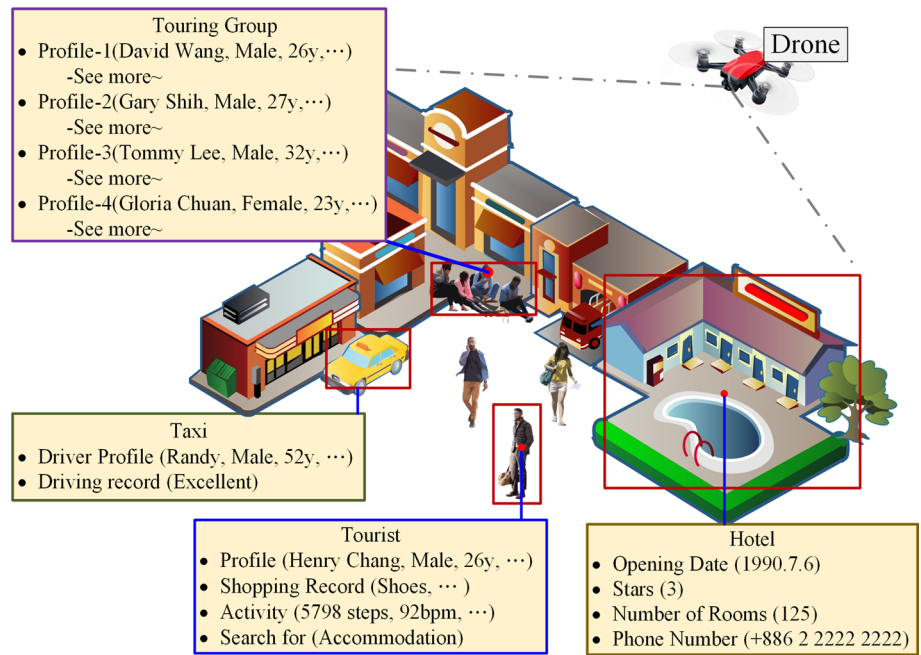


Fig. 3 Data fusion model

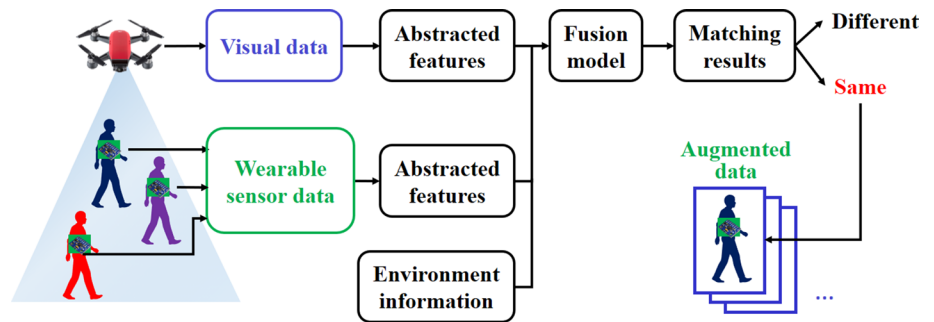
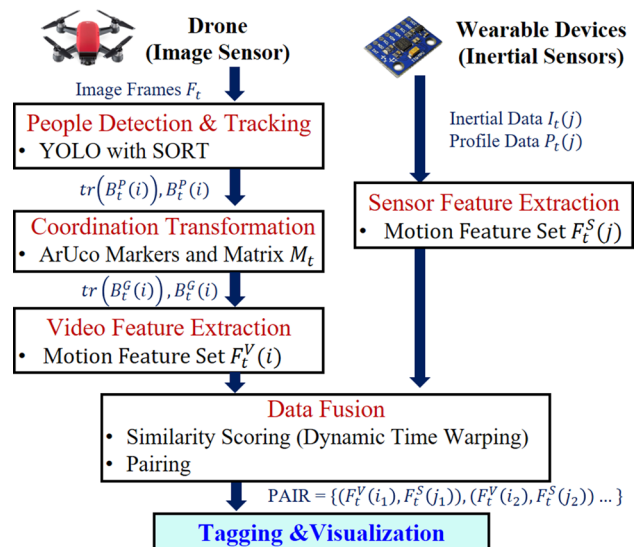


Fig. 4 Data processing flow



or situations where they dress up similarly. Figure 2 shows an imaginary smart surveillance scenario in a square, which may shed some light on how our system works. Figure 3 shows the general concept of our data fusion model, where both camera and wearable devices try to capture common features for fusion purpose. Figure 4 shows our data processing flow, which consists of six main modules.

3.1 People detection and tracking

The drone camera provides video streams. The image frame at time index t is denoted as F_t . We then use You Only Look Once (YOLO) [41] to retrieve all objects that are recognized as human. The bounding box of the i -th human object in frame F_t is denoted as $B_t^P(i)$, where P means pixel space. From continuous images, a human object's walking trajectory is derived from continuous bounding boxes by Simple Online and Realtime Tracking (SORT) [42]. We denote by $tr(B_t^P(i))$ the trajectory of $B_t^P(i)$ that happened before F_t . If $B_t^P(i)$ is a newly appeared object, then $tr(B_t^P(i)) = \emptyset$. So the trajectory up to time t is $tr(B_t^P(i)) + B_t^P(i)$ (+ means concatenation). Note that since we only concern about a person's location, a bounding box is only represented by its center (in pixel space).

3.2 Coordination transformation

Due to mobility, the drone's view angle and location may keep changing, making camera coordinate change as well. Therefore, the above trajectory $tr(B_t^P(i)) + B_t^P(i)$ may be inconsistent. To maintain a fixed coordinate, we place four ArUco markers [12] with a visible size on the ground. An ArUco marker is a synthetic square composed by a wide black border and an inner binary matrix which determines its identifier. The black border facilitates fast detection and the binary codification enables quick identification with a certain error detection and correction capability. These markers allow us to transform a pixel location to a global ground space with respect to these markers' locations. A transformation matrix M_t for time t will be calculated. The ground coordinate is written as $M_t(B_t^P(i)) = B_t^G(i)$, and the trajectory of $B_t^G(i)$ before F_t after such a transformation is written as $tr(B_t^G(i))$. That is, the trajectory up to time t in the ground space is $tr(B_t^G(i)) + B_t^G(i)$. Note that the ground markers and their arrangements are only to demonstrate our idea, they may be replaced by other techniques and even marker-less schemes.

3.3 Video feature extraction

From the ground trajectory $tr(B_t^G(i)) + B_t^G(i)$, we extract its motion features such as rotation, direction, behavior, and movement. We denote by $F_t^V(i) = \{F_t^V(i).f_1, F_t^V(i).f_2, \dots, F_t^V(i).f_r\}$ the set of motion features retrieved from the trajectory, where r is the number of features and V means "video".

3.4 Sensor feature extraction

For fusion purpose, we need to retrieve the same features from wearable sensor data as those from videos. For IoT device j , its inertial data and profile data at time t are denoted as $I_t(j)$ and $P_t(j)$, respectively. Since $P_t(j)$ is for visual augmentation only, we will focus on $I_t(j)$ in the following. Similarly, the time series data of sensor j is denoted by $tr(I_t(j)) + I_t(j)$. From the series, we extract the similar set of motion features $F_t^S(j) = \{F_t^S(j).f_1, F_t^S(j).f_2, \dots, F_t^S(j).f_r\}$, where S means "sensor".

3.5 Data fusion

The correlation of video and sensor data is calculated by a fusion model, which consists of two tasks: similarity Scoring and Paring. The Similarity Scoring sub-module calculates a similarity matrix indexed by each pair of $F_t^V(i)$ and $F_t^S(j)$ through dynamic time warping (DTW). The comparisons are based on the retrieved r motion features. Here the problems of clock asynchrony and inconsistent sampling for distributed sensors are handled by DTW. Then the Paring sub-module sorts these similarity scores, matches human objects with their wearable devices, and finally outputs its pairing result $PAIR = \{(F_t^V(i_1), F_t^S(j_1)), (F_t^V(i_2), F_t^S(j_2)), \dots\}$.

3.6 Tagging and visualization

Based on the pairing output *PAIR*, the final step is to integrate the result with a target application. We use tagging IoT information on human objects in a drone view as an example. This allows us to visualize not only people's IDs, but also their personal profiles, which may lead to a lot of interesting applications, such as military training, pedestrian monitoring, homeland security, group tracking, business data analysis, people flow analysis, and so on, as illustrated in Fig. 2.

We remark that the above process is all based on the users' willingness to collaborate with our system. Nevertheless, our fusion system is still able to distinguish those people who collaborate and who do not. For the latter, we can tag them as "unknown persons".

4 Fusion methodology

We have introduced our fusion architecture above. Below, we provide more technique details of some specific modules.

4.1 People detection and tracking

For the drone videos, we retrieve all human objects from each image by YOLO [41], and then establish frame-to-frame association to connect these human objects by SORT [42]. Let $B_t^p = \{B_t^p(1), B_t^p(2), \dots\}$ be the set of human objects detected in frame F_t . Since we only care about user locations, each $B_t^p(i)$ is written as a coordinate $(x_t^p(i), y_t^p(i))$, which is the center of its bounding box in the image. SORT is applied to associate the same person in adjacent frames. However, when a person leaves the camera view or is partially occluded, his ID is destroyed. If he re-enters the view, a new unique ID will be created. Thus, the moving trajectory of the same person is not always the same and may be partitioned and assigned to multiple IDs by SORT. We denote by $tr(B_t^p(i))$ the trajectory of $B_t^p(i)$ that happened before F_t . If $B_t^p(i)$ is a new object to SORT, $tr(B_t^p(i)) = \emptyset$. Otherwise, its trajectory so far is $tr(B_t^p(i)) + B_t^p(i)$.

4.2 Coordination transformation

To maintain a consistent coordinate system, we place four ArUco markers [12] at four corners on the ground. Note that this requirement can be removed by some modern computer vision technologies. Figure 5 shows some ArUco markers [13]. We use these markers as invariants for projecting a pixel in a video frame onto our viewing plane (i.e., the ground space). Given a bounding box $B_t^p(i)$, which is a pixel, the transformation process involves the following steps.

4.2.1 Marker detection

Given an image frame where at least four ArUco markers are visible, we use the function *detectMarkers* of OpenCV for detection. Note that we only need to detect two markers if the recognition area is of a symmetrical shape. Let the four markers be found at the pixel coordinates, $P_1(x_1, y_1)$, $P_2(x_2, y_2)$, $P_3(x_3, y_3)$ and $P_4(x_4, y_4)$. During marker deployment, we also have their coordinates in the ground say, $P'_1(x'_1, y'_1)$, $P'_2(x'_2, y'_2)$, $P'_3(x'_3, y'_3)$ and $P'_4(x'_4, y'_4)$, respectively.

4.2.2 Transformation matrix generation

Using the locations of these markers on the ground as invariants, it is possible to compute a transformation matrix from the pixel space to the ground space. SIFT [43] is a feature detection algorithm in computer vision to detect and describe local features in images. It is widely used in applications such as object recognition, robotic mapping and navigation, and image stitching. It is also available in OpenCV. We use *findEssentialMat* in OpenCV to derive the transformation matrix M_t .

Fig. 5 ArUco markers



4.2.3 Coordinate transformation

By M_t , we then conduct the coordinate transformation:

$$B_t^G(i) = M_t \cdot (B_t^P(i)). \quad (1)$$

For example, Fig. 6a is a drone image. After transforming every pixel to the ground space, the ground image looks as Fig. 6b, where the four markers do form a square. Note that at least two markers are needed for the transformation.

4.3 Video feature extraction

Recall that $tr(B_t^G(i))$ is the trajectory of $B_t^G(i)$ at the ground space before time t . From trajectory $tr(B_t^G(i)) + B_t^G(i)$, we derive three motion-related features as follows:

1. *Acceleration feature* $F_t^V(i), f_1$: Since the sampling rate is fixed, the second derivation of $tr(B_t^G(i)) + B_t^G(i)$ is acceleration. So we set this feature value to $d(x_3, x_2) - d(x_2, x_1)$, where function $d()$ returns the Euclidean distance of two points and x_1, x_2 , and x_3 are the last three points in $tr(B_t^G(i)) + B_t^G(i)$, in that order.
2. *Orientation feature* $F_t^V(i), f_2$: This is the facial direction of a human object. We suggest that this feature can be derived by the last few points of $tr(B_t^G(i)) + B_t^G(i)$ together with the body shape in the bounding box at t .
3. *Rotation feature* $F_t^V(i), f_3$: This is the change of angle of the last three points in $tr(B_t^G(i)) + B_t^G(i)$. We suggest that this feature can be calculated by $a(\vec{x_3x_2} - \vec{x_2x_1})$, where function $a()$ returns the angle of a vector.

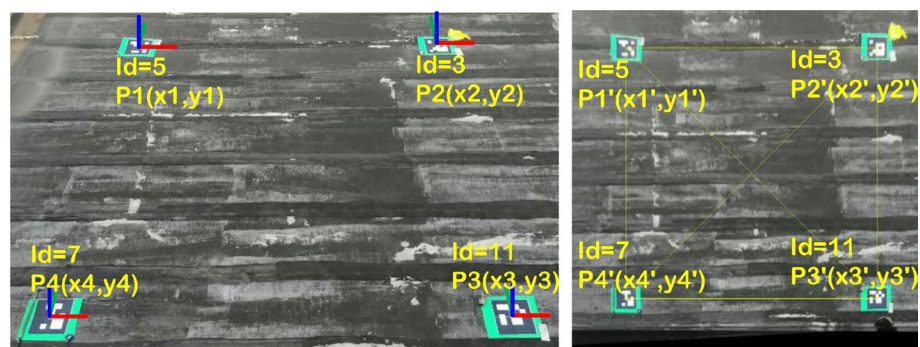
The above three features are derived for each bounding box $B_t^G(i)$. We make some remarks below. First, since a trajectory may not contain enough data points because of occlusion and object detection errors, we will mark the corresponding feature values as "unknown" if this is the case. Second, since people's walking behaviors may not be detectable if the video sampling rate is set to 30 fps, some down-sampling may be conducted on the raw data.

4.4 Sensor feature extraction

We assume that our wearable IoT devices have built-in inertial sensors, including accelerometer, gyro, and magnetometer field sensors. This allows us to monitor and retrieve users' motion-related features. For each wearable device j and its inertial data $I_t(j)$ at time t , we derive the same three features as follows.

1. *Acceleration feature* $F_t^S(j), f_1$: We obtain the acceleration feature directly from the 3-axis accelerometer after removing the component of gravity.
2. *Orientation feature* $F_t^S(j), f_2$: Device orientation is the position of the device in space relative to the Earth's magnetic north pole. We obtain the orientation feature directly from the geomagnetic field sensor.

Fig. 6 Coordinate transformation for drone view



(a) Before transformation

(b) After transformation

3. *Rotation feature* $F_t^S(j).f_3$: If wearable IoT device j is a newly appeared device or $|F_t^S(j).f_2 - F_{t-1}^S(j).f_2|$ is too small, the rotation feature $F_t^S(j).f_3$ is set to 0, to eliminate measurement errors of inertial sensor data. Otherwise, the rotation feature $F_t^S(j).f_3$ is set to $F_t^S(j).f_2 - F_{t-1}^S(j).f_2$.

4.5 Data fusion

The data fusion module calculates the correlation of video and sensor data, which consists of two tasks: Similarity Scoring and Paring. First, we measure the similarity score of any pair of normalized motion feature series, $F_t^V(i).f_k$ and $F_t^S(j).f_k$, $k = 1..r$. The scores for all features are then combined. By quantifying all-pair similarity scores, we are able to couple human objects with their IoT information. To achieve this goal, we need to address the asynchronous time and the warped data problems due to distributed sensors.

1. Time synchronization: Cameras and wearable devices have their own clocks and sampling frequencies. Communications may also suffer from errors and delays, making it hard to synchronize all these data.
2. Warped data series: In wireless communications, data loss is inevitable due to channel noise and network congestion. In the worst case, some data may be corrupted and can not be recovered.

DTW [44] is an algorithm for measuring similarity between two time series, which are sequences of observations, measured at successive times, spaced by time intervals. With DTW, we compare our time-specific feature points from videos and sensors regardless of the above two problems. With this, the Similarity Scoring sub-module generates a similarity matrix indexed by each pair of $F_t^V(i)$ and $F_t^S(j)$. To evaluate the difference between two time-series data $X(i)$ and $Y(j)$, $i = 1, \dots, |X|$ and $j = 1, \dots, |Y|$, DTW is derived based on dynamic programming:

$$D(i, j) = \min\{D(i-1, j), D(i-1, j-1), D(i, j-1)\} + d(X(i), Y(j)), \quad (2)$$

where $D(i, j)$ is the distance between $X[1 : i]$ and $Y[1 : j]$ with the initial condition $D(1, 1) = d(X(1), Y(1))$ and $d(X(i), Y(j))$ is the distance function between two data points. The goal of DTW is to find a mapping path between $X[1 : i]$ and $Y[1 : j]$, such that the total distance on this mapping path is minimized.

In our case, we evaluate the similarity for each feature between any pair of $F_t^V(i)$ and $F_t^S(j)$. These similarity scores must be normalized first. The purpose of normalization is to fairly treat each feature. The normalization factors are usually derived from experiments. Let us denote by $D(F_t^V(i), F_t^S(j))$ the combined similarity score (including all r features) between $F_t^V(i)$ and $F_t^S(j)$, where a smaller value represents higher similarity. The Pairing sub-module uses these similarity scores and matches human objects with their wearable devices by a greedy approach. Specifically, for each human object i , we find its best matched wearable device $Match(i)$ as follows:

$$Match(i) = \arg \min_{all j} (D(F_t^V(i), F_t^S(j))). \quad (3)$$

Then, the matched pair $(i, Match(i))$ is added to a set *PAIR*. This process is then repeated to find the next pair (paired ones will not be considered in the future process). The final set *PAIR* is the output. The pairing results couple human objects with their IoT devices, achieving our goal of tagging wearable device data on human objects in drone-recorded videos.

5 Performance evaluation

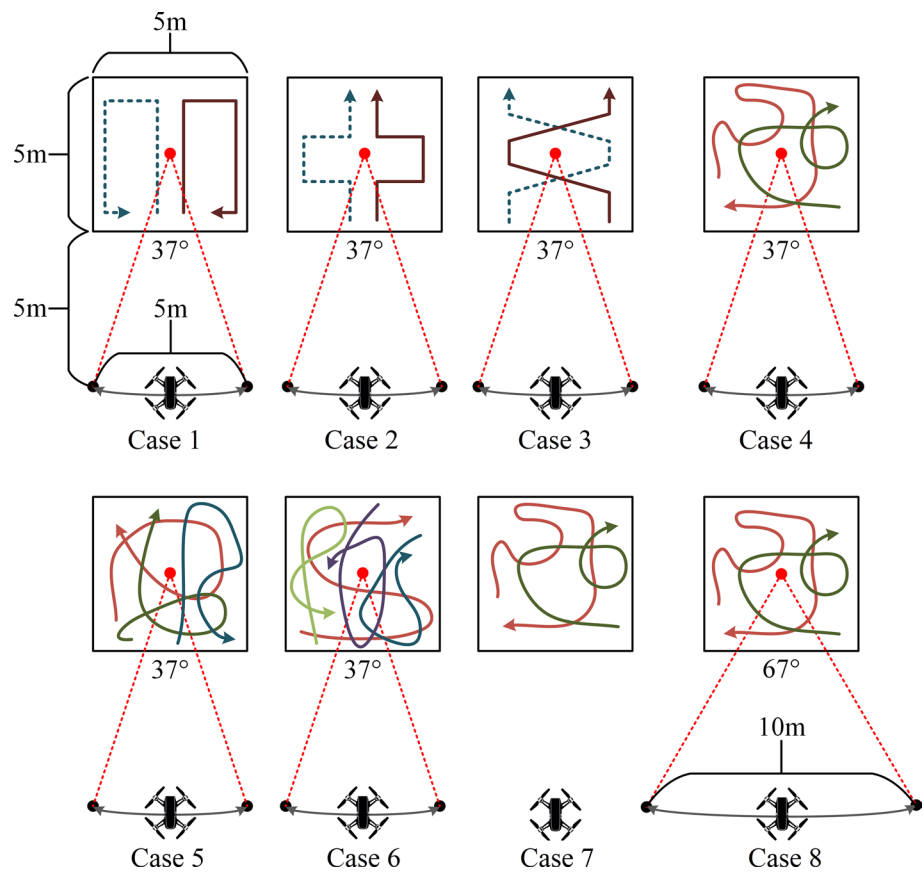
We implement a prototype of aerial surveillance system with the following components: (i) a DJI Spark, which has a camera with a 1/2.3 CMOS sensor of 11.8 megapixel resolution, a viewing angle of 81.9, and a sample rate of 30 fps, (ii) some wearable devices are simulated by smartphones (HTC 10) running the Android platform by setting their sensors' sample rate to 30 Hz, and (iii) a server equipped with an Intel Core i7-8700HQ processor, 32GB RAM, and a GeForce GTX1080 Ti GPU to conduct data fusion.

The experimental scene is shown in Fig. 7. We place four AuUco markers with the size of 42 cm × 42 cm on a square ground of 5 m × 5 m. There are a number of people walking in the square ground, all putting on wearable devices on their chests. An aerial camera keeps monitoring the square and continuously recording videos. Figure 8 shows eight

Fig. 7 Experiment scene and wearable device setup



Fig. 8 Experimental design: (i) Cases 1, 2, 3, 4 represent two peoples walking in four types of paths, (ii) Cases 4, 5, and 6 contain two, three, and four people, respectively, walking randomly, and (iii) the drone in Cases 7, 4, and 8 stay still, fly within 37° range, and fly within 67° range, respectively



experimental scenarios, which address three aspects: (i) different human walking paths, (ii) different number of people in drone view, and (iii) change of drone’s view angle and position. Each case is tested three rounds, and in each round, the drone flies four times while maintaining a height of 5 m. We adopt the *multiple object tracking accuracy* (MOTA) [45, 46] to calculate the accuracy of PID:

Fig. 9 PID accuracy

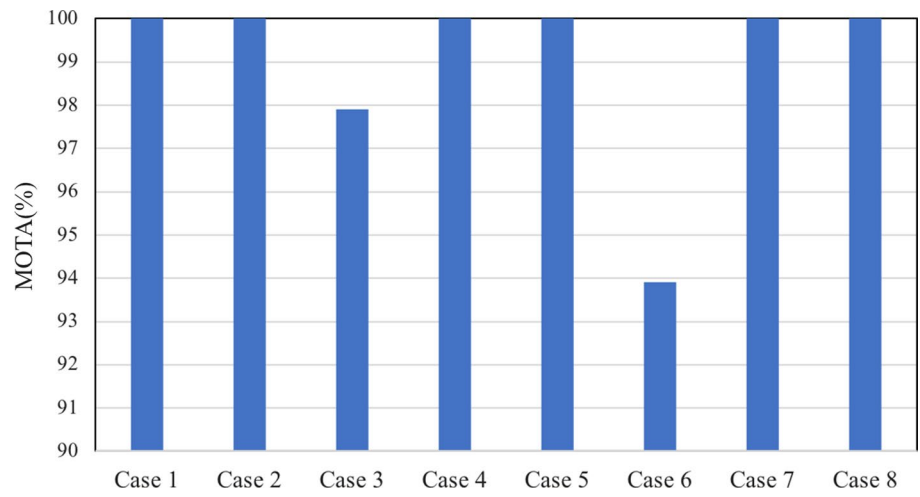
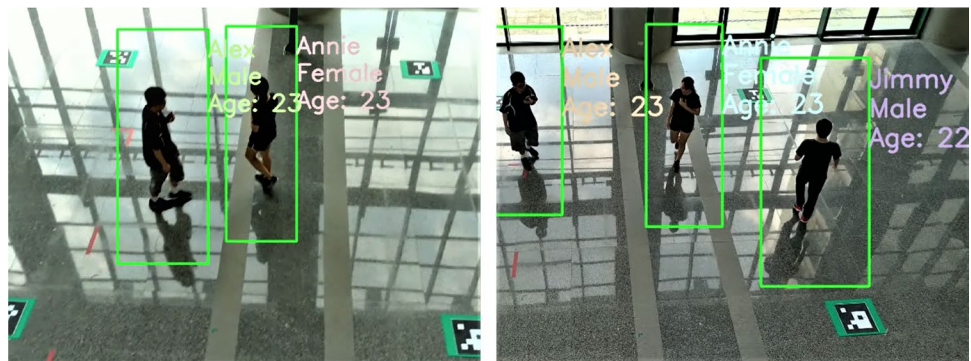
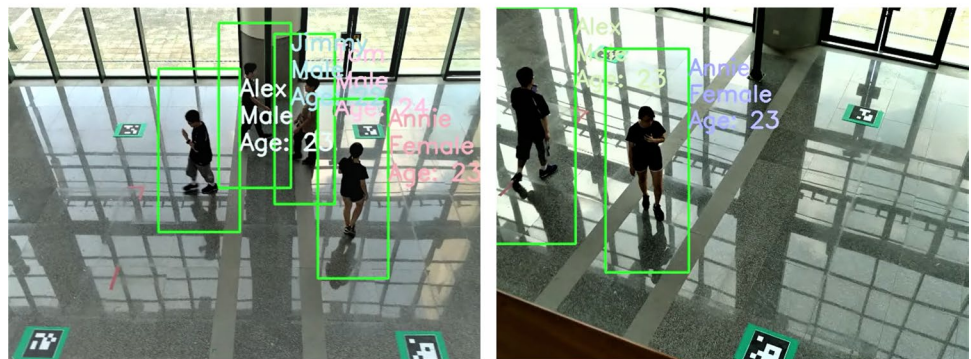


Fig. 10 Visualization of PID results



(a) Case 3.

(b) Case 4.



(c) Case 6.

(d) Case 8.

$$MOTA = \frac{\sum_{v_t} \text{correct identifications in } t}{\sum_{v_t} \text{all identifications in } t}. \tag{4}$$

6 Experimental results

Figure 9 shows the PID accuracy for the above 8 cases. Figure 10 shows some visualization results, where (a) contains two people walking in interleaving paths, (b) contains two people walking in random paths, (c) contains four people walking in random paths, and (d) is the case of drone flying within 67° angle.

6.1 Varying walking paths

As shown in Fig. 8, we design four types of walking paths, including circular (Case 1), non-interleaving (Case 2), interleaving (Case 3), and random (Case 4). In interleaving walk (Fig. 10a) or random walk experiments (Fig. 10b), the walking paths of volunteers interleave each other sometimes, causing occlusion in videos from time to time. However, since our data fusion model consider long-term inertial data, it is still possible to get correct pairing. As shown in Fig. 9, the average of MOTA achieves 99.5% for the cases 1, 2, 3, and 4.

6.2 Varying number of people

We also change the number of people walking in the field and analyze the fusion results. As shown in Fig. 9 (Cases 4, 5, and 6), the number of people walking in the field can affect the accuracy of our system. The accuracy of our system is 93.8% for four people walking randomly (Case 6). Figure 10c shows that when people walk in a crowd and cross each other, it is difficult to identify them by video only. But our system is still able to get correct pairing by taking advantage of sensor data.

6.3 Varying drone view angle and position

Further, we design different drone view angles to validate the robust of system under the following flight conditions: (i) the drone stays in the center position and continuously records videos for 1 min (Case 7), (ii) the drone flies within 37° angle range (Case 4), that is, the drone flies 5 m horizontally while maintaining a height of 5 m, and (iii) the drone flies within 67° angle range (Case 8), that is, the drone flies 10 m horizontally while maintaining a height of 5 m. Figure 10d shows some snapshots.

7 Conclusions

To conclude, this work presents a novel approach to fusing visual data from a drone camera and inertial data from wearable devices for person identification and tracking. We build a prototype to test the feasibility of the proposed system in a practical situation. To our best knowledge, this is the first work addressing tagging wearable IoT information on drone videos. We have tested changing position and view angle of a drone. Our approach requires no cumbersome data labeling process and does not rely on tracking human biological features. Our experiments show that even under occlusions or appearance changes, PID is still possible. The proposed PID and tracking system requires the authorization of accessing IoT data, and thus has potential applications in military training and construction site monitoring where people are required to wear badges. In the future, we would like to explore multiple intelligent IoT devices in larger surveillance regions with multiple drones under more complex environments.

Acknowledgements This work is supported in part by the Ministry of Science and Technology (MOST) under Grant MOST 108-2218-E-009-012, MOST Pervasive Artificial Intelligence Research (PAIR) Labs under contract MOST 109-2634-F-009-026, MOST 106-2221-E-009-028-MY3, and by the Ministry of Education (MOE) Center for Open Intelligent Connectivity.

Authors' contributions LDV, LYZ, and YCT wrote the main manuscript text. CHC, KLT, and KRW conduct the experiments and prepared all figures. All authors reviewed the manuscript. All authors read and approved the final manuscript.

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source,

provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Van LD, Chang CH, Tong KL, Wu KR, Zhang LY, Tseng YC. Demo: tagging IOT data in a drone view. In: Proceedings of the international conference on mobil computing and networking; 2019. p. 1–3.
2. Perera T, Priyankara A, Jayasinghe G. Unmanned arial vehicles (UAV) in smart agriculture: trends, benefits and future perspectives. In: Proceedings of the international research conference of Uva Wellassa University
3. Roberge V, Tarbouchi M, Labonté G. Fast genetic algorithm path planner for fixed-wing military UAV using GPU. *IEEE Trans Aerospace Electr Syst.* 2018;54(5):2105–17.
4. Qu T, Zang W, Peng Z, Liu J, Loh P. Construction site monitoring using uav oblique photogrammeter and bim technologies. In: Proceedings of the 22nd association for computer-aided architectural design research in asia conference; 2017. p. 654–62.
5. Manfreda S, McCabe M, Miller P, Lucas R, Pajuelo Madrigal V, Mallinis G, Ben Dor E, Helman D, Estes L, Ciraolo G. On the use of unmanned aerial systems for environmental monitoring. *Remote Sens.* 2018;641:1–28.
6. Nath RK, Bajpai R, Thapliyal H. IoT based indoor location detection system for smart home environment. In: Proceedings of the IEEE international conference on consumer electronics; 2018. p. 1–3.
7. Hassanalieragh M, Page A, Soyata T, Sharma G, Aktas M, Mateos G, Kantarci B, Andreescu S. Health monitoring and management using internet-of-things (IoT) sensing with cloud-based processing: Opportunities and challenges. In: Proceedings of the IEEE international conference on services computing; 2015. p. 285–92.
8. Paul A, Ahmad A, Rathore MM, Jabbar S. Smartbuddy: defining human behaviors using big data analytics in social internet of things. *IEEE Wireless Commun.* 2016;23(5):68–74.
9. Davis N, Pittaluga F, Panetta K. Facial recognition using human visual system algorithms for robotic and uav platforms. In: Proceedings of the IEEE conference on technologies for practical robot applications; 2013. p. 1–5.
10. Fysh M, Bindemann M. Person identification from drones by humans: insights from cognitive psychology. *Drones.* 2018;2(4):1–11.
11. Liu F, Zhang D, Shen L. Study on novel curvature features for 3D fingerprint recognition. *Neurocomputing.* 2015;168:599–608.
12. Chen J, Shen F, Chen DZ, Flynn PJ. Iris recognition based on human-interpretible features. *IEEE Trans Inf For Secur.* 2016;11(7):1476–85.
13. Garrido-Jurado S, Muñoz-Salinas R, Madrid-Cuevas FJ, Marín-Jiménez MJ. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recogn.* 2014;47(6):2280–92.
14. Chen B, Chen C, Wang J. Smart homecare surveillance system: behavior identification based on state-transition support vector machines and sound directivity pattern analysis. *IEEE Trans Syst Man Cybernetics Syst.* 2013;43(6):1279–89.
15. Qi M, Han J, Jiang J, Liu H, et al. Deep feature representation and multiple metric ensembles for person re-identification in security surveillance system. *Multimedia Tools Appl.* 2019;78:1–15.
16. Hsu SC, Wang YW, Huang CL. Human object identification for human-robot interaction by using fast r-cnn. In: Proceedings of the IEEE international conference on robotic computing; 2018. p. 201–4.
17. Wang T, Gong S, Zhu X, Wang S. Person re-identification by discriminative selection in video ranking. *IEEE Trans Pattern Anal Mach Intell.* 2016;38(12):2501–14.
18. Tang S, Andriluka M, Andres B, Schiele B. Multiple people tracking by lifted multicut and person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 3701–10.
19. Gong XW, Zhu SG. Person re-identification based on two-stream network with attention and pose features. *IEEE Access.* 2019;7(99):374–82.
20. Din C, Ta D. Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE Trans Pattern Anal Mach Intell.* 2018;40(4):1002–144.
21. Ge S, Zhao S, Li C, Li J. Low-resolution face recognition in the wild via selective knowledge distillation. *IEEE Trans Image Process.* 2019;28(4):2051–62.
22. Parkhi OM, Vedaldi A, Zisserman A. Deep face recognition. In: Proceedings of the British machine vision conference; 2015. p. 1–12.
23. Hsu HJ, Chen KT. Face recognition on drones: Issues and limitations. In: Proceedings of the workshop on micro aerial vehicle networks, systems, and applications for civilian use; 2015. p. 39–44.
24. face++. <https://www.faceplusplus.com/>.
25. Rekognition. <https://aws.amazon.com/tw/rekognition/>.
26. OpenCV. <https://opencv.org/>.
27. Das R, Piciuccio E, Maiorana E, Campisi P. Convolutional neural network for finger-vein-based biometric identification. *IEEE Trans Inf For Secur.* 2019;14(2):360–73.
28. O'Toole AJ, Phillips PJ, Weimer S, Roark DA, Ayyad J, Barwick R, Dunlop J. Recognizing people from dynamic and static faces and bodies: dissecting identity with a fusion approach. *Vision Res.* 2011;51(1):74–83.
29. Henschel R, Marcard TV, Rosenhahn B. Simultaneous identification and tracking of multiple people using video and imus. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops; 2020. p. 780–9.
30. Masullo A, Burghardt T, Damen D, Perrett T, Mirmehdi M. Who goes there? exploiting silhouettes and wearable signals for subject identification in multi-person environments. In: Proceedings IEEE/CVF international conference on computer vision workshop; 2019. p. 1599–607.
31. Masullo A, Burghardt T, Damen D, Perrett T, Mirmehdi M. Person re-id by fusion of video silhouettes and wearable signals for home monitoring applications. *Sensors.* 2020;20(9):2576–95.
32. Korany B, Karanam CR, Cai H, Mostofi Y. Xmodal-id: using wifi for through-wall person identification from candidate video footage. In: Proceedings of the international conference on mobile computing and networking; 2019. p. 1–15.

33. Naegeli T, Oberholzer S, Pluess S, Alonso-Mora J, Hilliges O. Flycon: real-time environment-independent multi-view human pose estimation with aerial vehicles. *ACM Trans Graph*. 2018;37(6):182.1–182.14.
34. Zhu P, Wen L, Bian X, Ling H, Hu Q. Vision meets drones: a challenge. *Comput Res Reposit*; 2018. p. 1–11.
35. Tzelepi M, Tefas A. Graph embedded convolutional neural networks in human crowd detection for drone flight safety. *IEEE Trans Emerg Topics Comput Intell*. 2019;99:1–14.
36. Layne R, Hospedales TM, Gong S. Investigating open-world person re-identification using a drone. In: *European conference on computer vision*; 2015. p. 225–40.
37. Portmann J, Lynen S, Chli M, Siegwart R. People detection and tracking from aerial thermal views. In: *Proceedings IEEE international conference on robotics and automation*; 2014. p. 1794–800.
38. Fradi H, Bracco L, Canino F, Dugelay JL. Autonomous person detection and tracking framework using unmanned aerial vehicles (uavs). In: *Proceedings European signal processing conference*; 2018. p. 1047–51.
39. Shen Q, Jiang L, Xiong H. Person tracking and frontal face capture with uav. In: *Proceedings of the international conference on communication technology*; 2018. p. 1412–6.
40. Nguyen HD, Na IS, Kim SH, Lee GS, Yang HJ, Choi JH. Multiple human tracking in drone image. *Multimedia Tools Appl*. 2019;78(4):4563–77.
41. Redmon J, Farhadi A. Yolov3: an incremental improvement; 2018. CoRR. abs/1804.02767. [arxiv:1804.02767](https://arxiv.org/abs/1804.02767)
42. Bewley A, Ge Z, Ott L, Ramos F, Upcroft B. Simple online and realtime tracking. In: *Proceedings of the IEEE international conference on image processing*; 2016. p. 3464–8.
43. Lowe DG et al. Object recognition from local scale-invariant features. In: *Proceedings of the IEEE international conference on computer vision*. 1999; 2:1150–7.
44. Berndt DJ, Clifford J. Using dynamic time warping to find patterns in time series. In: *Proceedings KDD workshop*; 1994. p. 359–70.
45. Bernardin K, Elbs A, Stiefelhagen R. Multiple object tracking performance metrics and evaluation in a smart room environment. In: *Proceedings IEEE international workshop on visual surveillance, in conjunction with ECCV*; 2006. p. 1–8.
46. Teixeira T, Jung D, Savvides A. Tasking networked cctv cameras and mobile phones to identify and localize multiple people. In: *Proceedings of the ACM international conference on ubiquitous computing*; 2010. p. 213–22.