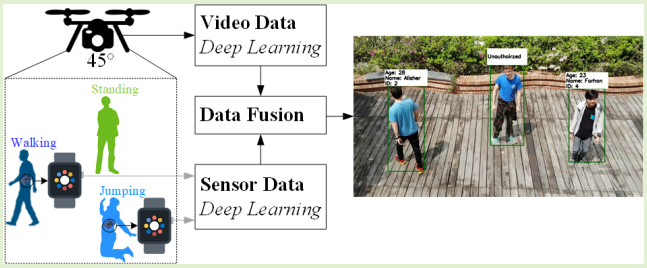
Person Tracking by Fusing Posture Data From UAV Video and Wearable Sensors

Alisher Mukashev, Lan-Da Van, *Senior Member, IEEE*, Susanta Sharma, M. Farhan Tandia, and Yu-Chee Tseng, *Fellow, IEEE*

***Abstract*—In this article, a novel framework that fuses the posture data taken by a drone (or unmanned aerial vehicle, UAV) camera and the wearable sensors data recorded by smartwatches is proposed. The framework is designed for continuously tracking persons in a drone view by analyzing location-independent human posture features and correctly tagging smartwatch identities (IDs) and personal profiles to video human objects, thus conquering the former work in requiring ground markers. Person detection, ID assignment, and pose estimation are integrated into our framework to obtain recognized human postures. These recognized postures are then paired with those from the wearable sensors. Through fusing common postures, such as standing, walking, jumping, and falling down, person tracking accuracy by UAV up to 95.36% can be attained in our testing scenarios.**



***Index Terms*—Action recognition, data fusion, drone, person identification, person tracking, wearable devices.**

I. Introduction

PERSON tracking is an important issue for most robotic platforms. Among these platforms, unmanned aerial vehicles (UAVs) have been adopted in many practical fields, including construction, agriculture, and environmental monitoring. Most of these applications rely on state-of-the-art computer vision technologies, such as object detection [1], [2], semantic segmentation [3], instance segmentation [4], and action recognition [5]. On the other hand, person tracking, trajectory tracing, and physical behavior analysis have greatly benefited from the advance of the Internet of Things (IoT) technologies.

Although computer vision and IoT do not seem to be highly related, we see a new opportunity by integrating UAV and IoT to solve the person-tracking problem scenario by fusing posture data from both drone videos and wearable IoT sensors. Previous works [6], [7], [8], [9], [10] focused on fusion of computer vision and IoT approaches, but they

Manuscript received 15 September 2022; accepted 12 October 2022. Date of publication 7 November 2022; date of current version 14 December 2022. This work was supported in part by the Ministry of Science and Technology (MOST) through the Pervasive Artificial Intelligence Research (PAIR) Labs under Grant MOST 109-2634-F-009-026, Grant MOST 109-2221-E-009-084-MY3, and Grant MOST 110-2634-F-A49-004. The associate editor coordinating the review of this article and approving it for publication was Dr. Xiaojin Zhao. *(Corresponding author: Alisher Mukashev.)*

The authors are with the Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan (e-mail: alishsuper.eed07g@nctu.eud.tw; ldvan@cs.nycu.edu.tw; susantasharma.cs06g@nctu.edu.tw; farhantandia.eic08g@nctu.eud.tw; yctseng@cs.nycu.edu.tw).

Digital Object Identifier 10.1109/JSEN.2022.3218373

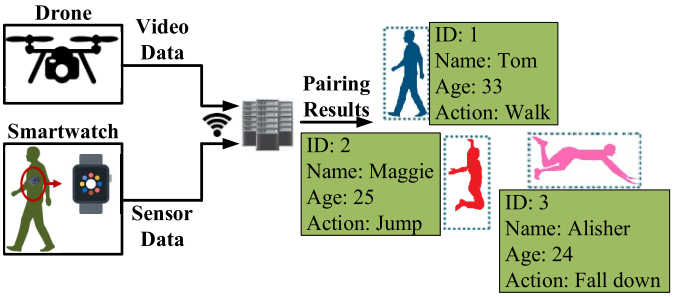


Fig. 1. Personal tracking scenario.

[6], [7], [8], [9] had the following constraints. First, in [6] and [7], ArUco markers must be used to transform a human object's location from a pixel space to a ground space. The proposed person-tracking system in this article conquers this problem by using only location-independent motion feature information. We use human body keypoints to obtain motion feature information. Person detection helps us to remove the false pose information and to enhance system performance. Second, the works [8], [9] applied static RGB-D cameras to get visual features, while this work uses the camera on a drone, which is a highly dynamic platform with changing view angles and view distances. Therefore, high-level posture information, rather than raw motion values as in [8] and [9], needs to be explored in order to conduct dynamic surveillance on a drone. Based on these observations and state-of-the-art tools, this proposed system elegantly integrates computer vision tools and IoT technologies to achieve person tracking.

Our person-tracking scenario is shown in Fig. 1. There is a drone conducting surveillance. A number of people are in the

drone view and some of them may put on their smartwatches. A fusion server collects both drone videos and smartwatch sensor data for posture analysis. Behaviors such as standing, walking, jumping, and falling down from both posture data sources are fused. In the end, we are able to correctly tag IoT data (e.g., smartwatch IDs and personal profiles) on human objects captured by a drone view. The technical novelties of this article are summarized as follows.

1) We use postures, which are location-independent features, to exclude the need of ground markers.
2) The proposed person-tracking framework elegantly fuses multimodality sensors (wearable device and drone camera) that observe the same postures.
3) The results can continuously tag smartwatch IDs and personal profiles, such as name and age, on the human objects, thus bypassing the ID switch problem caused by a video-based solution in surveillance applications.
4) Our framework is highly modularized, allowing future extension to other sensors or tools.

II. Related Work

Vision recognition plays an important role in a series of UAV tasks and applications. The work [11] used an autonomous drone swarm approach for multiview human poses estimation. The work [12] utilized a fully convolutional neural network for human crowd detection from drone-captured images. The work [13] presented an algorithm and a dataset for pedestrian detection with UAV. The work [14] presented real-time human and gesture detection by an on-board computing UAV for rescue applications. The work [15] focused on colorwise safety helmet detection and counting of workers by an edge-controlled outdoor autonomous UAV in construction sites. The work [16] explored how to perform action recognition with a few aerial videos. The work [17] provided rich UAV-based datasets for action recognition research. However, these works focused on human detection in UAV videos, but cannot provide further information about the detected humans such as identities (IDs). Several works have studied person tracking on UAVs. The work [18] combined frontal face perception and color feature descriptors to track people. The work [19] presented a new particle filter-based algorithm to track down human targets in aerial thermal images. In [20], a visual data stream from drones for detecting and tracking people was applied. However, these solutions [18], [19], [20] are all sensitive to environmental factors and are vulnerable to target appearance changes. The work [21] proposed a multiperson tracking and identification solution while considering the features from video and IMUs. The work [22] used a joint graphical model to track persons. However, the video streams of [21], [22] are acquired from static cameras, so the solution may not be easily applicable to dynamic drone views. Bio-features such as fingerprint and iris were used in [23], [24], and [25]. Although combining multiple features helps improve accuracy, acquiring these bio-features requires close contact with specific devices. Due to changing height and viewing angle of drone, applying bio-features to drone is more difficult.

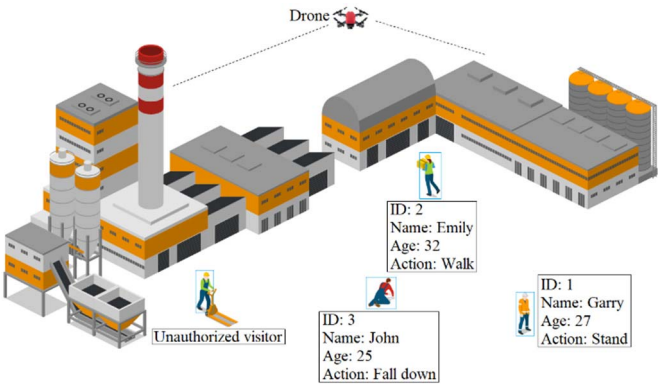


Fig. 2. Imaginary scenario in a factory space.

Recently, a lot of skeleton action recognition solutions are proposed based on deep learning and hand-crafted techniques. The work [26] presented co-occurrence feature learning using regularized long short-term memory (LSTM) networks, which is an RNN-based model. The work [27] analyzed human actions by a temporal CNN. The works [27] and [28] studied skeleton sequences for 3-D action recognition.

Considering action recognition by wearable sensors, the work [29] provided a review on this topic. The work [30] utilized the streaming accelerometer data from a smartwatch to detect falls. The work [31] proposed a kernel fusion-based extreme learning machine to adapt the model to various sensor locations on a human body, so the generalizability of the model can be attained. The work [32] performed a comprehensive performance comparison of human action recognition between different deep learning models on large-scale datasets. The work [33] proposed probabilistic neural networks and an adjustable fuzzy clustering to improve incremental learning abilities of their human activity recognition model by wearable sensors. This helps to learn features from new training data without previously used training data.

III. System Model

Fig. 2 shows our system application scenario. We consider a controlled environment with a crowd of people monitored by a drone. Some users may put on their smartwatches, but some may not. We intend to not only track the people but also tag their smartwatch IDs and personal profiles. The notations used in this article are defined in Table I. The proposed person tracking system is shown in Fig. 3. There are two input data flows from the drone camera and smartwatches, and one data fusion module, all of which are detailed below.

A. UAV Video Data Flow

The drone continuously monitors the field and processes its video data. The data preprocessing block conducts person detection, ID assignment, and pose estimation. In particular, integrating ID assignment and pose estimation can enhance person tracking accuracy. The motion image frame taken at time t by the drone is denoted by M_t^D. In order to reduce the possibility of getting the nonhuman joint information, we do person detection to identify all human bounding boxes

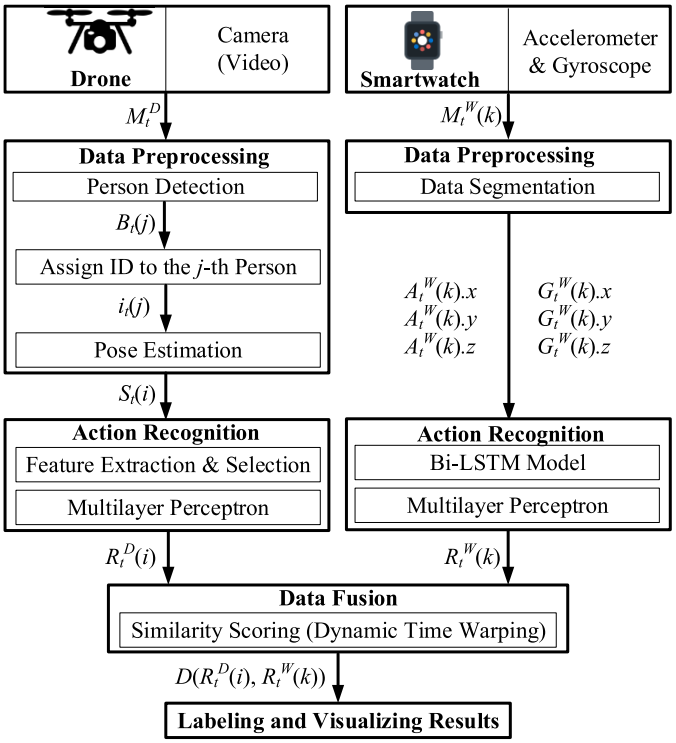


Fig. 3. Proposed person tracking system architecture.

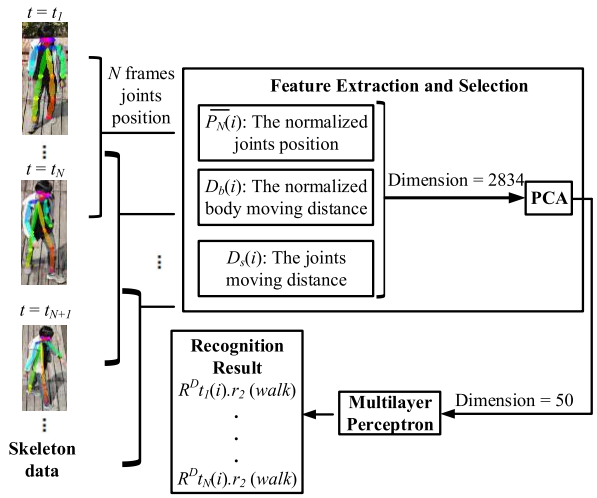


Fig. 4. Workflow of the video-based action recognition [34].

$B_t(j)$ in M_t^D first. Next, an object tracking algorithm for a video sequence is applied to tackle the occlusion problem by considering motion and appearance information and to assign a unique ID number $i_t(j)$ to each human object (note that these IDs are sequence numbers, not smartwatch IDs). Then, we detect the human skeleton (joint positions) in each bounding box $B_t(j)$, denoted as $S_t(i)$. This completes the data preprocessing part.

Next, the action recognition workflow [34] is shown in Fig. 4. There are two submodules: feature extraction and selection, and multilayer perceptron (MLP). A sliding window is applied to aggregate the skeleton data within N continuous frames [34] in Fig. 4. These skeleton sequences are fed into these two submodules to obtain the activity type of each skeleton sequence in a window. In our model, each skeleton has 18 joints, spreading among head, neck, arms, and legs. Each joint position is represented by (x, y) in the image space. Since bounding box sizes are different, it is necessary to scale these coordinates to the same unit before further processing. Since the head, which contains five joints, is not helpful for activity recognition, we remove all joints on the head. From N continuous frames, we hope to retrieve more salient features that may help distinguish human action types. Our goal is to distinguish four action types: standing, walking, jumping, and falling down. We consider features $\overline{P_N}(i)$, $D_b(i)$, and $D_s(i)$, which are defined in Table I. Considering these features together, the feature vector dimension is formulated in the following equation:

$$\begin{aligned} F_D &= \text{Dimension of } \overline{P_N}(i) + (\text{Dimension of } D_b(i)) \cdot 10 \\ &\quad + \text{Dimension of } D_s(i) \\ &= 13 \cdot 2 \cdot N + 2 \cdot (N-1) \cdot 10 + 13 \cdot 2 \cdot (N-1) \end{aligned} \quad (1)$$

TABLE I
SYMBOL DEFINITION

Symbol	Description
M_t^D	The motion image frame taken at time t by the drone.
$B_t(j)$	The bounding box of the j-th person at time t.
$i_t(j)$	The ID number i assigned to the j-th person at time t.
$S_t(i)$	The skeleton joint set of the person with ID number i at time t, denoted by $\{i_t(j): [(x_1, y_1), (x_2, y_2), \ldots, (x_c, y_c)]\}$.
$R_t^D(i)$	The motion recognition result of the person with ID number i at time t from the video $\{R_t^D(i).r_1, R_t^D(i).r_2, \ldots, R_t^D(i).r_p\}$, where p is the number of postures.
$P_N(i)$	The joint positions of the person with ID number i of concatenated N frames.
H	The average height of the skeleton, which equals the length from neck to thigh.
$\overline{P_N}(i)$	The normalized joint positions of the person with ID number i of concatenated N frames, which equal $\frac{[P-\mu \cdot P]}{H}$, where μ is an arithmetic mean, and $P \in P_N(i)$.
$D_b(i)$	The normalized body moving distance of the person with ID number i between adjacent frames, where the normalization equals the move of the neck divided by H.
$D_s(i)$	The skeleton joint moving distances of the person with ID number i between adjacent frames.
F_D	The feature vector dimension.
$M_t^W(k)$	The motion sensor data of the k-th person, taken at time t by the smartwatch.
$A_t^W(k)$	The accelerometer data of the k-th person at time t $\{A_t^W(k).x, A_t^W(k).y, A_t^W(k).z\}$.
$G_t^W(k)$	The gyroscope data of the k-th person at time t $\{G_t^W(k).x, G_t^W(k).y, G_t^W(k).z\}$.
$R_t^W(k)$	The motion recognition result of the k-th person at time t from the smartwatch $\{R_t^W(k).r_1, R_t^W(k).r_2, \ldots, R_t^W(k).r_p\}$, where p is the number of postures.
$D(R_t^D(i), R_t^W(k))$	The DTW distance for the matched pair of the motion recognition result with ID number i from the video and motion recognition result of the k-th person from the smartwatch.

TABLE II
CLASSIFICATION ACCURACY ANALYSIS CONSIDERING PCA DIMENSION AND NUMBER OF FRAMES

# of Frames / PCA Dim	20	25	30	35	40	45	50
20	81.75	77.96	79.72	81.76	82.31	81.49	87.47
30	94.82	95.21	94.34	89.37	91.52	92.65	92.77
40	95.04	96.18	97.28	97.01	96.44	96.24	93.64
50	94.81	96.23	96.79	97.82	98.24	98.12	97.17
60	94.33	95.79	96.12	97.34	98	97.95	97.91
70	95.79	95.24	95.71	95.86	96.85	97.58	97.8

TABLE III
CLASSIFIER PERFORMANCE ANALYSIS

Method	Main Settings [34]	Classification Accuracy (%)
kNN	k=5	94.88
SVM	Kernel type – linear	61.13
SVM	Kernel type – RBF	72.05
MLP	Number of layers = 3, number of neurons = 100	98.24
Random Forest	Depth=30, trees=100	96.43

where 13 joints and 2 dimension positions are adopted in (1). In order to make the number of features $D_b(i)$ approach to other terms, we increase its features in the learning process by repeating each item 10 times, as it is shown in (1). For example, when $N = 40$, $F_D = 2834$. Then, the principal component analysis (PCA) algorithm is adopted to reduce the feature vector dimension. We analyzed the number of frames and PCA dimension in terms of classification accuracy in Table II. As we can see from Table II, the optimized PCA dimension and the number of frames are 50 and 40, respectively.

Next, we apply machine learning algorithms, such as MLP, k-nearest neighbor (kNN), support vector machine (SVM), and random forest to classify each feature vector. To evaluate the performance of each classifier [34], [35], we ran experiments on our own dataset, where most main settings follow the values of [34]. According to Table III, the MLP algorithm performs better, so we adopt this classifier in our work.

Going through the MLP, the motion recognition result R_t^D can be obtained. We define the motion recognition result $R_t^D(i)$ of the person(s) with ID number i from the drone view as shown in (2). We choose standing, walking, jumping, and falling down as they are the most frequently occurring actions in our daily life. The result "*Unknown*" is to take the other actions, camera instability, and action switching into consideration

$$R_t^D(i) = \begin{cases} R_t^D(i) \cdot r_1, & \text{Stand} \\ R_t^D(i) \cdot r_2, & \text{Walk} \\ R_t^D(i) \cdot r_3, & \text{Jump} \\ R_t^D(i) \cdot r_4, & \text{Falldown} \\ R_t^D(i) \cdot r_5, & \text{Unknown.} \end{cases} \quad (2)$$

B. Smartwatch Data Flow

Wearable sensor data including standing, walking, jumping, and falling down is collected via a smartwatch. More specifically, we collect tri-axial accelerometer data and tri-axial gyroscope inertial sensor data for action recognition. The real sensor data waveforms and corresponding action types are shown in Fig. 5. Jumping data are obtained by performing jumping continuously, and falling data are obtained by performing falling-stand-falling and so forth. The difference between actions can be observed in terms of their magnitudes and periods. From our data, the magnitude order from large to small accelerations should be falling down, jumping, walking, and standing. On the other hand, walking and jumping have periodical waveforms from gyroscope.

Human action recognition $R_t^W(k)$ for wearable sensor data $M_t^W(k)$ is conducted by input data segmentation, one-layer bidirectional long-short term memory (Bi-LSTM) [36] with 200 neurons to extract the features from the hand movement dependencies and is classified by one-layer MLP with 100 neurons in Fig. 6. Bi-LSTM (i.e., an improved version of LSTM [37]) consisting of two LSTMs is suitable to obtain more context than the standard LSTM. Herein, one LSTM is used to learn the context from the past to the future, while the other one is used to learn the context from the future to the past. However, the standard LSTM only captures the previous context without regard to the future context. In order to train our Bi-LSTM model, we need to split each collected data into a number of windows (i.e., data segmentation whose outputs are denoted by $\{A_t^W(k) \cdot x, A_t^W(k) \cdot y, A_t^W(k) \cdot z\}$ for tri-axial accelerometer data and $\{G_t^W(k) \cdot x, G_t^W(k) \cdot y, G_t^W(k) \cdot z\}$ for tri-axial gyroscope inertial sensor data). Through analyzing different window sizes, we find the optimized window size to be about 120 data points, which comes out from 6 s data at 20 Hz. The Bi-LSTM learns to map and predict each window sensor data to an activity as shown in Fig. 6. The output feature maps as shown in Fig. 7 are randomly selected from the output of Bi-LSTM. We can see that these four actions have different feature maps such that the posture detection can be well performed by our model.

C. Data Fusion

The goal of the data fusion module in Fig. 3 is to generate the best matching among $R_t^D(i)$ and $R_t^W(k)$. Ideally, each video sequence (i.e., $R_t^D(i)$) can be matched to the sensor data (i.e., $R_t^W(k)$) that belongs to the same person. Therefore, we can augment user k's smartwatch ID and personal profile on the video object with ID number i to realize our person tracking scenario in Fig. 1. However, in practice, there are several difficulties. First, person(s) will leave and reenter the drone view such that an ID switch problem occurs. Second, the persons within the drone view sometimes/frequently have occlusion with each other such that IDs may switch. Third, people may have the same action types in some cases. Fourth, the video data coming from a drone and the sensor data coming from smartwatches may not synchronize in time. These are all difficult issues to be tackled in our matching task.

Herein, we use different postures of video data and wearable data in the fusion framework to bypass the ID switch effect for the first and second difficulties. Herein, the smartwatch IDs will be used as our tagging IDs in the drone view. For the third difficulty, herein, the different postures for different

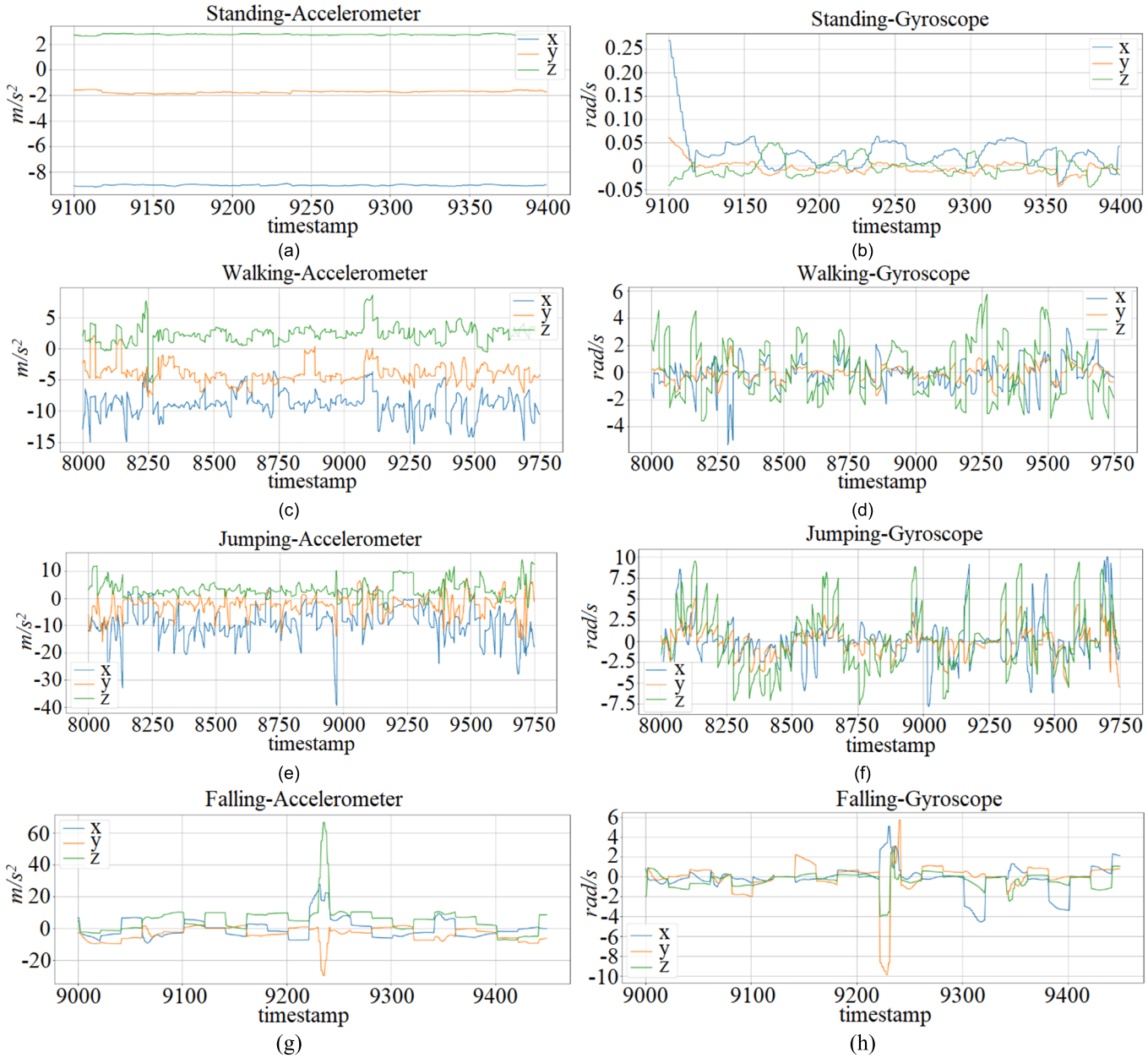


Fig. 5. Capture of accelerometer and gyroscope data.

people are requested before they might have the same actions. Once the proposed system pairs the best match, the system could continuously recognize and track the different people who do the same action without occlusion in the remaining time. This is because the dynamic time warping (DTW) [38] can keep the minimal distance for matched pairs in the remaining time. How to correctly synchronize distributed data and occasional interrupted data is another issue. Herein, similar to [6], [7], [8], [9], the DTW [38] scheme is used to solve this issue, where the detailed derivation and descriptions can be referred to [39]. Here, we would like to point out that due to the use of the tracking algorithm, we can trace back the data prior to $R_t^D(i)$. That is, we may find a sequence $\{R_{t-N}^D(i),\ R_{t-N+1}^D(i), \ldots, R_t^D(i)\}$ that are recognized as the same ID number i. Similarly, for $R_t^W(k)$ because it is always from the same smartwatch, we can also trace back to the sequence $\{R_{t-N}^W(k),\ R_{t-N+1}^W(k), \ldots, R_t^W(k)\}$ belonging to user k. DTW can be applied to the above sequences for each pair of video object(s) with ID number i and user k. A pair will be determined as a best matched pair by the minimal DTW distance ($D(R_t^D, R_t^W)$) between action sequences from video data ($R_t^D(i)$) and sensor data ($R_t^W(k)$) for all pairs

$$D\left(R_t^D, R_t^W\right) = d\left(R_t^D, R_t^W\right) + \min \begin{cases} D\left(R_{t-1}^D, R_t^W\right) \\ D\left(R_{t-1}^D, R_{t-1}^W\right) \\ D\left(R_t^D, R_{t-1}^W\right) \end{cases} \quad (3)$$

where $d(R_t^D, R_t^W)$ denotes the distance between two different sequences and is defined as follows:

$$d(i,j) = \begin{cases} d_0 = 0, & \text{if } R_t^D = R_t^W \\ d_1 = 1, & \text{if } R_t^D \neq R_t^W. \end{cases} \quad (4)$$

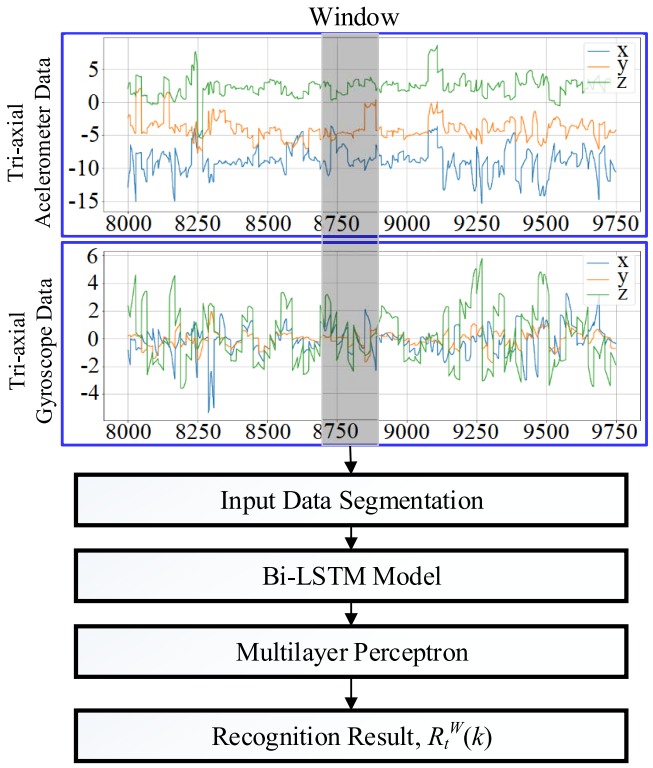


Fig. 6. Workflow of the sensor-based walking action recognition.

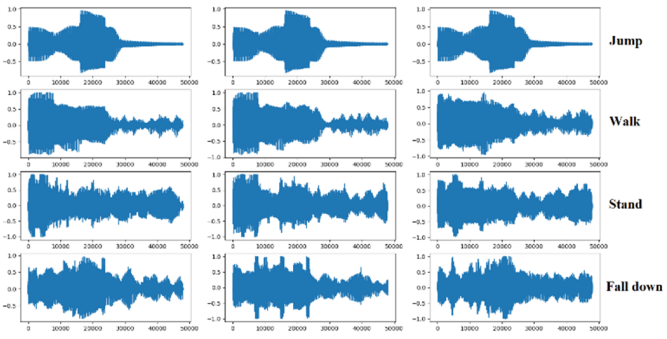


Fig. 7. Feature maps for four different actions.

TABLE IV
EXAMPLE OF BEST MATCHED PAIRS

	$R_t^D(i=1)$	$R_t^D(i=2)$
$R_t^W(k=1)$	$4 \cdot d_0 = 0$	$4 \cdot d_1 = 4$
$R_t^W(k=2)$	$4 \cdot d_1 = 4$	$4 \cdot d_0 = 0$

For example, in Table IV, there are two people with two smartwatches, where one is standing and jumping, and the other is walking and falling down. Assume that the video action sequence with $i = 1$ obtained by the drone is $R_t^D(i = 1) = \{\text{Stand, Stand, Jump, Stand}\}$, and the video action sequence with $i = 2$ is $R_t^D(i = 2) = \{\text{Walk, Walk, Walk, Falldown}\}$. Assume that the action sequence obtained by smartwatches from the first person is $R_t^W(k = 1) = \{\text{Stand, Stand, Jump, Stand}\}$, and the action sequence from the second person is $R_t^W(k = 2) = \{\text{Walk, Walk, Walk, Falldown}\}$. We calculate the DTW distance among them to generate a similarity matrix as shown in Table IV. It is obvious that the best-matched pairs are $\{R_t^D(i = 1), R_t^W(k = 1)\}$ and $\{R_t^D(i = 2), R_t^W(k = 2)\}$. According to the paring results, the smartwatch IDs and personal profiles can be correctly tagged on drone videos.

IV. Evaluation Results

A. Dataset

Since there is not a lot of available drone view data and wearable sensor data in the open source dataset, we collect video data and sensor data by ourselves to prove the proposed system concept. We record the video with a drone for seven people performing actions including standing, walking, jumping, and falling down from different viewing angles for the dataset. A total of around 24 400 samples are collected, where 70% dataset is used for training and 30% dataset is used for validation.

To make sure that the collected sensor data are uniformly distributed, we collect the smartwatch data from the left hand and the right hand of four different individuals for the dataset. A total of 64 000 samples are collected, consisting of tri-axis accelerometer data and tri-axis gyroscope data for four actions as shown in Fig. 5 as an example. About 80% dataset is used for training and 20% dataset is used for validation. Our model used Adam optimizer with learning rate equals 0.0001.

B. Environment Setup

We have built a prototype to validate our model. The following hardware specifications are adopted in our prototype.

1) *UAV:* DJI Spark with a camera of 11.8-Mpixel resolution, a viewing angle of 81.9°, and a frame rate of 30 fps.
2) *Smartwatch:* Ticwatch Pro and/or Huawei watch 2 equipped with a tri-axis accelerometer and other sensors.
3) *Fusion Server:* Intel Core i7-8750 CPU with 16 GB RAM and a GeForce RTX2070 GPU.

The following tools are used for data preprocessing.

1) Sensor Manager API [40] is used for accessing the wearable sensors data $M_t^W(k)$.
2) YOLO v3 [1], SORT [41], DeepSORT [42] are used to detect and track all human objects with their bounding boxes $B_t(j)$ and IDs $i_t(j)$ for each image frame M_t^D.
3) OpenPose [43] uses the two-branch CNN to predict confidence maps and part affinity fields to associate the joints into human skeletons.

Fig. 8 shows a test scenario with four persons for Case 2 of Fig. 9, where five testing cases are described as follows.

Case 1: 60° angle with four persons and two smartwatches as shown in Fig. 9(a).

Case 2: 60° angle with four persons and three smartwatches as shown in Fig. 9(b).

Case 3: 45° angle with three persons and two smartwatches as shown in Fig. 9(c).

Case 4: 45° angle with four persons and three smartwatches as shown in Fig. 9(d).

Case 5: The drone changes the degree angle from 60° to 45° and moves horizontally 2 m with four persons and two smartwatches as shown in Fig. 9(e).

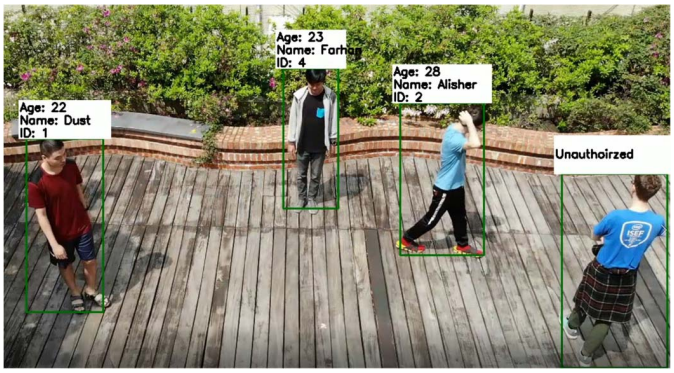


Fig. 8. Testing scenario for Case 2 of Fig. 9.

TABLE V
CONFIGURATION DEFINITION OF THE PROPOSED FRAMEWORK

Configuration	Description
Config. 1	YOLO + SORT + OpenPose + Smartwatch
Config. 2	YOLO + DeepSORT + OpenPose + Smartwatch

Herein, we develop two configurations (Config. 1 and Config. 2) as defined in Table V for each case. The main difference between Configuration 1 and Configuration 2 is to use SORT and DeepSORT. The former is based on Kalman filter and Hungarian algorithm. The latter uses deep association metric. Our framework allows to modify the configuration with available different tools.

C. Evaluation Results

To evaluate the person tracking performance among five cases and two configurations, we adopt multiple object tracking accuracy (MOTA) as follows:

$$\text{MOTA} = \frac{\sum_t \text{Correct identifications at time } t}{\sum_t \text{All identifications at time } t}. \quad (5)$$

The performance results are shown in Table VI using our own dataset. From Table VI, we use the posture to exclude the assisted ground markers and fuse posture data from drone views and wearable sensors to attain a dynamic person tracking and to tag smartwatch IDs and personal profiles in the drone view to bypass the ID switch problem. Due to modularity of the proposed framework, we can develop two configurations for persona tacking. In order to compare YOLO + SORT and YOLO + DeepSORT in terms of MOTA, we use clothes color as the reference answer to calculate MOTA for YOLO + SORT and YOLO + DeepSORT. The best performance among five cases in our experiment is 95.36%. Unlike YOLO + SORT and YOLO + DeepSORT, our proposed system could continuously and correctly tag smartwatch IDs and personal profile information to video human objects due to the fusion of drone view and smartwatch data. Thus, the proposed Configuration 1 and Configuration 2 show comparable performance in terms of MOTA in Table VI.

Fig. 10 shows the person tracking performance versus experiment running time. We pick up the best performance curve for Case1, Case 3, and Case 5. It is observed that, at the very beginning of the experiment, errors are high because the sequences to be compared are too short and people may perform the same actions. However, over time, sequences (belonging to some persons) start to accumulate, so MOTA increases. We observe that the longer the experiment takes place, the higher MOTA. Within one minute, MOTA can attain 80%–90% for Case 1, Case 3, and Case 5. In Fig. 10, when the tracked person may sometimes be blocked by another person (i.e., occlusion) or the person may leave the camera view, the

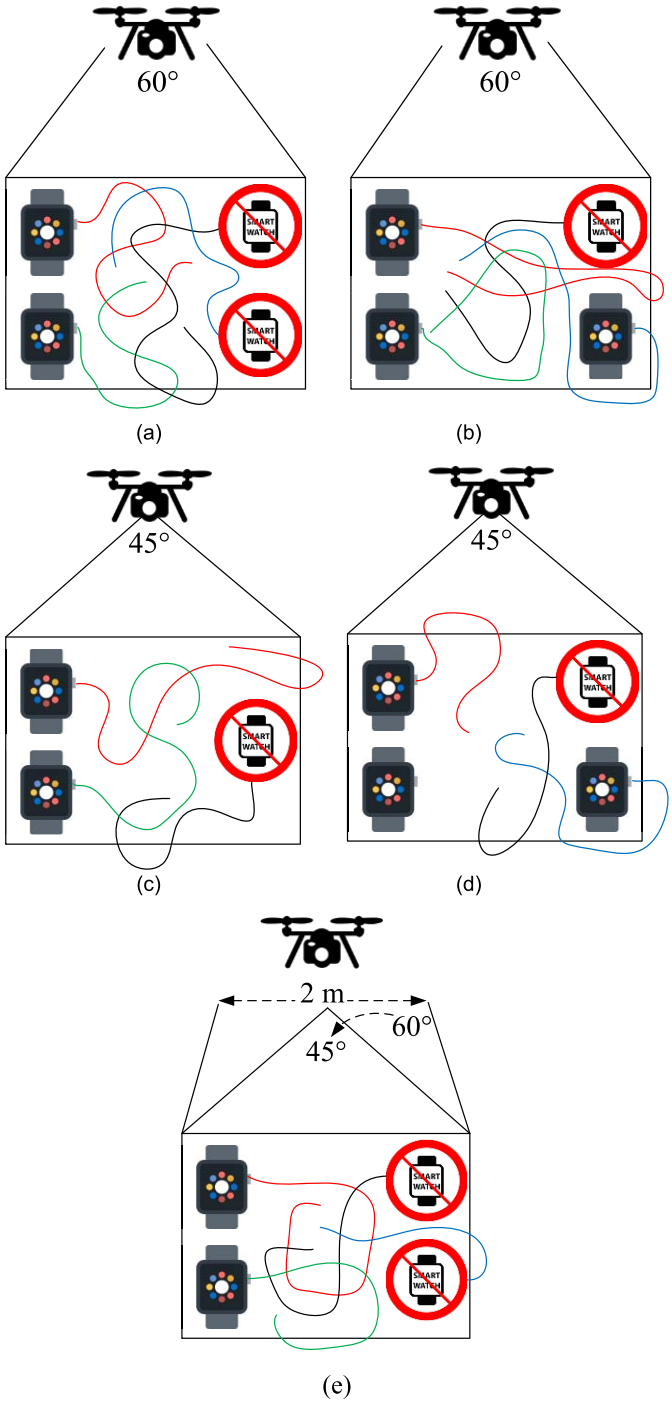


Fig. 9. (a) Case 1. (b) Case 2. (c) Case 3. (d) Case 4. (e) Case 5.

TABLE VI
EVALUATION AND COMPARISON RESULTS AMONG FIVE CASES

Method	MOTA (%)	Personal Profile
Case 1, YOLO+SORT	66.36	No
Case 1, YOLO+DeepSORT	71.36	No
Case 1, Config. 1 (Our)	87.87	Yes
Case 1, Config. 2 (Our)	93.71	Yes
Case 2, YOLO+SORT	50.25	No
Case 2, YOLO+DeepSORT	51.21	No
Case 2, Config. 1	81.96	Yes
Case 2, Config. 2	91.00	Yes
Case 3, YOLO+SORT	71.38	No
Case 3, YOLO+DeepSORT	71.9	No
Case 3, Config. 1	93.51	Yes
Case 3, Config. 2	95.36	Yes
Case 4, YOLO+SORT	27.45	No
Case 4, YOLO+DeepSORT	54.65	No
Case 4, Config. 1	87.77	Yes
Case 4, Config. 2	93.85	Yes
Case 5, YOLO+SORT	50.48	No
Case 5, YOLO+DeepSORT	50.75	No
Case 5, Config. 1	84.77	Yes
Case 5, Config. 2	89.74	Yes

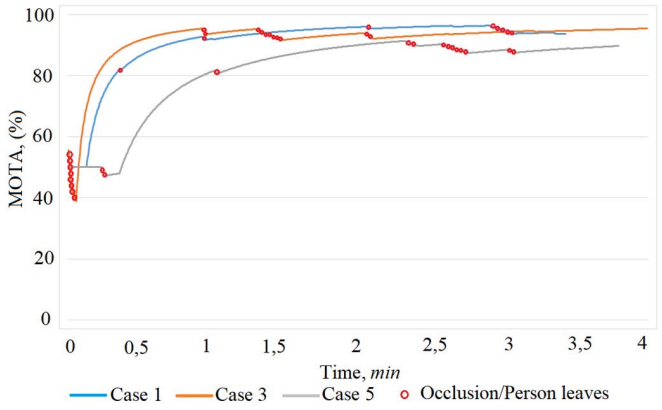


Fig. 10. MOTA versus experiment running time.

system performance may slightly drop due to smartwatch ID switches. The proposed system can quickly recover afterward. The proposed system allows not only to track a person, but also to implicitly identify an unauthorized person who does not wear a smartwatch, when people perform different actions. In order to solve the same actions issue, the different postures for different people are requested before they may have the same action. Once the proposed system pairs the best match, the system could continuously recognize and track the different people who do the same action without occlusion in the remaining time. Fig. 11 shows the real testing photos to prove the proposed person-tracking framework can tackle the ID switch problem and exclude the assisted ground markers.

Since it is not easy to completely test the number of people under all different environment settings, instead, we design a data fusion simulator to consider the following factors: 1) time interval (5 min); 2) number of people (from 4 to 10); and 3) number of action features (4 denotes stand, walk, jump, fall down). However, this simulator does not consider the factors,

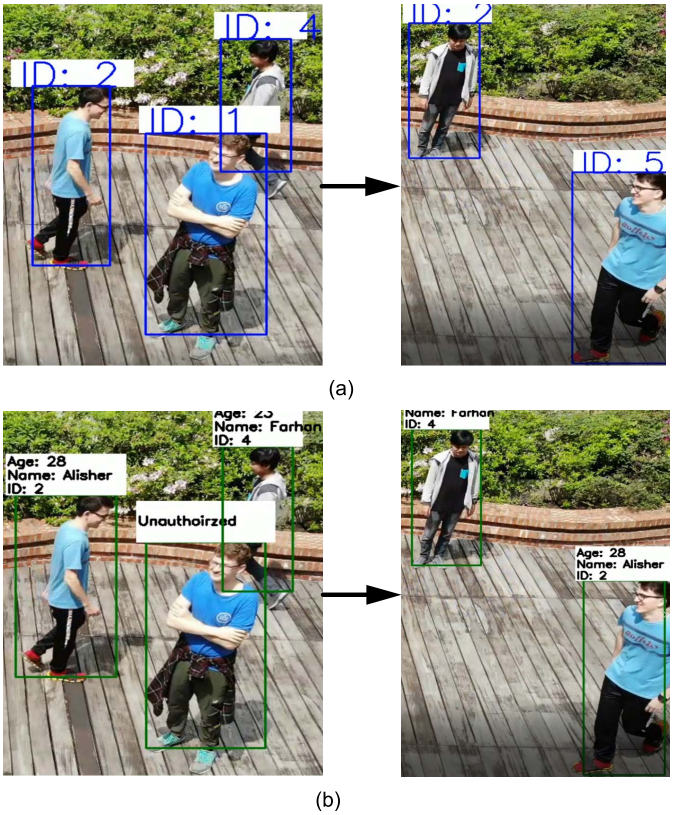


Fig. 11. Real testing results for Case 3: (a) YOLO + SORT and (b) Config. 2.

TABLE VII
MOTA ANALYSIS CONSIDERING NOISE AND NUMBER OF PERSONS

Noise Level / # of Persons	0%	10%	20%
4	97.33	95.58	86
5	95.67	91.07	76.67
6	95.5	89.78	74.06
7	95.57	92.1	69.67
8	94.42	87.21	71.54
9	93.37	86.67	61.11
10	94.53	84.13	51.63

including human size, occlusion, and without smartwatch. Herein, the data fusion simulator shows MOTA analysis under different noise levels and number of persons in Table VII, where the noise level is defined as the mismatch percentage in the sequences of video and sensor data. Obviously, more persons and larger noise degrade the MOTA performance.

Table VIII shows a qualitative and quantitative comparison among different person tracking systems. In terms of qualitative comparison, compared with our previous works [6], [7], this work can exclude the assisted ground markers. Compared with [42] and [22], our work uses the drone to dynamically and continuously track the people through fusion of drone video and sensor data. Although the work [18] has a drone view, our proposed system allows multiple people tracking even with the existence of occlusions. The work [21] considers the features from static camera video and wearable sensors for multiple persons tracking and identification. In terms of

TABLE VIII
PERSON TRACKING METHODS COMPARISON

Method	Drone View	IoT Wearable Devices	Dataset	# of Param.	Marker Assisted
YOLOv3 + SORT + Cellphones [6-7]	Yes	Yes	Own Dataset	65.25 M*1	Yes
VGG16 + DeepSORT [42]	No	No	MOT16	141.16 M*2	No
YOLOv3 + MTCNN [18]	Yes	No	Own Dataset	> 65.25 M	No
Videos + IMUs Sensors [21]	No	Yes	Own Dataset	-	No
Joint Graphical Model [22]	No	No	FBMS59	-	No
YOLOv3 + SORT + OpenPose + Smartwatch (Our)	Yes	Yes	Own Dataset	117.56 M*3	No
YOLOv3 + DeepSORT + OpenPose + Smartwatch (Our)	Yes	Yes	Own Dataset	120.36 M*4	No

*1: [44]
*2: 138.36M+2.8M= 141.16M
*3: 65.25 M + 52.31 M = 117.56 M
*4: 65.25 M + 2.8 M + 52.31 M = 120.36 M

quantitative comparison, we add the number of parameters. Although our configurations have more parameters than most methods [6], [7], [18] in Table VIII, this work can reach: 1) use the posture to avoid using the assisted ground markers; 2) propose a person tracking framework by fusion of posture data from drone views and wearable sensors; 3) tag smartwatch IDs and personal profiles including name and age information on the human objects of the drone view to bypass the ID switch problem; and 4) modularize the proposed framework to easily allow the change of tools if needed.

V. CONCLUSION

In this article, we use the posture-based multisensor data fusion to avoid using the assisted ground markers and to bypass the ID switch problem for the proposed person tracking framework. This framework is capable of tracking people through the fusion of the posture data from UAV video and smartwatches and tagging smartwatch IDs and personal profiles to drone videos. In video data processing flow, the combination of person detection, ID assignment, pose estimation, and MLP is developed for posture recognition. In smartwatch data processing flow, another combination of Bi-LSTM and MLP is developed. A fusion module is developed to integrate the posture information from two data processing flows by DTW for person tracking. Through the proposed framework, we explore two configurations. Our result can greatly improve the visualization effect of UAV videos. In the near future, we will consider online processing for the multiple persons tracking.

REFERENCES

[1] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1302–1310.
[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.
[4] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, Venice, Italy, 2017, pp. 2961–2969.
[5] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 1–9.
[6] L.-D. Van, L.-Y. Zhang, C.-H. Chang, K.-L. Tong, K.-R. Wu, and Y.-C. Tseng, "Things in the air: Tagging wearable IoT information on drone videos," *Discover Internet Things*, vol. 1, no. 1, pp. 1–13, Feb. 2021.
[7] L.-D. Van, C.-H. Chang, K.-L. Tong, K.-R. Wu, L.-Y. Zhang, and Y.-C. Tseng, "Demo: Tagging IoT data in a drone view," in *Proc. 25th Annu. Int. Conf. Mobile Comput. Netw.*, Los Cabos, Mexico, Oct. 2019, pp. 1–3.
[8] R. Y.-C. Tsai, H. T.-Y. Ke, K. C.-J. Lin, and Y.-C. Tseng, "Enabling identity-aware tracking via fusion of visual and inertial features," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 2260–2266.
[9] 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. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 2021–2026.
[10] S. Majumder and N. Kehtarnavaz, "Vision and inertial sensing fusion for human action recognition: A review," *IEEE Sensors J.*, vol. 21, no. 3, pp. 2454–2467, Feb. 2021.
[11] T. Nägeli, S. Oberholzer, S. Plüss, J. Alonso-Mora, and O. Hilliges, "Flycon: Real-time environment-independent multi-view human pose estimation with aerial vehicles," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1–14, Dec. 2018.
[12] M. Tzelepi and A. Tefas, "Graph embedded convolutional neural networks in human crowd detection for drone flight safety," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 2, pp. 191–204, Apr. 2021.
[13] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision meets drones: A challenge," 2018, *arXiv:1804.07437*.
[14] C. Liu and T. Szirányi, "Real-time human detection and gesture recognition for on-board UAV rescue," *Sensors*, vol. 21, no. 6, p. 2180, 2021.
[15] S. Sharma, A. V. Venkata Susmitha, L.-D. Van, and Y.-C. Tseng, "An edge-controlled outdoor autonomous UAV for colorwise safety helmet detection and counting of workers in construction sites," in *Proc. IEEE 94th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2021, pp. 1–5.
[16] W. Sultani and M. Shah, "Human action recognition in drone videos using a few aerial training examples," 2019, *arXiv:1910.10027*.
[17] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, "UAV-Human: A large benchmark for human behavior understanding with unmanned aerial vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 16266–16275.
[18] Q. Shen, L. Jiang, and H. Xiong, "Person tracking and frontal face capture with UAV," in *Proc. IEEE 18th Int. Conf. Commun. Technol. (ICCT)*, Oct. 2018, pp. 1412–1416.
[19] J. Portmann, S. Lynen, M. Chli, and R. Siegwart, "People detection and tracking from aerial thermal views," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 1794–1800.
[20] H. Fradi, L. Bracco, F. Canino, and J.-L. Dugelay, "Autonomous person detection and tracking framework using unmanned aerial vehicles (UAVs)," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 1047–1051.
[21] R. Henschel, T. von Marcard, and B. Rosenhahn, "Simultaneous identification and tracking of multiple people using video and IMUs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Long Beach, CA, USA, Jun. 2019, pp. 780–789.
[22] M. Keuper, S. Tang, Y. Zhongjie, B. Andres, T. Brox, and B. Schiele, "A multi-cut formulation for joint segmentation and tracking of multiple objects," 2016, *arXiv:1607.06317*.
[23] F. Liu, D. Zhang, and L. Shen, "Study on novel curvature features for 3D fingerprint recognition," *Neurocomputing*, vol. 168, no. 1, pp. 599–608, 2015.
[24] J. Chen, F. Shen, D. Z. Chen, and P. J. Flynn, "Iris recognition based on human-interpretable features," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 7, pp. 1476–1485, Jul. 2016.
[25] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*. Boston, MA, USA: Springer, 2008.

[26] W. Zhu et al., "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *Proc. Assoc. Advancement Artif. Intell.*, 2016, pp. 3697–3703.
[27] T. S. Kim and A. Reiter, "Interpretable 3D human action analysis with temporal convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1623–1631.
[28] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, Aug. 2017.
[29] M. Cornacchia, K. Ozcan, Y. Zheng, and S. Velipasalar, "A survey on activity detection and classification using wearable sensors," *IEEE Sensors J.*, vol. 17, no. 2, pp. 386–403, Jan. 2017.
[30] A. H. Ngu, P. T. Tseng, M. Paliwal, C. Carpenter, and W. Stipe, "Smartwatch-based IoT fall detection application," *Open J. Internet Things (OJIOT)*, vol. 4, no. 1, pp. 87–98, 2018.
[31] Z. Wang, D. Wu, R. Gravina, G. Fortino, Y. Jiang, and K. Tang, "Kernel fusion based extreme learning machine for cross-location activity recognition," *Inf. Fusion*, vol. 37, pp. 1–9, Sep. 2017.
[32] N. Y. Hammerla, S. Halloran, and T. Ploetz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," 2016, *arXiv:1604.08880*.
[33] Z. Wang, M. Jiang, Y. Hu, and H. Li, "An incremental learning method based on probabilistic neural networks and adjustable fuzzy clustering for human activity recognition by using wearable sensors," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 4, pp. 691–699, Jul. 2012.
[34] *Multi-Person Real-Time Recognition Repository*. [Online]. Available: https://github.com/felixchenfy/Realtime-Action-Recognition
[35] D. R. Amancio, C. H. Comin, D. Casanova, G. Travieso, O. M. Bruno, and F. A. Rodrigues, "A systematic comparison of supervised classifiers," *PLoS ONE*, vol. 9, no. 4, pp. 1–14, Apr. 2014.
[36] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[38] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proc. KDD Workshop*, 1994, pp. 359–370.
[39] K. Bernardin, A. Elbs, and R. Stiefelhagen, "Multiple object tracking performance metrics and evaluation in a smart room environment," in *Proc. 6th IEEE Int. Workshop Vis. Surveill.*, 2006, pp. 1–8.
[40] *Sensor Manager Documentation*. [Online]. Available: https://developer.android.com/reference/android/hardware/SensorManager
[41] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3464–3468.
[42] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3645–3649.
[43] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.
[44] *Wolfram Neural Net Repository*. [Online]. Available: https://resources.wolframcloud.com/NeuralNetRepository/resources/YOLO-V3-Trained-on-Open-Images-Data/

Alisher Mukashev received the bachelor's degree in electrical engineering and the master's degree in information and communication technologies and systems from the Tomsk State University of Control Systems and Radioelectronics, Tomsk, Russia, in 2015 and 2017, respectively, and the bachelor's degree in English language from the National Research Tomsk Polytechnic University, Tomsk, in 2015.

He is currently pursuing the Ph.D. degree in computer science from the National Yang Ming Chiao Tung University, Hsinchu, Taiwan. He has lead Industrial projects with CTCI and YungTay during his Ph.D. study. His research interests include artificial intelligence, machine learning, deep learning, and computer vision.

Mr. Mukashev won the Vladimir Potanin Scholarship in 2017.

Lan-Da Van (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the National Taiwan University (NTU), Taipei, Taiwan, in 2001.

Since February 2006, he joined the faculty of the Department of Computer Science, National Yang Ming Chiao Tung University (NYCU), Hsinchu, Taiwan, where he is currently a Professor. He serves the Associate Chief Director of MIRC, NYCU. His research interests are intelligent/VLSI algorithms, architectures, chips, systems, and the applications for digital signal processing and adaptive/machine leaning computation. This includes the design of low-power/high-performance/cost-effective adaptive filter, computer arithmetic, independent component analysis (ICA), multi-dimensional filter, transform, 3-D graphics systems, intelligent elevator systems, and UAV and wearable data fusion systems.

Dr. Van was a recipient of the Best Poster Award in the iNEER Conference for Engineering Education and Research (iCEER) in 2005. In 2014, he received the Best Paper Award in the IEEE International Conference on Internet of Things (iThings2014). He was also a recipient of the Teaching Award of the Computer Science College, National Chiao Tung University, in 2014. He served as the Chairperson for the IEEE NTU Student Branch in 2000. In 2001, he has received the IEEE Award for outstanding leadership and service to the IEEE NTU Student Branch. From 2009 to 2010, he served as an Officer for the IEEE Taipei Section. He was the Special Session Co-Chair of the 2018 IEEE International Conference on Digital Signal Processing (DSP 2018). In 2019, he served as the Publicity Co-Chair for the IEEE International System-on-Chip Conference (SOCC 2019). In 2020, he served as the Tutorial Co-Chair for the IEEE International System-on-Chip Conference (SOCC 2020) and the TPC Co-Chair for the IEEE Asia Pacific Conference on Circuits and Systems (APCCAS 2020). In 2021, he served as the Special Session Co-Chair for the 2021 IEEE International Symposium on Circuits and Systems (ISCAS 2021), the Tutorial Co-Chair of the IEEE International System-on-Chip Conference (SOCC 2021), the Special Session Co-Chair for the International SoC Design Conference (ISOCC 2021), and an International Steering Committee Member of the IEEE Asia Pacific Conference on Circuits and Systems (APCCAS 2021). In 2022, he serves as the Special Session Co-Chair for the 2022 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS 2022), the Technical Program Committee Co-Chair for the International SoC Design Conference (ISOCC 2022), and the Program Co-Chair for the IEEE International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC 2022). He serves as the Chair-Elect/Secretary of the IEEE CASS VLSI Systems and Applications Technical Committee (VSA-TC) since 2022. He served as an Associate Editor for the IEEE ACCESS from 2018 to 2022, IEEE TRANSACTIONS ON COMPUTERS from 2014 to 2018, and has been serving as an Associate Editor for the IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING since 2022 and *ACM Computing Surveys* since 2020.

Susanta Sharma received the B.Tech. degree in electronics and communication engineering from the IERCEM Institute of Information Technology, Techno India Group, Habra, India, in 2012, and the master's degree in electronics engineering from Chang Gung University, Taoyuan, Taiwan, in 2017, Taiwan. He is currently pursuing the Ph.D. degree in computer science from the National Chiao Tung University, Hsinchu, Taiwan.

From 2013 to 2015, he has worked with the Indian Statistical Institute, Kolkata, India, as a Project Fellow, where he worked in SoDAR for Environment monitoring. He has gone through several real-time project experiences with the most advanced technologies in this present world, which includes AI, IoT, AR/VR (Microsoft HoloLens, HTC VIVE), robotics, deep learning, biomedical signal processing (like ECG, EEG, and EMG), UAVs, and so on. He has led industrial projects, like CTCI, MOST, and NCSIST, Taiwan, during his Ph.D. study at NYCU, Taiwan. He was also a Regular Recipient of the International Student Scholarship and a Tuition Fee Waiver during his master's and Ph.D. study till date in Taiwan.

M. Farhan Tandia received the B.Eng. degree in electrical engineering from Universitas Gadjah Mada, Yogyakarta, Indonesia, in 2018. He is currently pursuing the M.S. degree in electrical engineering and computer science with the National Yang Ming Chiao Tung University, Hsinchu, Taiwan.

He joined research team at the Nature Intelligence Laboratory that focus on applied computer vision, deep learning, machine learning, and UAV integration.

Yu-Chee Tseng (Fellow, IEEE) received the Ph.D. degree in computer and information science from The Ohio State University, Columbus, OH, USA, in January 1994.

He was/is Chairperson, from 2005 to 2009, and the Dean, since 2011, with the College of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan. His research interests include mobile computing, wireless communication, and the Internet of Things. His H-index is more than 50.

Dr. Tseng has been awarded as an NCTU Chair Professor since 2011 and the Y. Z. Hsu Scientific Chair Professor from 2012 to 2013. He received the Outstanding Research Award from the National Science Council in 2001, 2003, and 2009; the Best Paper Award from the International Conference on Parallel Processing in 2003; the Elite I. T. Award in 2004; and the Distinguished Alumnus Award from The Ohio State University in 2005; and the Y. Z. Hsu Scientific Paper Award in 2009. He served/serves on the editorial boards for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, and IEEE INTERNET OF THINGS JOURNAL.