18th International Conference on Information Fusion
Washington, DC - July 6-9, 2015

Combining Passive Visual Cameras and Active IMU Sensors to Track Cooperative People

Wenchao Jiang
Missouri University of Science and Technology, MO, USA, 65401
Email: wjm84@mst.edu

Zhaozheng Yin
Missouri University of Science and Technology, MO, USA, 65401
Email: yinz@mst.edu

***Abstract*—We attack the problem of persistently tracking cooperative people such as children, the elderly or patients by combining passive tracking and active tracking techniques. Passive tracking uses visual signals from surveillance cameras, but vision based people tracking becomes a hard problem in challenging scenarios such as long-term/heavy occlusion, people changing their movement patterns during occlusion, or people temporarily moving out of the visual field. Active tracking uses sensor signals from Inertial Measurement Unit (IMU) carried by targets themselves. IMU-based tracking is independent of visual signals, so it keeps working when people are visually occluded and offers clues where the target could be, helping the visual tracking to reidentify the target. Meanwhile, when visual signals on people are available, visual tracking can calibrate IMU-based tracking to avoid sensor drift. The experimental results show that the IMU and visual tracking are complementary to each other and their combination performs robustly on tracking cooperative people in many challenging scenarios.**

I. Introduction

A. Problem

People tracking has a wide range of applications such as tracking people in public crowded environments for security surveillance and tracking family members to avoid losing loved ones. Typically, people tracking techniques can be classified as "passive" or "active" tracking. *Passive tracking* utilizes devices that are not carried by people, such as surveillance cameras. *Active tracking* locates targets by sensors carried by the *cooperative* targets themselves, such as the Global Positioning System (GPS), WiFi receiver and Inertial Measurement Unit (IMU). This paper attacks the problem of *persistently* tracking cooperative targets (e.g., children, teens, the elderly, patients with autism/alzheimers/dementia) by combining passive and active tracking.

B. Related Work

1) Passive Visual Tracking: Passive vision-based people detection and tracking have been studied for several decades [2][3]. The challenges are to track people persistently through occlusion or clutter. For partial occlusion, Wu and Nevatia [14] represented humans as an assemble of four body parts and combined body part detectors and human detectors to track humans when they were occluded. Papadakis et al. [10] formed the representation of visible and occluded parts and segmented the two parts by graph cuts. Tang et al. [12] trained an occlusion-aware person detector, which was a joint model of detecting single person as well as pairs of persons under varying degrees of occlusion. For full occlusion, previous efforts focused on predicting targets' positions when they are occluded and this was realized by Kalman filter [16], [9], [13] or assuming the target keeps a constant velocity [15], [5].

Fig. 1. Visual people tracking and its challenges. (a) Successful tracking; (b) The target is occluded by other people; (c) The target is occluded by a tree over a long period; (d) The target moves out of the field of view temporarily.

When there is heavy occlusion, large appearance change, nearby clutter or pedestrians temporarily moving out of the field of view, as shown in Fig.1, it is challenging for a merely vision-based tracking algorithm to persistently track people without failure. This problem becomes worse when people change their moving patterns (e.g., speed, direction) when occluded, which voids the linear filtering based prediction approaches.

2) Active Sensor Tracking: It is intuitive to track people with GPS considering its wide application in vehicle navigation. However, the accuracy of a common GPS module is not high enough (15 meters on average as reported in [1]). Furthermore, obstructions such as city canyons or tall trees outdoors and walls/ceilings indoors weaken the signals transmitted between GPS receivers and positioning satellites, making the GPS-based tracking unreliable in these GPS-denied environments. WiFi is another choice to locate people but the coverage area of most WiFi hotspots is less than 50

meters, limiting its application in people tracking outdoors.

Inertial Measurement Unit (IMU), consisting of gyroscope, magnetometer and accelerometer, is a good choice to track people by Dead Reckoning (DR) which adds the current displacement vector to the previous estimated location. DR is built upon three components: step detection, speed estimation and forward moving direction determination. Previously, steps are detected by setting a threshold on the value of acceleration [11], [6], but the threshold depends on a person's movement patterns such as running and fast/slow walking. Speed equals to the product of a predefined calibrated coefficient and the amplitude of acceleration [11], [6], [7], [4], but the calibration coefficient is hard-coded and person-dependant. The orientation of acceleration is used to determine the forward moving direction [11], [6], [8], but finding the accurate transformation from sensor movement directions on human body (e.g., in a pocket) to the walking/running direction of a person in the world coordinate is difficult. Furthermore, the DR-based approach is prone to drift if small errors on each step are accumulated over a long period.

3) Vision and IMU Fusion: Previous work has explored the possibility to fuse vision and IMU for people tracking or navigation [20][18][19][17]. These work usually fixed a camera on the target's body and utilized vision information for motion estimation. By involving the motion information from vision and Dead Reckoning result from IMU into the Kalman filter framework, a more accurate tracking results can be obtained. These work is single-direction fusion, that is only vision can aid IMU for people tracking. Our work sets up the stationary surveillance camera out of the target's body and investigated how IMU and vision tracking can assist each other and form a persistent people tracking system.

C. Motivation

Visual tracking can obtain the movement trajectories/patterns of people, thus it can calibrate IMU-based active tracking to avoid sensor drift. It is challenging for visual tracking to handle heavy occlusion, but active people tracking methods have no problems of occlusion because they do not rely on visual signals, thus the occlusion problem of visual tracking can be compensated by active sensor tracking.

D. Proposal

Since visual tracking and IMU-based active tracking are complementary, i.e., not only can IMU assist visual tracking when the target is occluded, but also the challenges of IMU tracking (calibration and drift) are alleviated when visual signals are available, we propose a novel people tracking system combining passive visual tracking and active sensor tracking. The visual signal is from stationary surveillance cameras and IMU devices from cooperative people are used for active tracking.

E. System Overview

Our cooperative tracking system consists of three parts:

(1) Passive Visual Tracking (Section 2): Given a stationary surveillance camera, a scene-specific pedestrian detector is trained to improve the detection performance. An adaptive scale selection algorithm is proposed to further improve the pedestrian detection performance and reduce computational cost. Mode-seeking algorithm is applied to the detection confidence map for people tracking.

(2) Active IMU Tracking (Section 3): A Discrete Fourier Transform (DFT)-based step detection method is proposed, which does not need preset person-dependant thresholds. The calibration coefficient in speed estimation is obtained by visual tracking instead of manual setting. More accurate forward moving direction is obtained by a principle frequency component filter.

(3) Integration of Visual and IMU Tracking (Section 4): When the target is visible, its visual trajectory calibrates and adjusts its IMU trajectory. When the target is occluded, IMU tracking keeps working and offers clues for visual tracking to re-identify the missed target.

II. Passive Visual Tracking

In this section, a scene-specific and scale-adaptive pedestrian detector is firstly introduced, then visual people tracking based on detections is described.

A. Training a Scene-specific Pedestrian Detector

Histogram of Oriented Gradient (HOG) feature along with Support Vector Machine (SVM) has been widely used to perform pedestrian detection in images. A large dataset (both positive samples and negative samples) are usually needed to train a general pedestrian detector which is very time-consuming and the detector may not work well on scenarios different from the training dataset [2].

In a fixed scene, the viewpoints from which people can be observed and the scales of people in images are limited. Moreover, the negative samples are limited (they are just the background in the scene!). The critical problem in people detection is how to classify those background samples that are very likely to be mistakenly classified as people samples. If the detector can correctly classify those background samples whose feature vectors are near the decision boundary, it is sufficient to classify other background samples which are largely different from people samples. In this paper, we propose a new iterative training algorithm to deal with the problem.

As illustrated in Fig.2, the positive samples (images framed in red) are the manually cropped pedestrian images from videos taken on the specific scene, which include pedestrian images with different walking gaits and scales that can be seen from the specific viewpoint. The pool of positive samples is not changed during the iterative training. The negative sample pool initially consists of randomly cropped backgrounds from images taken on the specific scene in different weather and illumination conditions (images framed in blue). The negative sample pool expands gradually during the iterative training.

The iterative training algorithm is performed in the following steps: when a new pedestrian detector is available after SVM training, it will be applied to classify background images

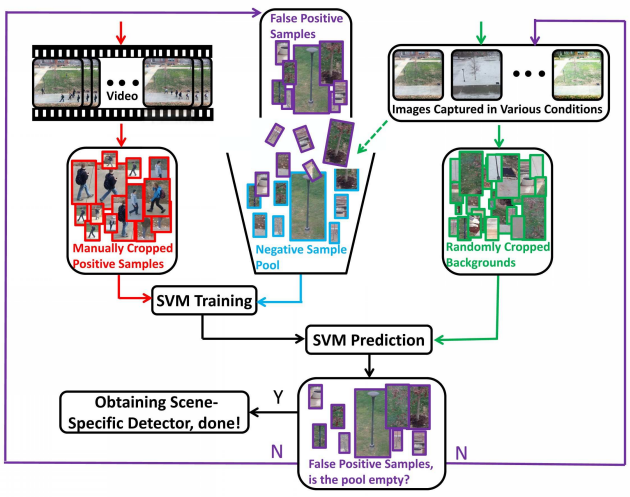

Fig. 2. Training a scene-specific pedestrian detector.

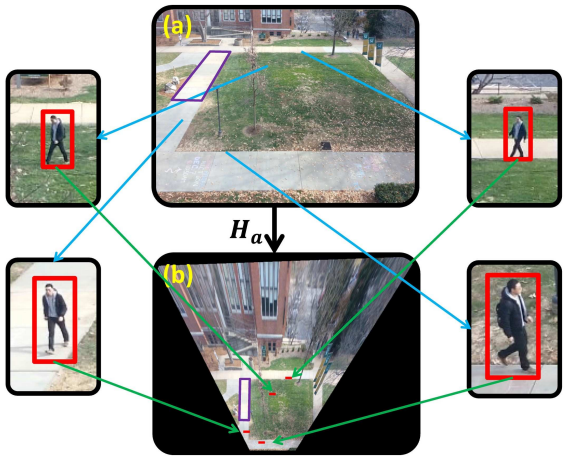

Fig. 3. Adaptive scale selection.

randomly cropped from images taken on the specific scene (images framed in green); the false positive samples (images framed in purple) are put into the negative sample pool and the SVM will be updated for the next iteration; training stops when the number of false positive samples is zero. Every time the SVM is updated, the detector is more robust to classify those samples that are misclassified previously.

B. Adaptive Scale Selection

In common people detection algorithms, for every input frame, different scales defined by the height and width of rectangles need to be searched in the image exhaustively to detect all pedestrians. In a fixed scene, although the same person may display different scales at different locations in the image (e.g., Fig.3(a)). However, if we transform the image into the top-down view by a homography matrix $\mathbf{H}_a$, the width of the pedestrian rectangle is almost a constant (red lines in Fig.3(b) are the warped rectangular bottoms from four detections). Thus, if we fix the ratio of the height and width of a detection rectangle and determine the standard scale S_{std} by the length of the bottom side of the warped rectangle, people's scales in every region of the specific scene can be estimated in advance, i.e., we know which scale in the original image we should use to detect pedestrians rather than performing the exhaustive scale search.

$\mathbf{H}_a$ is estimated by four pairs of point correspondences (e.g., the four corners of the purple rectangle in Fig.3). $\mathbf{H}_a$ is a constant for a fixed scene and it only needs to be updated when the viewpoint changes.

C. Tracking by Detection

Based on the target's location in the previous frame $t-1$, we apply our scene-specific and scale-adaptive pedestrian detector within a local region around the previous location to detect the target in the current frame t. Fig.4 shows some pedestrian images and their confidence maps corresponding to SVM scores of the pedestrian detector. The white in a confidence map denotes high score (confidence) of people detection. The target's location in frame t is determined by seeking the mode (the position with maximal confidence in the confidence map).

Fig. 4. Visual tracking. (a) and (c) are the pedestrian images. (b) and (d) are their confidence maps, respectively.

If the target is not occluded (Fig.4(a)), there is a single global peak in the confidence map, thus the target can be correctly tracked. However, when the target is occluded by other pedestrians (e.g., Fig.4(c)), there are multiple peaks in the corresponding confidence map. It is possible that the non-target pedestrian is detected and tracked mistakenly. Therefore, when occlusion, clutter and disappearance of the target happen, we refer to IMU-based active tracking to correct the visual tracker and reidentify the lost target.

III. Active IMU-based Tracking

IMU includes accelerometer, magnetometer and gyroscope, which measures tri-axis acceleration, the strength of magnetic field and tri-axis angular velocities, respectively. As shown in Fig.5, our IMU tracking is based on Dead-Reckoning (DR) which adds the displacement vector, $v_n \frac{\mathbf{u}_n}{\|\mathbf{u}_n\|}$, to the previously estimated location $\mathbf{p}_n$. v_n and $\mathbf{u}_n$ are the speed and forward moving direction in step n, respectively. Our IMU tracking approach consists of three components: step detection, speed estimation and forward moving direction determination.

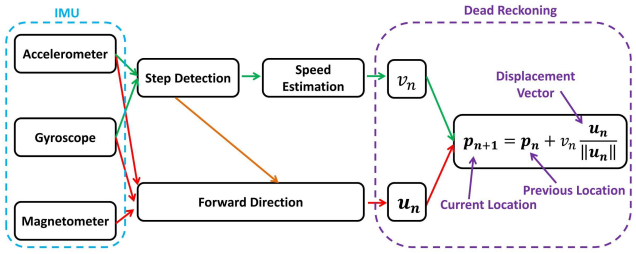

Fig. 5. Flow chart of our IMU-based pedestrian tracking.

A. Step Detection

Speed and heading direction require to be estimated on a complete step period for DR, so the accurate beginning and end of each step is needed. In [11], step is detected in the time domain by finding local maximum and minimum of acceleration data and a threshold is set to rule out false positives. However, the threshold depends on the speed and is person-specific. When speed greatly changes, missed detection of steps increases rapidly. Different thresholds need to be chosen for different targets.

In this paper, a step detection algorithm based on adaptive sliding window and Discrete Fourier Transform (DFT) is introduced, which is inspired by the following observations: (1) The movement pattern of a walking person is periodic. Therefore, DFT can be applied to find the number of periods (i.e., the number of steps) in a certain sliding window. (2) Magnetic field is sensitive to heading direction change, so it is not suitable for step detection. Instead, angular velocity and acceleration are ideal because they do not depend on the forward direction. (3) Only one axis signal is not reliable for step detection. All 6 axes of gyroscope and accelerometer are considered in our step detection by DFT.

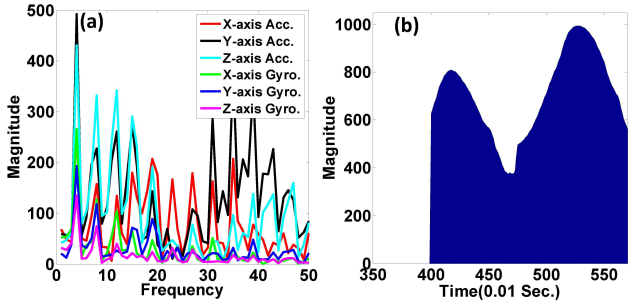

Fig. 6. Step detection. (a) DFT results of 4-second signal with 400 samples. (b) The variance metric vs. signal length to detect the accurate step period.

Fig.6(a) shows the results after applying DFT to six signals of accelerometer and gyroscope over a time sliding window L of 400 samples. In our IMU device, sensor data are collected at the rate of 100 samples per second, so 400 samples of data implies data collected in 4 seconds. The horizontal axis in Fig.6(a) is the frequency which is related to the number of periods within the time sliding window. The vertical axis in Fig.6(a) is the corresponding magnitude. The frequency component related to the step number should have high magnitude while frequency components corresponding to noise should have low magnitude. We compute the principle frequency of all six signals, f^*, by

$$f^* = \arg\max_f \sum_{i=1}^{6} |F_i(f;L)| \quad (1)$$

where $|F_i(f;L)|$ denotes the magnitude of frequency component f of the ith signal within the time sliding window L. Note that f^* is an integer in DFT. In Fig.6(a), $f^* = 4$, but is there exactly 4 steps during this time sliding window (400 samples)? The answer is possibly NO. If there are 3.8 or 4.3 steps in the sliding window L, the corresponding principle frequency will be rounded to 4. We need to search the accurate beginning and end sampling moments of complete steps in the signals to estimate the speed for Dead-Reckoning (DR). Otherwise, DR will deviate from the truth quickly due to the accumulated error. Observing that the principle frequency has a large difference compared to its neighboring frequencies, we propose a new metric M_L, the magnitude variance of the principle frequency compared with its neighboring frequencies, to search the accurate steps:

$$M_L = \sum_{i=1}^{6} var([|F_i(f^*\text{-}1;L)|, |F_i(f^*;L)|, |F_i(f^*\text{+}1;L)|]) \quad (2)$$

We gradually increase the time sliding window L. For each L, we compute f^* by Eq.1 and then compute M_L by Eq.2. Fig.6(b) shows the plot of M_L versus L. We can see the first peak is around 420 with $L = [1, 420]$, which means that there are 4 steps in 420 samples (4.2 seconds), i.e., each step period is about 105 samples. If we keep increasing L, we will find another peak around 525 in $L = [1, 525]$ which means that there are 5 steps in 525 samples. The peaks in Fig.6(b) indicate that at these points, the magnitude of the principle frequency has the largest difference compared to its neighboring frequencies. Thus, we can detect the exact number of steps by adapting this time sliding window technique.

B. Speed Estimation

Practically, walking/running speed varies from person to person. Even for the same person, the moving speed may not be a constant over time. Integration on acceleration to obtain speed accumulates errors very fast, making it impractical for speed estimation. Observing that the magnitude of movement is approximately proportional to speed, we propose to use the maximal difference of angular velocity to measure the movement intensity. The measurement is only valid in complete movement pattern periods, which is at least one step. That is one of the reasons why we need accurate step detection and speed is calculated in the unit of step. The speed for step n is defined as

$$v_n = \alpha(max_{s\in[s_b^{(n)},\ s_e^{(n)}]} \| \mathbf{a}_s \| - min_{s\in[s_b^{(n)},\ s_e^{(n)}]} \| \mathbf{a}_s \|) \quad (3)$$

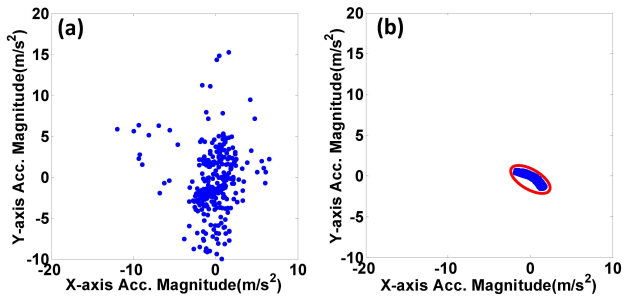

Fig. 7. Determine the forward moving direction. (a) The acceleration in a short period (4 steps) is projected to the horizontal plane in the world coordinate. (b) The acceleration corresponding to the principle frequency in a short period is projected. The semi-major axis of the ellipse represents the forward moving direction.

where $s_b^{(n)}$ and $s_e^{(n)}$ denote the beginning and end sample of the nth step and $\| \mathbf{a}_s \|$ is the magnitude of angular velocity at sample s. α is the calibration factor depending on specific persons. When visual signal is used, α is determined by a similarity warping matrix which will be introduced in Section 4.

C. Forward Moving Direction Determination

The 3D acceleration vectors in the IMU coordinate during each step can be projected to the horizontal plane in the world coordinate to infer the forward moving direction [11], [6]. This method works for professional IMUs. But for low cost IMUs such as the IMU module built in smatphones which is more likely to be influenced by noise, it performs poorly. Fig.7 shows the results when acceleration collected by a smartphone in 4 steps is projected to the world coordinate's horizontal plane. There is no obvious forward moving direction.

Considering the 3D acceleration vectors during a time sliding window as time-series signals, we transform them into the frequency domain. Since the principle frequency during the sliding window is already detected in the step detection process (Section 3.1), we treat all non-principle frequency components as noise and zero them out. Then, the filtered signal is transformed back to the time domain and is projected to the world coordinate's horizontal plane. As shown in Fig.7(b), the moving direction is obvious. Ellipse-fitting (i.e., 2D Principle Component Analysis) is applied to the projected principle acceleration and the semi-major axis of the ellipse represents the forward moving direction $[u_x \ u_y]$.

IV. Integration of Visual and IMU Tracking

As shown in Fig.8, our cooperative people tracking system is divided into three parts: initialization, tracking and re-identification.

A. Initialization

The target to be tracked is initially identified by human. Fig.8(a) and Fig.8(b) show the visual trajectory (red) and IMU trajectory (green) in the first sliding window L_1, respectively. The trajectory generated by IMU tracking is in the world coordinate, so it is a 2D curve in the horizontal plane viewed from top to down. Unlike IMU trajectory, visual trajectory is in the image coordinate depending on the specific camera viewpoint, thus they are not directly comparable. We warp the visual trajectory from scene-specific viewpoint to the top-down viewpoint by $\mathbf{H}_a$ (Section 2.2), as shown in Fig.8(c). Since the transformation between the warped visual trajectory (Fig.8(c)) and IMU trajectory (Fig.8(b)) is just rotation, translation and scaling (i.e., similarity transformation), we match the two trajectory curves by computing the similarity transformation matrix $\mathbf{H}_{s,k}$ in sliding window L_k using the least square procedure:

$$\arg\min_{\mathbf{H}_{s,k}} \sum_t (\mathbf{H}_{s,k}\mathbf{T}_t^{(v,k)} - \mathbf{T}_t^{(s,k)})^2 \quad (4)$$

where $\mathbf{T}_t^{(v,k)}$ and $\mathbf{T}_t^{(s,k)}$ denote the uniformly sampled point t on the warped visual trajectory and sensor trajectory in sliding window L_k, respectively. The initialization step is performed in the first sliding window, so $k = 1$. Fig.8(d) shows the result of IMU trajectories matched to visual trajectories. For better visualization, we can warp the top-down viewpoint to the scene-specific viewpoint by the inverse of $\mathbf{H}_a$. Therefore, two matrices, $\mathbf{H}_a$ (homography transformation) and $\mathbf{H}_{s,k}$ (similarity transformation), make visual and IMU trajectories compatible. $\mathbf{H}_a$ does not change unless the scene-specific viewpoint changes. $\mathbf{H}_{s,k}$ keeps being updated during each sliding window of the cooperative tracking.

B. Tracking

The initialization step only needs to be performed once, then our system goes to the normal tracking. Fig.8(f) and (g) show the trajectories based on visual and IMU tracking, respectively, in sliding windows $L_1 \sim L_k$. Then, $\mathbf{H}_{s,k-1}$ and $\mathbf{H}_a$ are applied to warp IMU and visual trajectories to the top-down viewpoint. The average distance d between trajectories in Fig.8(h) is calculated. If $d < d_{thr}$, visual and IMU trajectories are matched, then new $\mathbf{H}_{s,k}$ is computed using Eq.4 and we go to the next sliding window. In our tracking system, we set $d_{thr} = 80$inches.

C. Re-identification

Two cases lead to the re-identification: (1) The target disappears in visual tracking such as moving out of the visual field or being occluded by other objects; (2) Visual and IMU trajectories do not match each other, which may be caused by tracking drift (i.e., track a non-target pedestrian).

As shown in Fig.8(i), IMU keeps tracking the target even the target is occluded by a tree. The green curve is the IMU trajectory. Meanwhile, visual pedestrian detector tries to detect pedestrians in a search region estimated by IMU (yellow circle in Fig.8(i)). If detected, the pedestrian will be tracked by visual tracking for Δt frames (Fig.8(k)-Fig.8(m)). In this system, Δt is set as 150 frames (5 seconds). If any visual tracking failure happens within the Δt frames, we go back to the IMU-tracking (Fig.8(i)). If the tracking within the Δt frames succeeds, the

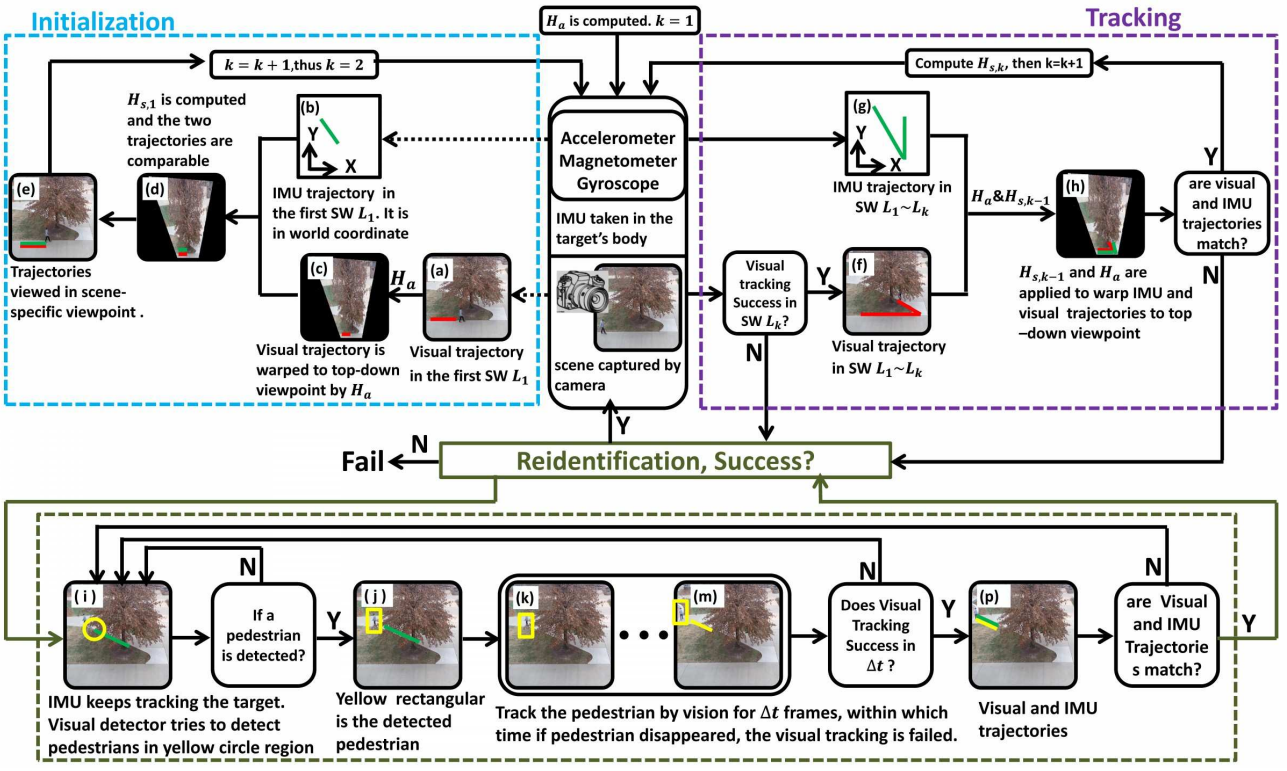

Fig. 8. The flow chart of the cooperative tracking system. SW: Sliding Window

average distance d between IMU and visual trajectories during Δt is computed to judge if they match. If $d < d_{thr}$, the target pedestrian is re-identified and we go back to the normal tracking again. Otherwise, we go back to the IMU-tracking (Fig.8(i)) for re-identification.

The above cooperative people tracking system elucidates why visual tracking and IMU tracking are "complementary". First, when visual tracking fails, IMU tracking keeps working and offers the clue where the target could be, helping visual tracking reidentify the target. Secondly, the visual trajectory corrects the bias of speed and forward direction estimation in IMU tracking by the similarity matrix $\mathbf{H}_{s,k}$. The calibration coefficient in Eq.3 is also compted by $\mathbf{H}_{s,k}$ once we know the length of matched visual and IMU trajectories. As we keep updating $\mathbf{H}_{s,k}$, visual tracking rebuilds the relationship with IMU tracking and rectifies the deviation of IMU-based tracking trajectory.

V. Experiments

To test the effectiveness of our cooperative tracking system, we apply it for people tracking in daily environments. Consumer electronics such as smartphones embedded with IMU modules are selected as the IMU signal collector. The IMU module in a smartphones is low cost and sensitive to noise. If our system works well using smartphones, we believe it will work using expensive and professional IMU devices. In addition, the popularity of smartphones offers more possibilities of applications of our tracking system. We developed an App to collect IMU signals when the target is moving or standing. The IMU signals are transmitted back to a groundstation by GSM. Meanwhile, a stationary surveillance camera collects visual signal of the target person. The visual signal is taken at 30 frames per second and the sampling frequency of IMU signal is 100 samples per second. The data transmitted between a smartphone and the groundstation is about 13.5 MB per hour. To synchronize the two signals, for every frame of the video, the nearest IMU signal is found according to the timestamp provided by the smartphone system.

A. Evaluation

We recorded four videos in different conditions to test the performance of our cooperative people tracking system. Fig.9 shows the visual and IMU trajectories from our tracking system.

· Video 1 was taken in an occlusion environment with a small slope. The target person was occluded by a tree twice for 9 and 16 seconds, respectively. Fig.9(a)(b) show that the target is successfully tracked by our system in long term and heavy occlusions.

· In video 2 (Fig.9(c)(d)), the target changed his speed from walking to sudden run and then stopped when hidden by the tree. Ten seconds later, the target began to walk in a direction different from his previous direction. This case is difficult for vision-based tracking algorithm because the target changes his

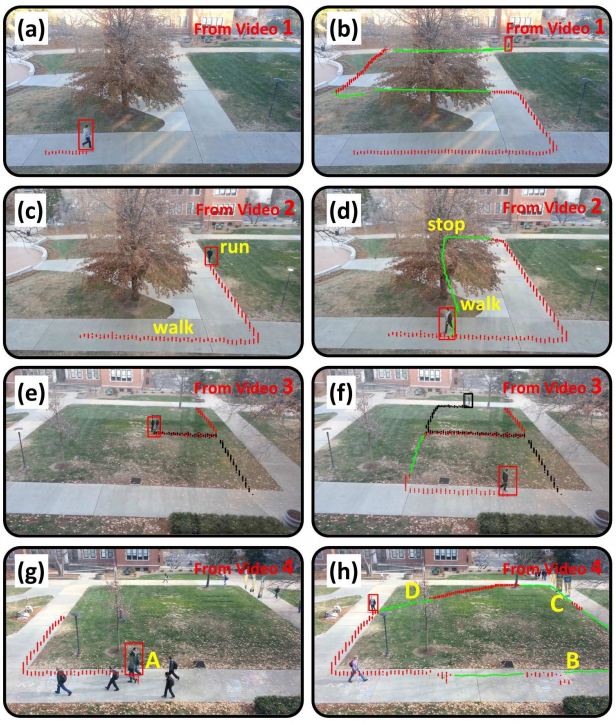

Fig. 9. Trajectories of the target person. Red curves are visual trajectories and green curves are IMU trajectories when visual tracking fails. (a)(b) Screenshots from video 1. (c)(d) Screenshots from video 2. (e)(f) Screenshots from video 3. (g)(h) Screenshots from video 4.

speed and forward direction when occluded.

· Video 3 (Fig.9(e)(f)) is an example of associating IMU trajectories with visual ones when multiple cooperative people carrying IMUs were in the scene. One target was occluded by the other. When they departed towards different directions, visual tracking failed because of the clutter of similar appearance. However, IMU tracking tracked and re-identified the target successfully.

· In video 4 (Fig.9(g)(h)), the target was occluded by moving pedestrians at location A and moved out of visual field from location B. At locations C and D, the target was occluded by background objects frequently. Despite the complex scenarios, our cooperative tracking system can persistently track the target.

Table I summarizes the quantitative evaluation on our cooperative tracking system. The two trajectories are synchronized by comparing their timestamps. The trajectory error is computed by the average difference between a tracked trajectory and its ground truth. The ground truth is labelled by a human annotator in each frame of the videos. The average errors of vision and IMU tracking in our cooperative people tracking are 37 inches and 44.3 inches, respectively. Visual tracking fails when the target is heavily occluded in the first time in each video, but our proposed tracking system can persistently track the target by combining visual and IMU tracking. The experiments validate that visual tracking combined with IMU tracking can achieve both accuracy and persistency. Here we did not provide the quantitative IMU-only tracking results, which is because IMU-only tracking needs manually set parameters such as the speed coefficient α in Eq.3 and the orientation offset [11]. These handcrafted parameters are person-specific and will largely affect the tracking performance. In the next subsection, we compare the *shape* of IMU tracking results which does not need manual parameters.

TABLE I
PEOPLE TRACKING RESULTS. FV: NUMBER OF FRAMES SUCCESSFULLY TRACKED ONLY BY VISION. FS: NUMBER OF FRAMES SUCCESSFULLY TRACKED BY OUR COOPERATIVE PEOPLE TRACKING SYSTEM. VTAE: VISUAL TRAJECTORY AVERAGE ERROR. ITAE: IMU TRAJECTORY AVERAGE ERROR

Video	#frames	FV	FS	VTAE(in)	ITAE(in)
1	2009	580	2009	50.3	66.0
2	1940	717	1940	35.15	45.2
3	1063	625	1063	22.4	15.8
4	1857	392	1857	33.1	36.1
avg				37.0	44.3

B. Comparison

We have seen IMU tracking can assist visual tracking in Section 5.1. We use video 1 as an example to compare different IMU tracking methods and shows the benefit of visual tracking to help IMU tracking. Fig.10(a) is the ground truth of the target trajectory in video 1, which is obtained by warping the target's trajectories in the scene-specific viewpoint to the top-down viewpoint using $\mathbf{H}_a$. All trajectories in Fig.10 are in the horizontal plane of the world coordinate. Fig.10(b) is based on the PCA2D (i.e., 2D Principle Component Analysis) method introduced in [11] which detects step in the time domain. There are many misdetections on step and direction by this approach and the tracked trajectory drifts away from the ground truth largely. Fig.10(c) is the result by our IMU tracking method without any assistance from the visual tracking, which is better but still drifts away from ground truth a little. Fig.10(d) shows the trajectory results of IMU tracking assisted by the visual tracking. When visual tracking is combined, visual trajectories constantly adjust the orientation and scale of IMU trajectories with $\mathbf{H}_{s,k}$. The IMU trajectory in Fig.10(d) is very close to the ground truth.

VI. CONCLUSION

To persistently track cooperative people such as children and patients in challenging scenarios, we present a novel tracking system combining the visual and Inertial Measurement Unit (IMU) signals, obtained from surveillance cameras and IMU devices carried by the targets themselves, respectively. Not only can IMU assist visual tracking when the target is occluded, but also the challenges of IMU tracking (calibration and drift) are alleviated when visual signals are available. Experimental results show that visual and IMU tracking are complementary to each other and their integration achieves

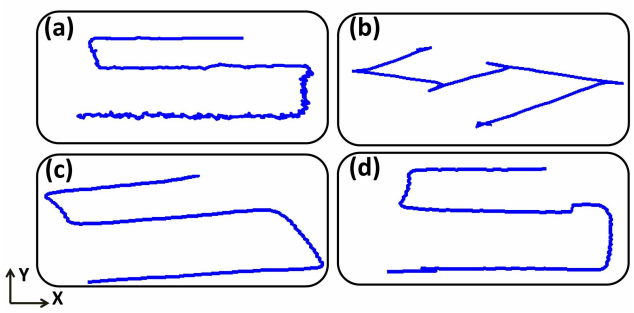

Fig. 10. IMU trajectories generated by three different approaches. (a) Ground truth trajectory; (b) Trajectory by time-domain step detection [11]; (c) Trajectory by our DFT approach; (d) Trajectory by our DFT approach assisted by the visual signal.

very good performance on persistent people tracking under challenging daily environments.

REFERENCES

[1] http://www8.garmin.com/aboutGPS/

[2] P. Dollar et al. "Pedestrian detection: An evaluation of the state of the art". IEEE Transaction on Pattern Analysis and Machine Intelligence, 34(4): 743-761, 2012.

[3] M. Enzweiler and D. Gavrila "Monocular pedestrian detection: Survey and experiments". IEEE Transaction on Pattern Analysis and Machine Intelligence, 31(12): 2179-2195, 2009.

[4] R. Feliz et al, "Pedestrian tracking using inertial sensors". Journal of Physical Agents, 3(1): 35-42, 2009.

[5] Y. Hua et al. "Occlusion and motion reasoning for long-term tracking". ECCV, 2014.

[6] Y. Jin et al. "A robust dead-reckoning pedestrian tracking system with low cost sensors". Pervasive Computing and Communications, 2011.

[7] M. Kourogi et al. "Personal positioning based on walking locomotion analysis with self-contained sensors and a wearable camera". International Symposium on Mixed and Augmented Reality, 2003.

[8] K. Kunze et al. "Which way am I facing: Inferring horizontal device orientation from an accelerometer signal". ISWC, 2009.

[9] M. Mirabi and S. Javadi. "People tracking in outdoor environment using Kalman filter". Intelligent Systems, Modelling and Simulation, 2012.

[10] N. Papadakis and A. Bugeau. "Tracking with occlusions via graph cuts". IEEE Transactions on Pattern Analysis and Machine Intelligence. 33(1): 144-157, 2011.

[11] U. Steinhoff and B. Schiele. "Dead reckoning from the pocket-an experimental study". Pervasive Computing and Communications, 2010.

[12] S. Tang et al. "Detection and tracking of occluded people". International Journal of Computer Vision, Published online. November 2013.

[13] B. Villiers et al. "Mean shift object tracking with occlusion handling". IAPR, 2012.

[14] B. Wu and R. Nevatia. "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors". International Journal of Computer Vision. 75(2): 247-266, 2010.

[15] J. Yan et al. "A novel occlusion-adaptive multi-object tracking method for road surveillance applications". Chinese Control Conference, 2013.

[16] B. Yuan et al. "Video tracking of human with occlusion based on Meanshift and Kalman filter". Applied Mechanics and Materials, 2013.

[17] Hide, Chris, Tom Botterill, and Marcus Andreotti. "Low cost vision-aided IMU for pedestrian navigation". Ubiquitous Positioning Indoor Navigation and Location Based Service (UPINLBS), 2010.

[18] Panahandeh, Ghazaleh, Dave Zachariah, and Magnus Jansson. "Exploiting ground plane constraints for visual-inertial navigation". Position Location and Navigation Symposium (PLANS), 2012.

[19] Panahandeh, Ghazaleh, Magnus Jansson, and Seth Hutchinson. "IMU-camera data fusion: Horizontal plane observation with explicit outlier rejection". Indoor Positioning and Indoor Navigation (IPIN), 2013.

[20] Barcelo, Guillem Casas, Ghazaleh Panahandeh, and Magnus Jansson. "Image-based floor segmentation in visual inertial navigation". Instrumentation and Measurement Technology Conference, 2013.