Accurate and Real-Time 3-D Tracking for the Following Robots by Fusing Vision and Ultrasonar Information

Mengmeng Wang, Yong Liu, *Member, IEEE*, Daobilige Su, Yufan Liao, Lei Shi, *Member, IEEE*, Jinhong Xu, and Jaime Valls Miro, *Member, IEEE*

***Abstract*—Acquiring the accurate three-dimensional (3-D) position of a target person around a robot provides valuable information that is applicable to a wide range of robotic tasks, especially for promoting the intelligent manufacturing processes of industries. This paper presents a real-time robotic 3-D human tracking system that combines a monocular camera with an ultrasonic sensor by an extended Kalman filter (EKF). The proposed system consists of three submodules: a monocular camera sensor tracking module, an ultrasonic sensor tracking module, and the multisensor fusion algorithm. An improved visual tracking algorithm is presented to provide 2-D partial location estimation. The algorithm is designed to overcome severe occlusions, scale variation, target missing, and achieve robust redetection. The scale accuracy is further enhanced by the estimated 3-D information. An ultrasonic sensor array is employed to provide the range information from the target person to the robot, and time of flight is used for the 2-D partial location estimation. EKF is adopted to sequentially process multiple, heterogeneous measurements arriving in an asynchronous order from the vision sensor, and the ultrasonic sensor separately. In the experiments, the proposed tracking system is tested in both a simulation platform and actual mobile robot for various indoor and outdoor scenes. The experimental results show the persuasive performance of the 3-D tracking system in terms of both the accuracy and robustness.**

***Index Terms*—Extended Kalman filter (EKF), following robot, three-dimensional (3-D) tracking, ultrasonic, visual tracking.**

Manuscript received September 28, 2017; revised December 21, 2017; accepted March 4, 2018. Date of publication March 28, 2018; date of current version June 12, 2018. Recommended by Technical Editor J. M. P. Geraedts. This work was supported in part by the National Natural Science Foundation of China under Grant U1509210 and Grant U1609210 and in part by the National Key Research and Development Program of China under Grant 2017YFB1302003. *(Corresponding author: Yong Liu.)*

M. Wang, Y. Liu, Y. Liao, and J. Xu are with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027, China (e-mail: mengmengwang@zju.edu.cn; yongliu@iipc.zju.edu.cn; liaoyufan@iipc.zju.edu.cn; xujinhong1992@qq.com).

D. Su, L. Shi, and V. M. Jaime are with the Centre for Autonomous Systems, University of Technology Sydney, Sydney, N.W.S. 2007, Australia (e-mail: daobilige.su@gmail.com; Lei.Shi-1@uts.edu.au; Jaime.VallsMiro@uts.edu.au).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMECH.2018.2820172

I. Introduction

THE following robot is one of the important autonomous devices in the intelligent manufacturing of industries. The core technology for the following robot is the accurate three-dimensional (3-D) position tracking for a moving object that the robot should follow in real time. It is also an essential building block of many advanced applications in the robotic areas, such as human–computer interaction, robot navigation, mobile robot obstacle avoidance, service robots, and industrial robots. For example, an autonomous transmitting robot tracks and follows a specific worker in order to provide cargo transportation services or to accomplish other manufacturing tasks in the complex indoor or outdoor factory environments. It is crucial to estimate the accurate positions of the target continuously for subsequent actions. To track the target person across various complex environments, robots need to localize the target and discriminate that specific target from other objects. In this context, localizing and tracking a moving target become critical and challenging for many indoor and outdoor robotic applications [1]–[4].

Tracking moving target for mobile robots has been a popular research topic in recent years, and many methods using various sensors have been developed [3], [5]–[8]. Among them, visual tracking enjoys a good population. It is an active research area in computer vision community and obtains significant progress over the past decade [9]–[13]. The visual tracking methods [14]–[16] usually employ the target's texture features in the sequenced RGB images to track the moving target. Recently, a group of correlation filter based discriminative trackers have made remarkable improvement in visual tracking field [10], [11], [17], [18], [19]. Due to the particularity of visual tracking, the correlation filter can be solved in the discrete Fourier transform (DFT) effectively and efficiently. These methods are skilled in many environments; however, they are not suitable for the 3-D target tracking in a robot platform, because they are not robust enough in the situations of severe occlusions and object missing in 2-D images. In addition, a monocular camera sensor can only obtain the 2-D position as it is insufficient to measure the range information from the robot to the following target. To introduce range information while retaining the advantages of visual tracking, an intuitive solution is to incorporate heterogeneous data from other sensors [6], [7], [20].

There are a number of different techniques to track target, especially for the human, with mobile robots. Using laser range finders with cameras for person tracking is frequent in the robotics community [3], [7], [20]. Laser scans are always used to detect the human legs at a fixed height. However, this cannot provide robust features for discriminating the different persons in the robot vicinity, while the detector fails when one leg occludes the other. Besides, the expensive laser range may also limit its implementation in popular industrial environments. The RGB-D sensors that can reconstruct 3-D features from 3-D point clouds are also used for the target tracking [1], [2], [21], [22]. However, the minimum distance requirement, narrow field of view, and sensitivity to the illumination variations of the RGB-D sensors limit this technique for robust target tracking applications.

Since ultrasonic sensors are widely used as range information detection, there are plenty of literature works that explore to localize and track targets by this kind of sensors [23]–[25]. However, the ultrasonic sensors may usually perform inaccurately when the target moves at the vertical direction of the sonar beam [25]. Another problem for sonar sensors is the ultrasonic reflection, which may cause invalid and incorrect results. In their approaches, the sonar array fires and receives along multiple paths, and it is likely to have reflection from the environments such as the walls, which leads to inaccurate tracking results. At the same time, if there are more than one users, i.e., more than one active sonar emitters in the sonar field of view, the passive sonar array cannot figure out the target person. Combining the sonar sensors with cameras is another popular research direction [5], [6], [26]. Most of these methods cascade the ultrasonic sensors and cameras together. They usually use the sonar sensors to detect the regions that might contain the target person in the sonar field of view. Corresponding regions in the images are then used as the priori searching areas for the target. This method may be invalid when the ultrasonic sensors lose the target, leading to the fact that the target is beyond the view of the camera sensor.

In this paper, we propose a new method for tracking the 3-D positions of a person by monocular vision and ultrasonic sequentially and independently in both indoor and outdoor environments with a robot platform. Due to the reliability and simplicity of the ultrasonic sensors, we fuse the partial position estimation from a camera and an ultrasonic sensor sequentially and exploit their respective advantages. Visual tracking processes videos captured from the camera sensor to estimate the target's locations in the image coordinate. Ultrasonic array sensor offers the range information of the target in the robot coordinate. The actual 3-D positions are estimated by merging these two heterogeneous information sources. This sensor configuration is an alternative to more complex and costly 3-D human tracking systems for mobile robots. In summary, the contributions of this paper can be presented as follows:

1) An accurate 3-D human tracking system is proposed by fusing a vision sensor with an ultrasonic array sensor sequentially by the extended Kalman filter (EKF).
2) An improved real-time visual tracking algorithm is presented to handle the situations of severe occlusion, object missing, and redetection;
3) The estimated 3-D information is further exploited to improve the scale accuracy of the target in the image coordinate.

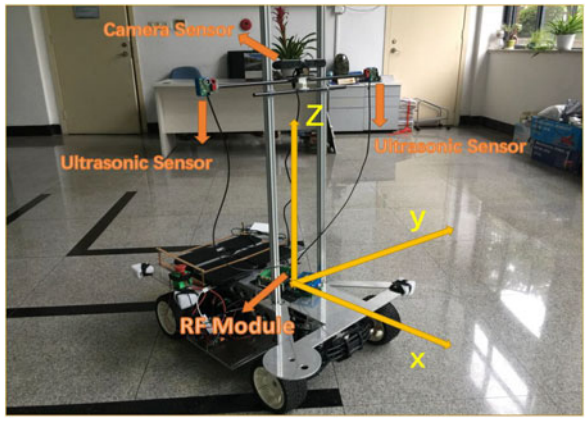


Fig. 1. The local robot coordinate. We employ a camera sensor (Xtion Pro LIVE) to acquire the ground truth of the 3-D position of the target during the tracking process in the experiment. Simultaneously, the RGB camera of Xtion is used as our monocular camera sensor for convenience.

In the experiment, we demonstrate the proposed method both in the robot platform and simulation. The experimental results show that our method performs accurately and robustly in the 3-D human tracking for the challenging conditions, such as occlusions, background clutters, scale variations, and even when the person is totally missing.

II. Method

The proposed 3-D tracking system can be decomposed into three submodules: a monocular camera sensor tracking module, an ultrasonic sensor tracking module, and a multisensor fusion algorithm. In this section, the details about these three submodules are presented.

The state of the target $\mathbf{x}_k = [x_k, y_k, z_k]^T$ is defined as the location of the sonar emitter wore by the target person. Here, the subscript k represents the kth time instant. All the variables are defined in the robot local coordinate frame as shown in Fig. 1. The target person is sensed by both a vision sensor and an ultrasonic sensor. To estimate the 3-D position of the person, the two data streams from the two sensors are fused sequentially using an EKF.

A. Monocular Camera Sensor Tracking Module

The monocular camera is installed on the top of the ultrasonic array sensor, which is attached on the mobile robots. The vision sensor measurement model $\mathbf{h}_C(\mathbf{x}_k)$ is a simple camera projection model as follows:

$$\mathbf{h}_C(\mathbf{x}_k) = [u_{Ck}, v_{Ck}]^T \tag{1a}$$

$$\begin{bmatrix} u_{Ck} \\ v_{Ck} \\ 1 \end{bmatrix} = \mathbf{A}[\mathbf{R}|\mathbf{t}] \begin{bmatrix} \mathbf{x}_k \\ 1 \end{bmatrix} \tag{1b}$$

where (u_{Ck}, v_{Ck}) is the target's location in the image coordinate, which is estimated by our visual tracking algorithm, and $[\mathbf{R}|\,\mathbf{t}]$ and $\mathbf{A}$ are the extrinsic and intrinsic parameter matrices of the camera separately.

For conventional visual tracking, the target is given in the first frame either from human annotation or a certain detector. In the proposed 3-D human tracking system, the initial bounding box is calculated by the 3-D to 2-D projection with the target person's height h and the initial 3-D position $\mathbf{x}_{\text{init}}$. Additionally, we assume the average width of a person is 0.4 m and the distance from the sonar emitter to the person's feet is 50% of his/her height h in all experiments. Then, the initial 3-D positions of left boundary $\mathbf{x}_{\text{linit}}$, right boundary $\mathbf{x}_{\text{rinit}}$, head $\mathbf{x}_{\text{hinit}}$, and feet $\mathbf{x}_{\text{finit}}$ of the target can be calculated by

$$\mathbf{x}_{\text{linit}} = \mathbf{x}_{\text{init}} + [0, 0.4/2, 0]^T \tag{2a}$$

$$\mathbf{x}_{\text{rinit}} = \mathbf{x}_{\text{init}} + [0, -0.4/2, 0]^T \tag{2b}$$

$$\mathbf{x}_{\text{hinit}} = \mathbf{x}_{\text{init}} + [0, 0, 0.5h]^T \tag{2c}$$

$$\mathbf{x}_{\text{finit}} = \mathbf{x}_{\text{init}} + [0, 0, -0.5h]^T. \tag{2d}$$

The initial width w_{init}, height h_{init}, and the center position of the target's bounding box $(u_{\text{init}}, v_{\text{init}})$ in the image is calculated as

$$w_{\text{init}} = u_{\text{rinit}} - u_{\text{linit}} \tag{3a}$$

$$h_{\text{init}} = v_{\text{finit}} - v_{\text{hinit}} \tag{3b}$$

$$u_{\text{init}} = (u_{\text{linit}} + u_{\text{rinit}})/2 \tag{3c}$$

$$v_{\text{init}} = (v_{\text{finit}} + v_{\text{hinit}})/2 \tag{3d}$$

where $u_{\text{rinit}}, u_{\text{linit}}$ are the u axis values in the image coordinate of $\mathbf{x}_{\text{linit}}$ and $\mathbf{x}_{\text{rinit}}$ calculated by (1b). Similarly, v_{finit} and v_{hinit} are the v-axis values in the image coordinate of $\mathbf{x}_{\text{hinit}}$ and $\mathbf{x}_{\text{finit}}$, respectively.

The presented visual tracking algorithm is based on the kernelized correlation filter (KCF) [10] tracker. We extend it with a novel criterion to evaluate the performances of the tracking results and develop a new scale estimation method that estimates the scale variations by combining the projection from the 3-D target position into the 2-D image coordinates with the visual scale estimations.

1) KCF Tracker: In this section, a brief exposition of KCF tracking algorithm is presented, which is described in detail in [10]. The goal is to learn an online correlation filter from a plenty of training samples of size $W \times H$. KCF considers all cyclic shifts $\mathbf{s}_{w,h}$, $(w,h) \in \{0, ..., W-1\} \times \{0, ..., H-1\}$ around the target as training examples. The desired correlation output $y_{w,h}$ is constructed as a Gaussian function with its peak located at the target center and smoothly decayed to 0 for any other shifts.

The optimal correlation filter $\mathbf{w}$ is obtained by a function that minimizes the squared error over samples $\mathbf{s}_{w,h}$ and their regression labels $y_{w,h}$

$$\min_{\mathbf{w}} \sum_{w,h} \left|\langle \varphi(\mathbf{s}_{w,h}), \mathbf{w}\rangle - y_{w,h}\right|^2 + \lambda \|\mathbf{w}\|^2 \tag{4}$$

where φ denotes the mapping to nonlinear feature space with kernel κ and the dot product of $\mathbf{s}$ and $\mathbf{s}'$ is $\langle \varphi(\mathbf{s}), \varphi(\mathbf{s}')\rangle = \kappa(\mathbf{s}, \mathbf{s}')$. λ is a regularization parameter that controls overfitting.

With the fact that all circulant matrices are made diagonal by the DFT and circulant kernels, the solution $\mathbf{w}$ can be represented as $\mathbf{w} = \sum_{w,h} \alpha_{w,h} \varphi(\mathbf{s}_{w,h})$, then the optimization goal is the variable α rather than $\mathbf{w}$:

$$\alpha = \mathscr{F}^{-1}\left(\frac{\mathscr{F}(\mathbf{y})}{\mathscr{F}(\mathbf{k}^{\mathbf{ss}}) + \lambda}\right) \tag{5}$$

where $\mathscr{F}$ and $\mathscr{F}^{-1}$ denote the DFT and its inverse. $\mathbf{k}^{\mathbf{ss}}$ is the kernel correlation of the target appearance model $\mathbf{s}$ with itself. Each cyclically shifted training sample $\mathbf{s}_{w,h}$ actually consists of certain feature maps extracted from its corresponding image region.

In the tracking process, a new image region $\mathbf{r}$ centered at the position of the last frame is cropped in the new frame. The position of the target is found in the maximum response of the output response map $f(\mathbf{r})$:

$$f(\mathbf{r}) = \mathscr{F}^{-1}\left(\mathscr{F}(\mathbf{k}^{\mathbf{sr}}) \odot \mathscr{F}(\alpha)\right) \tag{6}$$

where $\odot$ is the element-wise product and $\mathbf{k}^{\mathbf{sr}}$ represents the kernel correlation of $\mathbf{s}$ and $\mathbf{r}$.

Note that in KCF, α in (5) and the target appearance model $\mathbf{s}$ are updated continuously. The model will be corrupted when the object is occluded severely or totally missing and adapt to the wrong background or obstacle regions as shown in the third row of Fig. 2. This will lead to incorrect tracking results and missing the target in the following frames.

2) Analysis of the Response Map: In visual tracking, severe occlusion and missing are two of the main challenging factors that limit tracking performances. As mentioned above, the KCF tracker cannot avoid the model corruption due to the lack of the feedback of the tracking results.

The response map is the correlation response used to locate the position of target as in (6). It reveals the degree of confidence about the tracking results to some extent. The response map should have only one sharp peak and be smooth in all other areas when the detected target in the current frame is extremely matched to the correct target. The sharper the correlation peaks are, the better the location accuracy is. If the object is occluded severely or even missing, the whole response map will fluctuate intensely, whose pattern is significantly different from the normal response map, as shown in Fig. 2. If we continue to detect a target as normal, the results will be wrong mostly. So, we explore a novel criterion for severe occlusion, as well as retaining the advantages of KCF.

For correlation filter based classifier, the peak-to-sidelobe ratio (PSR) can be used to quantify the sharpness of the correlation peak. However, PSR is still not robust to partial occlusions [19]. We propose a novel criterion called PCE measure in the following equation:

$$\text{PCE} = \frac{|y_{\max}|^2}{E_y} \tag{7}$$

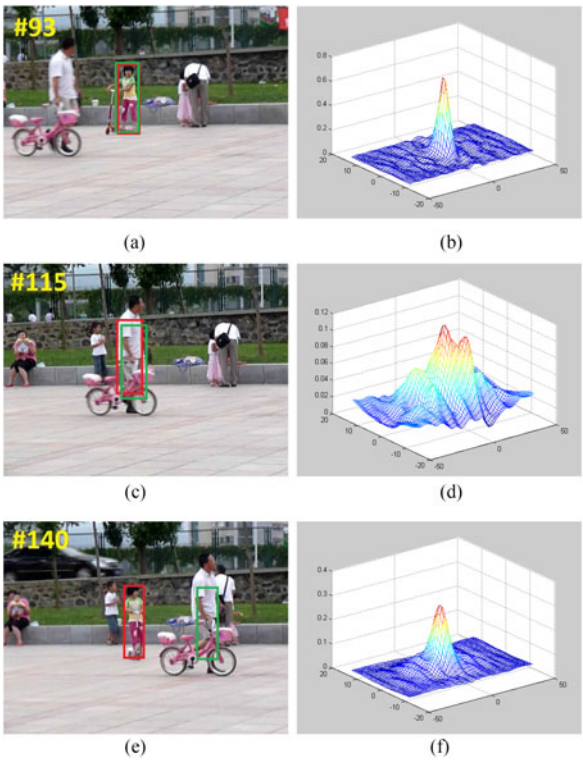


Fig. 2. First column shows the original frames from the vision sensors, and the second column reveals the corresponding response maps. The red bounding box represents the found target of our method, while the green one denotes the tracking result of KCF tracker. When the girl is fully occluded, the corresponding response map will fluctuate intensely. By introducing the proposed criterion peak-to-correlation energy (PCE), the target girl is redetected again by our method and the response map returns to the normal pattern. However, the KCF tracker loses the target due to the model corrupting during the occlusion. This video is from a visual tracking benchmark [15]. (a) Normal. (b) PCE = 0.786. (c) Occluded. (d) PCE = 0.087. (e) Redetected. (f) PCE = 0.671.

where $|y_{\max}|$ denotes the maximum peak magnitude, and the correlation response map energy E_y is defined as follows:

$$E_y = \sum_{w,h} |y_{w,h}|^2. \tag{8}$$

For sharper peak, i.e., the target apparently appearing in the visual field of the robot, E_y will get close to $|y_{\max}|^2$; thus, PCE will approach to 1. Otherwise, PCE will approach to 0 if the object is occluded or missing. When the PCE is lower than a predefined threshold, as shown in the second row of Fig. 2, the target appearance model and the filter model will not be updated.

3) Scale Estimation: When a robot tracks the target in front of it, the relative velocity of the robot and the target is changing all the time, and the size of the target in the image is varying according to the distance between the target and robot.

To handle scale variation $s_{2\text{-D}}$ in the 2-D visual tracking process, we employ the discriminative scale space tracking (DSST) [11] algorithm. First, the position of the object is determined by the learned translation filter with abundant features. Second, a group of windows with different scales is sampled around this position and correlated with the learned scale filter via coarse features. For each scale, the maximum value of its response map is measured as its matching score. The scale with the highest score is regarded as $s_{2\text{-D}}$. At the meantime, the standard variance $\sigma_{2\text{-D}}$ from $s_{2\text{-D}}$ is calculated as the uncertainty of $s_{2\text{-D}}$.

We also consider the scale states calculated from the 3-D position estimations. At the kth frame, the 3-D position $\mathbf{x}_k$ is estimated. Then, we can get the 3-D positions of the head $\mathbf{x}_{hk}$ and feet $\mathbf{x}_{fk}$ by (2b) and (2c) as the height h of the target is fixed during tracking.

We can obtain $\mathbf{v}_{hk}$ and $\mathbf{v}_{fk}$ by projecting $\mathbf{x}_{hk}$ and $\mathbf{x}_{fk}$ into the image space through (1b), where $\mathbf{v}_{hk}$ and $\mathbf{v}_{fk}$ are the v-axis values in the image space of head and feet, respectively. We assume that the scale variation of the height and width is synchronous. Then, the scale variation from the 3-D position is obtained from

$$s_{3\text{-D}} = (v_{fk} - v_{hk}) / v_{\text{init}} \tag{9}$$

where v_{init} is the initial height of the target calculated by (3d). Finally, the scale s_k of the kth frame is calculated as follows:

$$s_k = \left(\frac{s_{2\text{-D}}}{\sigma_{2\text{-D}}} + \frac{s_{3\text{-D}}}{\sigma_{3\text{-D}}}\right) \frac{\sigma_{2\text{-D}}\sigma_{3\text{-D}}}{\sigma_{2\text{-D}} + \sigma_{3\text{-D}}} \tag{10}$$

where $\sigma_{3\text{-D}} = \sqrt{\mathbf{P}_k(3,3)}$ is the uncertainty of $s_{3\text{-D}}$, and $\mathbf{P}_k$ is the covariance matrix of the kth estimated state.

B. Ultrasonic Sensor Tracking Module

In the proposed tracking system, we consider the traditional sonar array sensors to obtain the range information and improve the predicted accuracy of the tracking target. Traditional sonar array sensors always use time of flight (TOF) and triangulation to find the relative location of a target with respect to the source.

The active sonar emitter array that consists of one sonar sensor is designed as a human carrying portable user device (POD). The corresponding passive sensor receivers with two sonar units are attached equally spaced in front of the robot. In order to mate the sonar emitter and receiver, we introduce the radio frequency (RF) wireless module to transmit and receive radio signals between the POD and passive sensor array.

When the RF module on the robot receives the RF signal from the POD, it will start a timer and respond to the RF on the POD to make the sonar emitter launch an ultrasonic signal. Then, the time lapsed from when the timer starts until all the sonar units measure an incoming signal is the corresponding TOF. This time actually includes the response time caused by the RF module. We find out that this response time is fixed all the time and can be measured as Δt.

As shown in Fig. 3, the sonar emitter of the POD in the ultrasonic coordinate is denoted as (x_{Uk}, y_{Uk}). It can be calculated from

$$d_1 = v\,(t_1 - \Delta t) \tag{11a}$$

$$d_2 = v\,(t_2 - \Delta t) \tag{11b}$$

$$(x_{Uk} - 0)^2 + \left(y_{Uk} - \frac{D}{2}\right)^2 = {d_1}^2 \tag{11c}$$

$$(x_{Uk} - 0)^2 + \left(y_{Uk} - \left(-\frac{D}{2}\right)\right)^2 = {d_2}^2 \tag{11d}$$

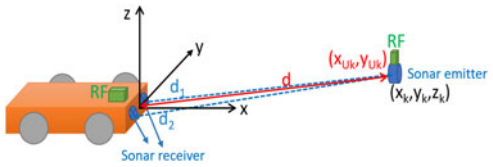


Fig. 3. Illustration of the ultrasonic sensor tracking module. The black coordinate on the robot denotes the local robot coordinate, and the black location (x_k, y_k, z_k) indicates the sonar emitter position in this coordinate. The red location (x_{Uk}, y_{Uk}) denotes the sonar emitter position in the ultrasonic coordinate, d_1, d_2 show the distances from the sonar emitter to the two sonar receivers, and d is the final distance between the robot and the sonar emitter.

where v denotes the sound velocity, which is generally considered to be 340 m/s. D indicates the distance between the two passive sonar receivers, and t_1 and t_2 are the TOF of them.

Transform to the kth system state $\mathbf{x}_k = [x_k, y_k, z_k]^T$ in the robot coordinate, the expected measurement $\mathbf{h}_U(\mathbf{x}_k)$ of the ultrasonic sensor is denoted as

$$\mathbf{h}_U(\mathbf{x}_k) = [x_{Uk}, y_{Uk}]^T \tag{12}$$

where $x_{Uk} = \sqrt{{x_k}^2 + {z_k}^2}$ and $y_{Uk} = y_k$.

C. Multisensor Fusion

As the sampling frequencies of the two sensors are different, the fusion algorithm will be run whenever any of them is updated. A standard EKF approach is utilized to fuse the measurements obtained from the ultrasonic sensor and the vision sensor. We adopt such a method to sequentially process the multiple, heterogeneous measurements arriving in an asynchronous order [27].

1) Prediction Step: As we have no knowledge of the target motion, a random walk or a constant velocity model can be used to predict the target location in the robot coordinate. In the case of random walk model, such as

$$\mathbf{x}_k = \mathbf{x}_{k-1} \tag{13a}$$

$$\mathbf{P}_k = \mathbf{G}\mathbf{P}_{k-1}\mathbf{G}^T + \mathbf{R}_k \tag{13b}$$

$$\mathbf{R}_k = \mathbf{R}(t_k - t_{k-1}) \tag{13c}$$

where $\mathbf{P}_k$ is the covariance matrix, $\mathbf{G}$ is the Jacobian of (13a) (a 3 by 3 identity matrix in the random walk model). $\mathbf{R}_k$ is the motion noise during the time step $(t_k - t_{k-1})$, so it is proportional to $(t_k - t_{k-1})$ by a constant noise level $\mathbf{R}$.

2) Correction Step: Whenever any measurement is available, the system state is updated by

$$\mathbf{K}_{*k} = \mathbf{P}_k\mathbf{H}_{*k}^T\left(\mathbf{H}_{*k}\mathbf{P}_k\mathbf{H}_{*k}^T + \mathbf{Q}_{*k}\right)^{-1} \tag{14a}$$

$$\mathbf{x}_k = \mathbf{x}_k + \mathbf{K}_{*k}\left(\mathbf{z}_{*k} - \mathbf{h}_*(\mathbf{x}_k)\right) \tag{14b}$$

$$\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_{*k}\mathbf{H}_{*k})\mathbf{P}_k \tag{14c}$$

where the $*$ in the subscript stands for either ultrasonic (U) or camera (C) measurement, we collectively call it *measurement next*. $\mathbf{h}_*(\mathbf{x}_k)$ is the sensor model that provides the predicted measurement. The camera sensor model $\mathbf{h}_C(\mathbf{x}_k)$ is defined in (1), and the sonar sensor model $\mathbf{h}_U(\mathbf{x}_k)$ is defined in (12). $\mathbf{H}_{*k}$ is the Jacobian matrix of $\mathbf{h}_*(\mathbf{x}_k)$. $\mathbf{z}_{*k}$ is the actual sensor measurement. $\mathbf{z}_{Ck}$ is estimated from the visual tracking algorithm by (6) and $\mathbf{z}_{Uk}$ is the mean distance predicted by (11). $\mathbf{Q}_{*k}$ is the measurement noise. For sonar sensor, $\mathbf{Q}_{Uk}$ is the covariance matrix from TOF. For the camera sensor, when the PCE is larger, the noise is smaller. The trend of PCE is opposite to the measurement noise $\mathbf{Q}_{Ck}$. Thus, we use the reciprocal of the PCE criterion and divide it by 100 to obtain the correct order of magnitude of the noise $\mathbf{Q}_{Ck}$. It is experimentally defined as

$$\mathbf{Q}_{Ck} = \begin{bmatrix} \frac{0.002}{\text{PCE}} & 0 \\ 0 & \frac{0.002}{\text{PCE}} \end{bmatrix}. \tag{15}$$

So, the input from the camera sensor will be $\mathbf{z}_{Ck}$ and $\mathbf{Q}_{Ck}$, both coming from the visual tracking algorithm. The input from the sonar sensor will be $\mathbf{z}_{Uk}$ and $\mathbf{Q}_{Uk}$, both coming from (11).

III. Experiments

To evaluate the performance of our multisensor 3-D human tracking system, sufficient experiments are carried out in both simulation environments and real world scenarios. The simulation is done by a robot simulator named virtual robot experimentation platform (V-REP), which is used for fast prototyping and verification to validate the accuracy of the proposed tracking system. We implement our 3-D tracking system in a new robot platform named Rock105, as shown in Fig. 1. It is a differential mobile robot system fitted with an autotipping mechanism [25]. The detailed tracking processes of all the experiments are shown in our video demo.

A. Implementation Details

All our experiments are performed using MATLAB R2015a on a 3.2 GHz Intel Core i7 CPU with 16 GB RAM. The robot operating system (ROS) has been employed as the software framework for the Rock105 platform, linked to MATLAB via the MATLAB-ROS bridge package. The camera sensor and the sonar receivers are both installed in front of the Rock105 in the same vertical plane facing front. The visual tracking algorithm runs at an average speed of 25 fps and the prediction of sonar runs at about 20 Hz. The height and width of each target will not impact the initialization of the proposed tracking system. We regard the height h as a parameter and set up with the actual height of the target in all experiments. The width is average as 0.4 m, which does not influence the results. The setup includes two parameters for ultrasonic sensor tracking. The first one is the distance between the two passive sonar receivers D, which is 0.6 m in Rock105. The other one is the response time of RF module Δt, which is 0.152 ms.

The setup of the visual tracking algorithm is detailed here. For the visual features, we employ the multichannel features based on histogram of oriented gradient (HOG) [28] with a cell size of 4 pixels, as well as color names (CNs) [17] with the first two dimensions. The threshold of the PCE criterion is set to 0.2 from experiments. When PCE is larger than 0.2, the target appearance and the filter models are updated. Otherwise, the target is perceived as occluding or missing, so the target appearance and the filter models are not updated. The regularization parameter λ in (4) is set to 0.0001.

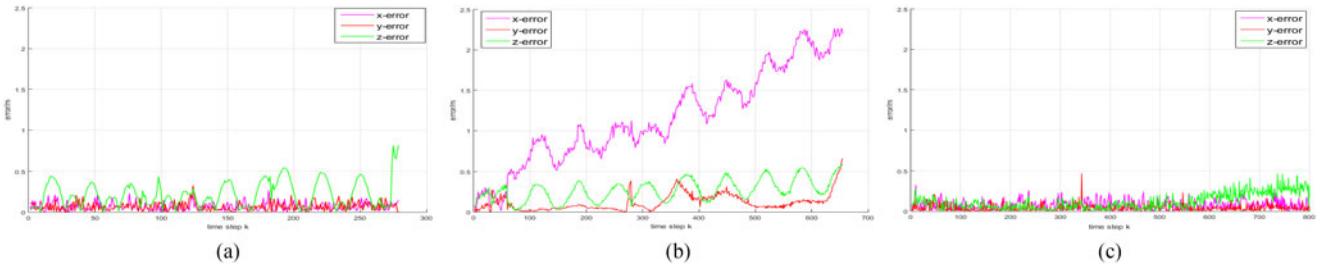


Fig. 4. Simulation outdoor scene. Error analysis for single sensor. (a) Ultrasonic only. (b) Camera only. (c) Camera + ultrasonic.

TABLE I
MEAN ERROR IN THREE AXES

Axis	SI(C+U)	SO(C+U)	SO(C)	SO(U)	RI(C+U)	RO(C+U)
x(m)	0.144	0.07521	1.1650	0.1413	0.168	0.2045
y(m)	0.1062	0.03776	0.1169	0.1364	0.1091	0.1262
z(m)	0.1085	0.1167	0.2679	0.4366	0.116	0.1231

S stands for simulation, R for Rock105, I for indoor, O for outdoor, C for the camera sensor, and U for the ultrasonic sensor.

In order to demonstrate the performance of the proposed 3-D tracking system, we test it in the simulation experiments and the Rock105 robotic platform in both indoor and outdoor environments. To illustrate that under normal walking paces and patterns the proposed tracking system is able to effectively track the target person, we apply a simple proportional controller in translation and orientation velocity to make the robot track automatically.

B. Experiments on Simulation Platform

In the simulated robotic platform, there are no disturbance and noise like the actual robots and scenarios. Thus, we do experiments in the simulated platform to verify the theoretical framework of the proposed 3-D tracking system. In the simulated robotic platform, a passive sensor array with two sonar units is attached equally spaced in front of the robot. The camera sensor is fixed below the sonar array.

1) Multisensors Validation: For safety, we compare the tracking results with individual sensor respectively only in the simulation outdoor scene to validate the necessity of the two sensors. Without the monocular camera sensor, the estimated accuracies of the z-axis are dramatically reduced, as shown in Fig. 4(a). The fluctuation of z-error is due to the target's movement on the z-axis that is designed to act as a simple sine curve. Without the sonar sensor, the estimation of the x-axis is incorrect, as shown in Fig. 4(b) due to the lack of information in this dimension. The tracking result is imponderable when the visual tracking algorithm loses the target. The corresponding mean errors are shown in Table I. Combining the ultrasonic sensor and the camera sensor together, the position of x and z axes can be well compensated and the position of the y-axis can also be enhanced accurately. The effectiveness is validated by the experiments. The tracking results on all three directions are getting more stable and accurate, as shown in Fig. 4(c).

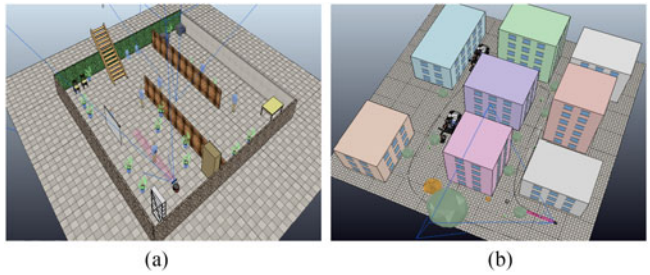


Fig. 5. Simulation scenes. (a) Indoor. (b) Outdoor.

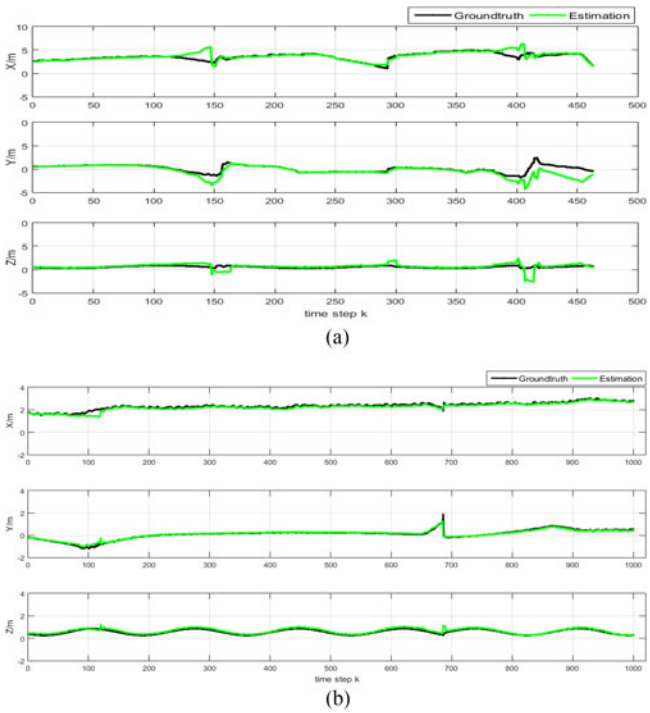


Fig. 6. 3-D tracking results of simulation experiments. (a) Indoor scene results. (b) Outdoor scene results.

2) Indoor and Outdoor Scenes: Before using the proposed method in our real-world robots, we test it in simulation platforms in both indoor and outdoor scenes. The indoor scene is constructed as an office room with many persons inside it like Fig. 5(a). The outdoor scene in Fig. 5(b) is built with plenty of buildings, trees, vehicles and people, just like a city area. The quantitative 3-D tracking results on the simulation scenes are shown in Fig. 6, where the green lines represent the ground truth

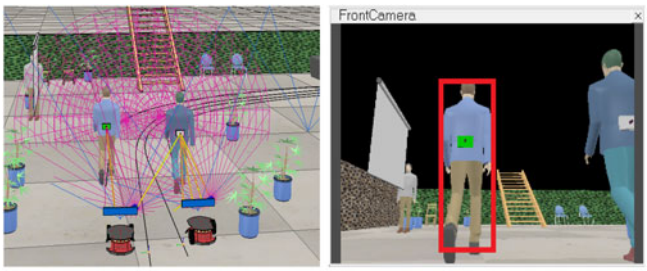


Fig. 7. Left shows the scene shot of multitarget scene. Right one denotes the camera view of the corresponding robot. (a) Multitarget scene. (b) Camera view.

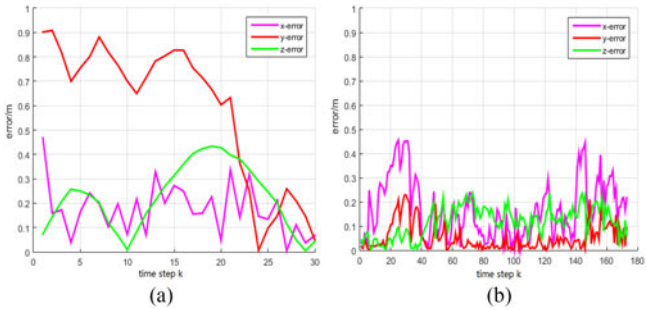


Fig. 8. Error analysis of multitarget interference. (a) Ultrasonic only. (b) Camera + ultrasonic.

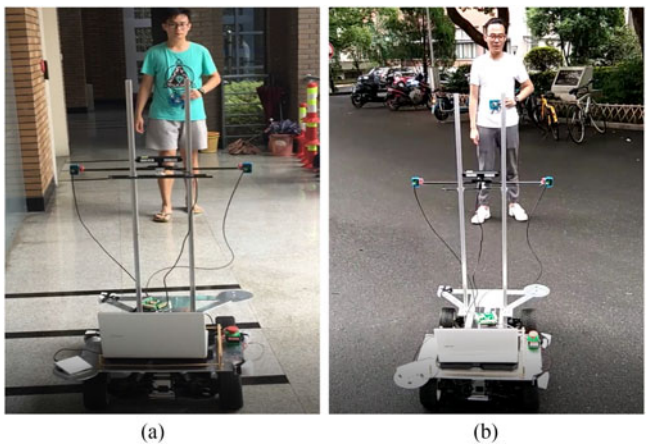


Fig. 9. Real-world scenes. (a) Indoor. (b) Outdoor.

of the target motions in three axes, and the black lines denote the estimation of the proposed tracking system. It can be observed that the tracking errors is markedly small since the two lines are closed to each other in all three axes.

3) Multitarget Interference: When there are two or more persons carrying with sonar PODs in the same scenario, the ultrasonic sensor might lose the correct target due to the distraction from other sonar PODs. We have experimented under this situation with only sonar sensor as shown in Fig. 7(a), there are two pairs of target-robot groups that are close to each other. As long as the RF module on the robot receives the RF signal from its corresponding POD, it will start a timer to measure the time lapsed from when the timer starts until the sonar units

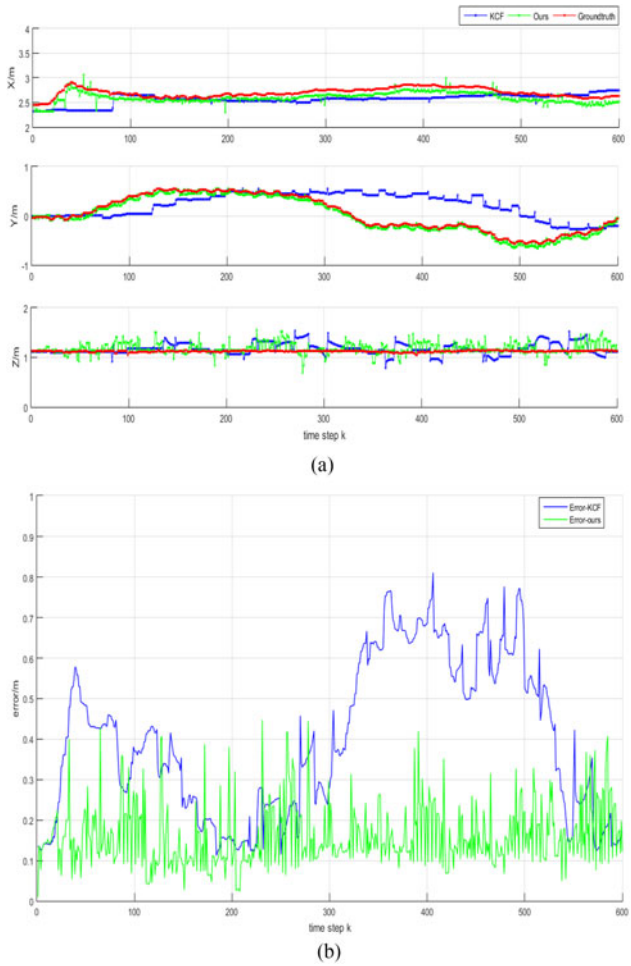


Fig. 10. Comparison with KCF tracker. (a) The results in all three axes. (b) Error comparison.

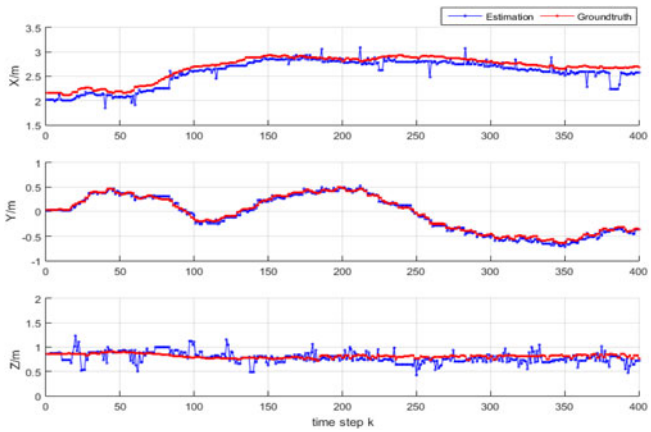


Fig. 11. 3-D tracking results on Rock105 in the outdoor scenario.

measure an incoming signal as the corresponding TOF. The RF modules are unique to each target–robot pair, while the ultrasonic active and passive sensors are not. Therefore, when the RF module starts a timer, if the sonar signal from another unmatched POD is received earlier than the matched one, the measured sonar signal is more likely to be a distraction. The

Fig. 12. Multitarget interference and illumination variation from the robot view. (a) Normal condition. (b) Interferential target occlusion. (c) Illumination variation.

error analysis for the left target-robot pair of Fig. 7(a) is shown in Fig. 8(a). The large error in the y-axis reflects the fact that the sonar sensor might be disturbed by nearby sonar PODs carried by other targets. In the proposed method, the camera sensor is employed to handle this problem, as shown in Fig. 7(b). The proposed visual tracking algorithm can track the target accurately without being disturbed by the nearby target under the circumstances. Thus, the proposed 3-D tracking method can still track the correct target with little error in all three axes, as shown in Fig. 8(b).

C. Experiments on Actual Mobile Robot

Further, we experiment the proposed 3-D tracking system in the Rock105 platform, as shown in Fig. 1. We introduce the skeleton tracking of Xtion PRO LIVE [1] to get the ground truth of the 3-D positions of the target during the tracking process through the OPENNI NITE2 tracker package. The position of the waist in the skeleton is regarded as the true position of the sonar POD carrying with the person. Simultaneously, the RGB camera of Xtion is used as our monocular camera sensor. As shown in Fig. 9, the indoor experiment is performed in the common corridor of our laboratory, while the outdoor experiment is conducted outside the robotics laboratory in the Zhejiang University.

1) Visual Tracking Algorithm Validation: The proposed monocular camera sensor tracking module is based on the KCF tracker [10]. However, the KCF tracker cannot handle the object occlusion and scale variation, which are important in the 3-D tracking system. We have made two main contributions beyond KCF. First, we explore a criterion to judge whether the target is occluded, and second, we combine the 2-D and 3-D scale estimation together to improve the scale accuracy. To validate the improvement of the proposed visual tracking model in our system, we conduct the experiment with both the KCF tracker and the proposed visual tracking algorithm on the Rock105 platform. The results of the experiments are shown in Fig. 10.

As shown in Fig. 10(a), the proposed visual tracking model outperforms KCF tracker especially in the y-axis. The errors in Fig. 10(b) are computed by the Euclidean distance between the estimated 3-D position and the ground truth. It is obvious that the errors of the proposed tracker are smaller than KCF tracker. It is mainly due to the robustness of the proposed PCE criterion, which can judge the wrong tracking results and avoid the model drifting problem.

2) Indoor and Outdoor Scenes: There are many challenges in these scenes such as illuminate variations, scale variations, part occlusions, severe occlusions, background clutters, and object missing. The target person is walking with the variations in all three ases to make the 3-D estimation more challenging. We show these challenges in our video demo with both indoor and outdoor experiments. Indoor results can be seen from Fig. 10, where the red lines show the ground truth of the target motions, while the green ones denote the estimation of our method. Outdoor results are shown in Fig. 11, where the red lines show the ground truth of the target motions, while the blue ones denote the estimation of our method. The tracking results of real-world robot experiments are actually impressive with very small errors as well. We calculate the mean error of all the experiments in Table I. The results illustrate a great performance of our proposed system.

3) Multitarget Interference: It is very common to encounter the situation with multitarget in the actual scenes. Therefore, we do another experiment with multitarget interference to show the robustness of the proposed 3-D tracking system. The interferential target also carries a sonar emitter and walks near the Rock105 robot. To make the experiment more challenging, the interferential target directly goes through and totally occludes the correct target three times, as shown in Fig. 12(b). Besides, the illumination varies sharply like Fig. 12(c). The quantitative 3-D tracking results are shown in Fig. 13, where the red lines show the ground truth of the target motions, while the blue ones denote the estimation of our method. The results apparently reflect that the proposed 3-D tracking system is robust to the multitarget interference. When the sonar sensor is disturbed by the interferential target, the proposed visual tracking algorithm can still track the target accurately under the circumstances and the proposed PCE criterion can handle the object occlusion problem robustly. Thus, the proposed 3-D tracking system can overcome the multitarget interference effectively.

[1] https://www.asus.com/3D-Sensor/Xtion_PRO/specifications/

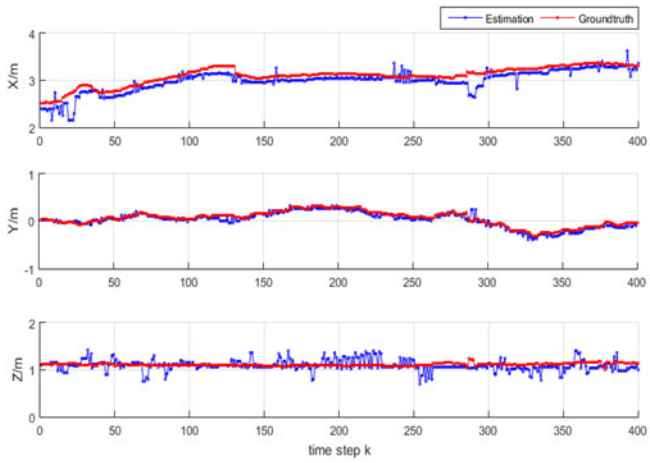


Fig. 13. 3-D tracking results of the multitarget interference experiment.

IV. Conclusion

In this paper, we address the problem of accurately estimating the 3-D position of the target in front of the mobile robot for tracking purposes, in both indoor and outdoor environments. Our approach fuses the partial location estimations from a monocular camera and an ultrasonic array. To improve the robustness of the tracking system, a novel criterion in the visual tracking model is introduced to overcome the problems of occlusions, scale variation, targets missing, and redetection. The ultrasonic sensor is used to provide the range-based location estimation. Information from two heterogeneous sources is processed with EKF sequentially to handle their different update rates. The estimated 3-D information is further exploited to improve the scale accuracy. The proposed approach is implemented and tested in both simulation and real-world scenarios. As the evaluation results show, the proposed algorithm is able to produce stable, accurate, and robust 3-D position estimations of the target in real time.

References

[1] A. P. Gritti *et al.*, "Kinect-based people detection and tracking from small-footprint ground robots," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2014, pp. 4096–4103.

[2] S. Knoop, S. Vacek, and R. Dillmann, "Sensor fusion for 3D human body tracking with an articulated 3D body model," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2006, pp. 1686–1691.

[3] N. Bellotto and H. Hu, "Multisensor-based human detection and tracking for mobile service robots," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 39, no. 1, pp. 167–181, Feb. 2009.

[4] X. Zhang, X. Chen, J. L. Alarcon-Herrera, and Y. Fang, "3-D model-based multi-camera deployment: A recursive convex optimization approach," *IEEE/ASME Trans. Mechatron.*, vol. 20, no. 6, pp. 3157–3169, Dec. 2015.

[5] G. Huang, A. B. Rad, Y.-K. Wong, and Y.-L. Ip, "Heterogeneous multi-sensor fusion for mapping dynamic environments," *Adv. Robot.*, vol. 21, no. 5–6, pp. 661–688, 2007.

[6] J.-H. Jean and J.-L. Wang, "Development of an indoor patrol robot based on ultrasonic and vision data fusion," in *Proc. IEEE Int. Conf. Mechatron. Autom.*, 2013, pp. 1234–1238.

[7] M. Kobilarov, G. Sukhatme, J. Hyams, and P. Batavia, "People tracking and following with mobile robot using an omnidirectional camera and a laser," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2006, pp. 557–562.

[8] X. Zhang, X. Chen, X. Liang, and Y. Fang, "Distributed coverage optimization for deployment of directional sensor networks," in *Proc. 54th IEEEE Conf. Decision Control*, 2015, pp. 246–251.

[9] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.

[10] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[11] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. British Mach. Vis. Conf.*, Nottingham, Sep. 1–5, 2014, p. 79.

[12] T. Germa, F. Lerasle, N. Ouadah, and V. Cadenat, "Vision and RFID data fusion for tracking people in crowds by a mobile robot," *Comput. Vis. Image Understanding*, vol. 114, no. 6, pp. 641–651, 2010.

[13] M. Wang, Y. Liu, and R. Xiong, "Robust object tracking with a hierarchical ensemble framework," in *proc. IEEE/RSJ Int. Conf. Intelligent Robots Syst*, Daejeon, 2016, pp. 438–445, doi: 10.1109/IROS.2016.7759091.

[14] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3074–3082.

[15] Y. Wu, J. Lim, and M.-H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.

[16] M. Kristan *et al.*, "The visual object tracking VOT2015 challenge results," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2015, pp. 1–23.

[17] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. Van de Weijer, "Adaptive color attributes for real-time visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1090–1097.

[18] S. Liu, T. Zhang, X. Cao, and C. Xu, "Structural correlation filter for robust visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 4312–4320.

[19] T. Liu, G. Wang, and Q. Yang, "Real-time part-based visual tracking via adaptive correlation filters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 4902–4912.

[20] J. Cui, H. Zha, H. Zhao, and R. Shibasaki, "Multi-modal tracking of people using laser scanners and video camera," *Image Vis. Comput.*, vol. 26, no. 2, pp. 240–252, 2008.

[21] C. Dondrup *et al.*, "Real-time multisensor people tracking for human-robot spatial interaction," in *Proc. Mach. Learn. Soc. Robot. Workshop*, 2015, pp. 26–31.

[22] M. Munaro and E. Menegatti, "Fast RGB-D people tracking for service robots," *Auton. Robots*, vol. 37, no. 3, pp. 227–242, 2014.

[23] R. Mahapatra, K. V. Kumar, G. Khurana, and R. Mahajan, "Ultrasonic sensor based blind spot accident prevention system," in *Proc. Int. Conf. Adv.Comput. Theory Eng.*, 2008, pp. 992–995.

[24] I. Ullah, Q. Ullah, F. Ullah, and S. Shin, "Integrated collision avoidance and tracking system for mobile robot," in *Proc. Int. Conf. Robot. Artif. Intell.*, 2012, pp. 68–74.

[25] D. Su and J. V. Miro, "An ultrasonic/RF GP-based sensor model robotic solution for indoors/outdoors person tracking," in *Proc. 13th Int. Conf. Control Autom. Robot. Vis.*, 2014, pp. 1662–1667.

[26] T. Wilhelm, H.-J. Böhme, and H.-M. Gross, "Sensor fusion for vision and sonar based people tracking on a mobile service robot," in *Proc. Int. Workshop Dynamic Perception*, 2002, pp. 315–320.

[27] H. Durrant-Whyte and T. C. Henderson, "Multisensor data fusion," in *Springer Handbook of Robotics*. New York, NY, USA: Springer-Verlag, 2008, pp. 585–610.

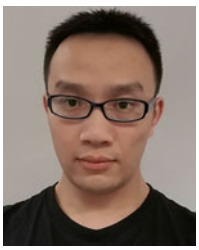
[28] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2005, vol. 1, pp. 886–893.

Mengmeng Wang received the B.S. degree in control science and enginering from Zhejiang University, Hangzhou, China, in 2015, where she is currently working toward the M.S. degree in control science and engineering.

Her research interests include visual object tracking and object detection.

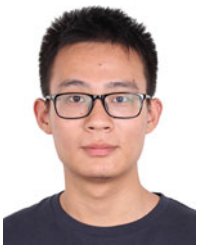
Yong Liu (M'11) received the B.S. degree in computer science and engineering and the Ph.D. degree in computer science from Zhejiang University, Zhejiang, China, in 2001 and 2007, respectively.

He is currently a Professor with the Institute of Cyber-Systems and Control, Department of Control Science and Engineering, Zhejiang University, Hangzhou, China. His research interests include machine learning and robotics vision.



Daobilige Su received the B.Eng. degree in mechatronic engineering from Zhejiang University, Hangzhou, China, in 2010, the M.Eng. degree in automation and robotics from Warsaw University of Technology, Warsaw, Poland, in 2012, and the M.Eng. degree in automation from the University of Genova, Genova, Italy through European Master on Advanced Robotics (EMARO) program in 2012. He is currently working toward the Ph.D. degree at the Centre for Autonomous System, University of Technology Sydney, Sydney, N.S.W., Australia.

His research areas include robotics, SLAM, robot audition, computer vision, and machine learning.



Yufan Liao received the B.S. degree in control science and enginieering from Zhejiang University, Hangzhou, China, in 2017, where he is currently working toward the M.S. degree in control science and engineering.

His research interests include visual object tracking and deep learning.

Lei Shi (M'10) received the B.Eng. degree in electrical engineering and automation from Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004, the M.Eng. degreee in control systems from the University of Sheffield, Sheffield, U.K. in 2005, and the Ph.D. degree in engineering (robotics) from the University of Technology Sydney (UTS), Sydney, Australia, in 2004, 2005, and 2014, respectively.

He is a currently a Research Fellow with the Centre for Autonomous Systems, UTS. His research interests include machine learning, applications of probability and statistics, conditional assessment, and sensors and signal processing.

Jinhong Xu received the B.S. degree in communication engineering from Zhejiang University of Technology, Hangzhou, China, in 2015, where he is currently working toward the M.S. degree in control science and engineering.

His research interests include mobile robots, computer vision, and SLAM.

Jaime Valls Miro (M'14) received his B.Eng. and M.Eng. degrees in computer science systems from the Valencia Polytechnic University, Spain, in 1990 and 1993 respectively. He received his Ph.D. degree from Middlesex University, England, in 1998. His thesis examined the use of dynamics for trajectory planning and optimal control of industrial manipulators. Before joining the Centre for Autonomous Systems at the University of Technology Sydney, he worked for 5 years as a software and control systems analyst in the underwater robotics industry for a London-based company. Currently, he is an Associate Professor with the Centre for Intelligent Mechatronic Systems (CIMS) at UTS. His research activities include autonomous mobile robot navigation and mapping (in particular in unstructured scenarios such as USAR), visual SLAM, assistive robotics, machine learning and human-robot interaction. He is a regular reviewer of scientific works published in the top international robotic journals and conferences.