Real-time 3D Human Tracking for Mobile Robots with Multisensors

Mengmeng Wang[1], Daobilige Su[2], Lei Shi[2], Yong Liu[1] and Jaime Valls Miro[2]

***Abstract*— Acquiring the accurate 3-D position of a target person around a robot provides fundamental and valuable information that is applicable to a wide range of robotic tasks, including home service, navigation and entertainment. This paper presents a real-time robotic 3-D human tracking system which combines a monocular camera with an ultrasonic sensor by the extended Kalman filter (EKF). The proposed system consists of three sub-modules: monocular camera sensor tracking model, ultrasonic sensor tracking model and multi-sensor fusion. An improved visual tracking algorithm is presented to provide partial location estimation (2-D). The algorithm is designed to overcome severe occlusions, scale variation, target missing and achieve robust re-detection. 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 Gaussian Process Regression is used for partial location estimation (2-D). 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 simulation platform and actual mobile robot for various indoor and outdoor scenes. The experimental results show the superior performance of the 3-D tracking system in terms of both the accuracy and robustness.**

I. Introduction

Tracking people in 3-D is a key ability for robots to effectively interact with humans. It is 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, a service robot tracks a specific person in order to provide certain services or to accomplish other tasks in office buildings, museums, hospital environments, or in shopping centers. It is crucial to estimate the accurate positions of the target continuously for subsequent actions. To track the target people across various complex environments, robots need to localize the target and discriminate him/her from other people. In this context, localizing and tracking a moving target become critical and challenging for many indoor and outdoor robotic applications [1]–[3].

Target tracking for mobile robots has been a popular research topic in recent years, and plenty of methods using various sensors have been developed [3]–[6]. Among them, visual tracking enjoys a good population. It is an extremely active research area in computer vision community and obtains significant progress over the past decade [7]–[11]. However, a monocular camera sensor is limited in providing the 2-D position because it is insufficient to measure the range information from the robot to the target. To introduce range information while retaining the advantages of visual tracking, an intuitive solution is to incorporate heterogeneous data from other sensors [5], [6], [12].

In this paper, we propose a new method for tracking the 3-D positions of a person by multi-sensors in both indoor and outdoor environments with a robotic 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. Above all, the contributions of our method are summarized 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 online visual tracking algorithm is presented to handle the situations of severe occlusion, object missing and re-detection;
3) The estimated 3-D information is further exploited to improve the scale accuracy of the target in the image coordinate;

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

II. Related Work

Our work is related to some specific research areas in computer vision and robotics, which are visual tracking, ultrasonic tracking and 3-D location estimation. We will give a brief exposition for each of them in this section.

A. Visual Tracking using Monocular Camera

Numerous visual tracking algorithms have been developed over the past few decades [7] [11] [8] [13]. Recently, a group of correlation filter based discriminative trackers have made remarkable improvement in visual tracking field [8]

[1]Mengmeng Wang and Yong Liu are with the State Key Laboratory of Institute of Cyber-Systems and Control, Zhejiang University, Zhejiang, 310027, China. Yong Liu is the corresponding author of this paper, email: yongliu@iipc.zju.edu.cn

[2]Daobilige Su, Lei Shi, and Jaime Valls Miro are with the Centre for Autonomous Systems (CAS), The University of Technology, Sydney, Australia.

[16] [9] [17] [18]. Considering the nature of visual tracking, the correlation filter can be solved in the Discrete Fourier Transform (DFT) effectively and efficiently. These methods are excellent in many environments but they are not suitable for the 3-D human tracking in a robot platform because they are not robust enough in the situations of severe occlusions and object missing. In this paper, an improved correlation filter based visual tracking algorithm was developed to provide enhanced robustness and performance in the application of mobile robots.

An exhaustive analysis is beyond this work. Thus we recommend [14] and [15] for a full understanding about this problem.

B. Ultrasonic Tracking

Ultrasonic sensors have been used extensively as time-of-flight range sensors in localizing the tracking targets [19]–[21]. However, one disadvantage of this type of sensor is that when the target moves at the vertical direction of the sonar beam, the calculated locations are usually inaccurate. Another problem with sonar sensors is the reflection from obstacles in the environment will usually cause invalid and incorrect results. Furthermore, relying on the time-of-flight measurement only, the receiver is unlikely able to discriminate multiple sources which means that the system does not work when multiple targets present.

C. Robotic 3-D Human Tracking

There are a number of different techniques to track persons with mobile robots. Using laser range finders with cameras for person tracking is an option in the robotics community [3], [6], [12]. Laser scans could be used to detect the human legs at a fixed height. However, this can not provide robust features for discriminating the different persons in the robot's vicinity, thus the detector always tends to fail when one leg occludes the other.

Combining the sonar sensors with cameras is another popular research direction [4], [5], [22]. They usually use the sonar sensors to detect the regions that might contain the target in the sonar's field of view. Corresponding regions in the images are then used as additional measurements.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.

3-D features from 3-D point clouds reconstructed by RGB-D sensors are used for 3-D human tracking as well [1], [2], [23], [24]. 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 human tracking applications.

III. Method

The proposed 3-D tracking system can be decomposed into three sub-modules, monocular camera sensor tracking model, ultrasonic sensor tracking model and multi-sensor fusion. In this section, the details about these three sub-modules are presented.

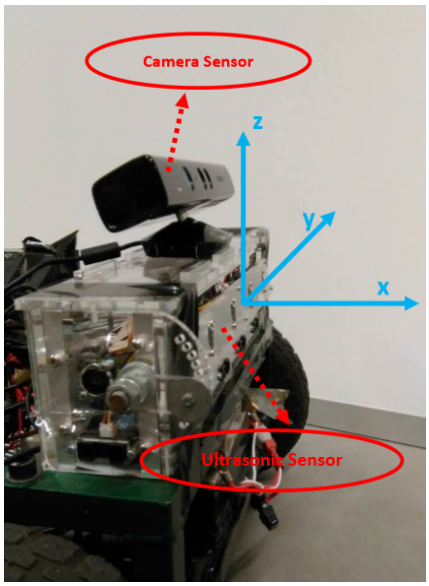


Fig. 1: The local robot coordinate. We employ Kinect XBOX 360 to acquire the ground truth of the 3-D position of the target during the tracking process in the experiment. Simultaneously, the RGB camera of Kinect is used as our monocular camera sensor for convenience.

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 k-th time instant. All the variables are defined in the robot's local coordinate frame as shown in Fig.1. The target people is sensed by both vision sensor and ultrasonic sensor. To estimate the 3-D position of the people, data acquired from the two sensors are fused sequentially using an EKF.

A. Monocular Camera Sensor Tracking Model

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 shown in Eq. 1.

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

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

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

For conventional visual tracking, the target is given in the first frame either from human annotation or 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 people's height h and the initial 3-D position $\mathbf{x}_{init}$. Additionally, we assume the average width of a person is 0.4 meters and the distance from the sonar emitter to the people's feet is 50% of his/her height h in all experiments.

Then the initial 3-D positions of left boundary $\mathbf{x}_{linit}$, right boundary $\mathbf{x}_{rinit}$, head $\mathbf{x}_{hinit}$ and feet $\mathbf{x}_{finit}$ of the target can be calculated by

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

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

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

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

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

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

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

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

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

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

The presented visual tracking algorithm is based on the Kernelized Correlation Filter (KCF) [8] tracker. We extend it with a novel criterion to evaluate the performance of the tracking results and develop a new scale estimation method which 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 detailedly in [8]. The goal is to learn an online correlation filter from 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 which minimizes the squared error over samples $\mathbf{s}_{w,h}$ and their regression labels $y_{w,h}$,

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

where φ denotes the mapping to non-linear feature space with kernel κ and the dot-products 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 some circulant kernels, the solution $\mathbf{w}$ can be represented as $\mathbf{w} = \sum_{w,h} \alpha_{w,h} \varphi\left(\mathbf{s}_{w,h}\right)$, 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 Eq. 5 and the target appearance model $\mathbf{s}$ is 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: Severe occlusion and missing target are two significant challenges in visual tracking. As mentioned above, the KCF tracker cannot avoid the model corrupting due to the lack of the feedback from the tracking results.

The response map is the correlation response used to locate the position of target as in Eq. 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, resulting in a pattern that is significantly different from the normal response map as shown in Fig. 2. Instead of reporting a target regardless of the response map pattern, we propose a novel criterion for severe occlusion while remaining the advantages of KCF.

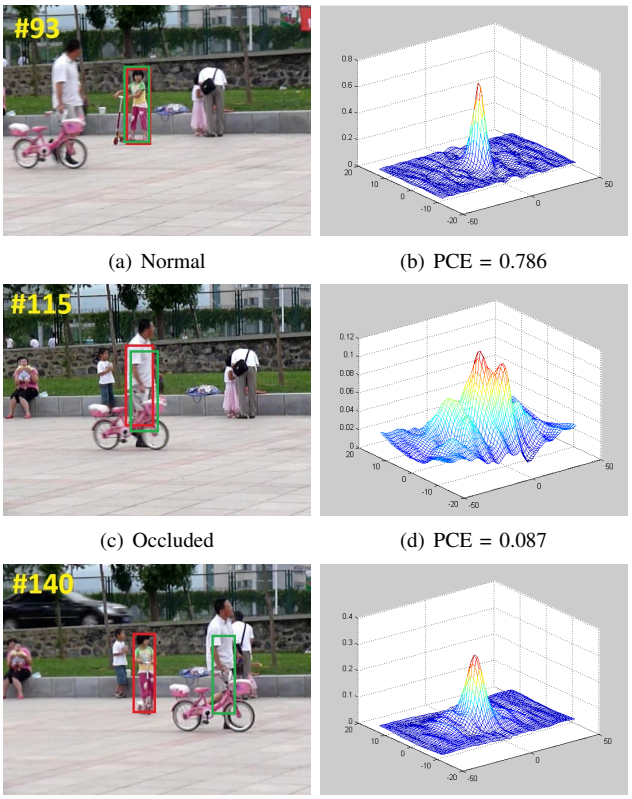
For correlation filter based classifier, the peak-to-sidelobe ratio (PSR) can be used to quantify the sharpness of the correlation peak. However, PSR is still not robust to partial occlusions [18]. Therefore, we propose a novel criterion called peak-to-correlation energy (PCE) as described in Eq. 7.

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

where $|y_{\max}|$ denotes the maximum peak magnitude, and the correlation response map energy E_y is defined as Eq. 8.

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

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



(a) Normal (b) PCE = 0.786

(c) Occluded (d) PCE = 0.087

(e) Re-detected (f) PCE = 0.671

Fig. 2: The first column shows the original frames from the vision sensors, 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 PCE, the target girl is re-detected 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.

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_{2D} in the 2-d visual tracking process, we employ DSST [9] algorithm. Firstly, the position of the object is determined by the learned translation filter with abundant features. Secondly, a group of windows with different scales are 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_{2D}. At the meantime, the standard variance σ_{2D} from s_{2D} is calculated as the uncertainty of s_{2D}.

We also consider the scale states calculated from the 3-D position estimations. At the k-th 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 Eq.2(c)(d) 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 Eq. 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

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

where v_{init} is the initial height of the target calculated by Eq. 3d. Finally, the scale s_k of the k-th frame is calculated using Eq.10.

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

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

B. Ultrasonic Sensor Tracking Model

Traditional sonar array sensors use time-of-flight (TOF) and triangulation to find the relative location of a target with respect to the source. In the proposed tracking system, the Gaussian Process Regression (GPR) techniques are used in sonar sensor tracking model to obtain the range information and improve the predicted accuracy of the tracking target [21].

The active sonar emitter array which consists of three sonar sensors is designed as a human carrying Portable User Device (POD). The corresponding passive sensor array with four sonar units is attached equally spaced in front of the robot. When the RF module on the robot receives the RF signal from the POD, it will start a timer. Then the time lapsed from when the timer starts until all the sonar units measure an incoming signal is the corresponding TOF.

For each sonar unit, GPR model trained with real data is built to predict sensor reading with corresponding covariance at a certain (x_{Uk}, y_{Uk}) location, where the subscript Uk denotes the k-th ultrasonic state. The final posterior probability for prediction is calculated by the Eq.11. The position with highest probability is chosen as the predicted location.

$$P(x_{Uk}, y_{Uk} | U_1, U_2, U_3, U_4) = \frac{\prod_{i=1}^{4} P(U_i | x_{Uk}, y_{Uk}) P(x_{Uk}, y_{Uk})}{P(U_1, U_2, U_3, U_4)} \tag{11}$$

where $P(U_i | x_{Uk}, y_{Uk})$, $i = 1, \ldots, 4$ are the learned GPR models for 4 sonar units. $P(x_{Uk}, y_{Uk})$ is the prior inference for unknown position (x_{Uk}, y_{Uk}) and is assumed as a uniform distribution. As the U_1 to U_4 are the actual observations, $P(U_1, U_2, U_3, U_4) = 1$.

Transform to the k-th 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

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

$$x_{Uk} = \sqrt{x_k{}^2 + z_k{}^2} \tag{12b}$$

$$y_{Uk} = y_k \tag{12c}$$

C. Multi-Sensor Fusion

A standard EKF approach is utilized to fuse the measurements obtained from the ultrasonic sensor and the vision sensor. As the update frequencies of the two sensors are different, the fusion algorithm will be triggered whenever any of them is updated. We adopt such a method to sequentially process the multiple, heterogeneous measurements arriving in an asynchronous order [25].

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,

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

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

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

where $\mathbf{P}_k$ is the covariance matrix, $\mathbf{G}$ is the Jacobian of Eq. 13a (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 the $(t_k - t_{k-1})$ by a constant noise level $\mathbf{R}$.

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

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

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

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

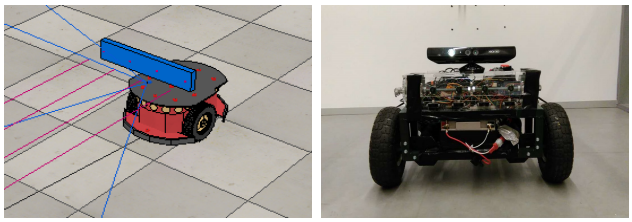
where the $*$ in the subscript stands for either ultrasonic (U) or camera (C) measurement, we collectively call it measurement next. $\mathbf{h}_*(\mathbf{x}_k)$ is the sensor model which provides the predicted measurement. The camera sensor model $\mathbf{h}_C(\mathbf{x}_k)$ is define in Eq.1 and the sonar sensor model $\mathbf{h}_U(\mathbf{x}_k)$ is defined in Eq.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 Eq.6 and $\mathbf{z}_{Uk}$ is the mean predicted by GPR in Eq.11. $\mathbf{Q}_{*k}$ is the measurement noise. For sonar sensor, $\mathbf{Q}_{Uk}$ is the covariance matrix from Gaussian process. For camera sensor, $\mathbf{Q}_{Ck}$ is defined as

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

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 GPR.

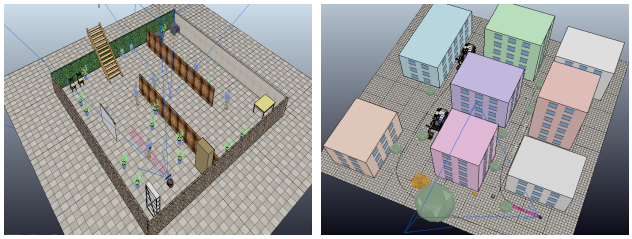
IV. Experiments

To evaluate the performance of our multi-sensor 3-D human tracking system, various experiments were carried out 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. The actual robot platform is named Garden Utility Transportation System (GUTS), shown in Fig.3b. It is a differential mobile robot system fitted with an auto-tipping mechanism [21]. The detailed tracking processes of all the experiments are shown in our video demo.



(a) Simulation (b) GUTS

Fig. 3: (a) Simulated and (b) Physical robot platforms. The Kinect on the GUTS is used to obtain the ground truth of the human as well as the monocular camera sensor.



(a) Indoor (b) Outdoor

Fig. 4: Simulation scenes.

A. Experimental Setup

All our experiments are performed using MATLAB R2015a on a 3.2 GHz Intel Core i7 PC with 16 GB RAM. The Robot Operating System (ROS) has been employed as the software framework for the GUTS platform, linked to MATLAB via the MATLAB-ROS bridge package. The visual tracking algorithm runs at an average speed of 25 FPS and the GPR prediction of sonar runs at about 5 Hz. The ultrasonic sensor tracking algorithm remains the same as described in our previous work [21]. The setup of the visual tracking algorithm is introduced below.

Multichannel features based on Histogram of Oriented Gradient (HOG) [26] with a cell size of 4 pixels, as well as color names (CN) [16] with the first 2 dimensions, are used in our method. 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 Eq.4 is set to 0.0001.

In order to demonstrate the performance of the proposed 3-D tracking system, we test it in the simulation experiments and the GUTS robotic platform in both indoor and outdoor

TABLE I: The arithmetic mean error in three axis. *S* is short for simulation, *G* for GUTS, *I* for indoor, *O* for outdoor, *C* for camera sensor, and *U* for ultrasonic sensor.

Axis	SI(C+U)	SO(C+U)	SO(C)	SO(U)	GI(C+U)	GO(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

environments. To illustrate that under normal walking paces and patterns the proposed tracking system is able to effectively track the target people, we apply a simple proportional controller in translation and orientation velocity to make the robot track automatically.

B. Experimental Results

The simulated robotic platform is depicted in Fig.3a. A passive sensor array with four sonar receiver units is mounted equally spaced in front of the robot. The camera is fixed below the sonar array. Indoor scene is constructed as an office room with many people walking inside it as shown in Fig.4a. Outdoor scene in Fig.4b is built with plenty of buildings, trees, vehicles and people to imitate a city area.

To validate the necessity of the two sensors, we compare the tracking results with individual sensor respectively only in the simulation outdoor scene for safety. Without the monocular camera sensor, the estimated accuracies of z axis is dramatically reduced. Without the sonar sensor, the estimation of x axis is incorrect 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.

To evaluate the performance of the proposed 3-D tracking system in reality, more experiments were conducted with the GUTS platform as shown in Fig. 3b. We introduce the skeleton tracking of Kinect XBOX 360 to collect the ground truth of the 3-D positions of the target during tracking process through the OpenNI tracker in ROS. The position of the waist in the skeleton is regarded as the true position of the sonar POD carried by the person. Simultaneously, the RGB camera on the Kinect is used as our monocular camera sensor. As shown in Fig.5, the indoor experiment is performed in the common corridor of our laboratory while the outdoor experiment is conducted outside the main building of University of Technology Sydney.

There are many challenges in these scenes such as illuminate variations, scale variations, part occlusions, severe occlusions, background clutters and object missing. The target people is walking with the variations in all three axes to make the 3-D estimation more challenging. We show these challenges in our video demo. The quantitative 3-D tracking results are shown in Fig. 6a,b for simulation scenes, Fig. 6c,d for real-world scenarios.

The results illustrate a great performance of our method. As shown in Fig. 6a,b for the indoor and outdoor scenes of the simulation experiments, the black lines represent the ground truth of the target motions in three axes, the green 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. Also, in Fig. 6c,d for the indoor and outdoor experiments in GUTS, the red lines show the ground truth of the target motions while the blue ones denote the estimation of our method. We calculate the mean errors of all the experiments in TABLE I.

(a) Indoor (b) Outdoor

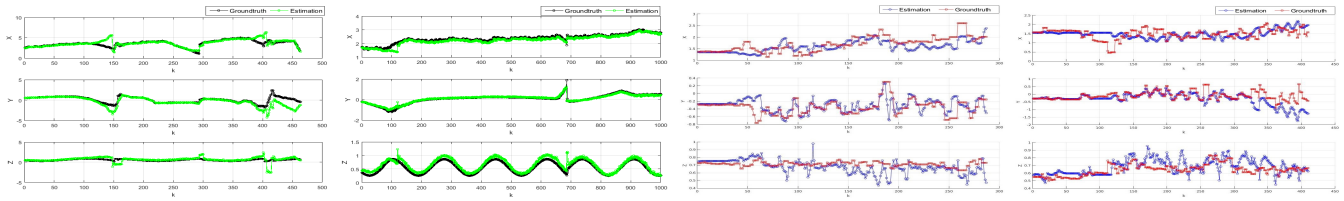
Fig. 5: Real world scenarios.

V. Conclusion

In this paper, we address the problem of accurately estimating the 3-D position of the target around 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 re-detection. 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.

Acknowledge

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 Natural Science Foundation of Zhejiang Province under Grant LR13F030003.



(a) Simulated Indoor Scene (b) Simulated Outdoor Scene (c) Actual Indoor Scene (d) Actual Outdoor Scene

Fig. 6: 3-D tracking results of simulation and actual experiments. Results are best viewed on high-resolution displays.

REFERENCES

[1] A. P. Gritti, O. Tarabini, J. Guzzi, G. A. Di Caro, V. Caglioti, L. M. Gambardella, and A. Giusti, "Kinect-based people detection and tracking from small-footprint ground robots," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4096–4103, IEEE, 2014.

[2] S. Knoop, S. Vacek, and R. Dillmann, "Sensor fusion for 3d human body tracking with an articulated 3d body model," in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pp. 1686–1691, IEEE, 2006.

[3] N. Bellotto and H. Hu, "Multisensor-based human detection and tracking for mobile service robots," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 167–181, 2009.

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

[5] J.-H. Jean and J.-L. Wang, "Development of an indoor patrol robot based on ultrasonic and vision data fusion," in *2013 IEEE International Conference on Mechatronics and Automation*, pp. 1234–1238, IEEE, 2013.

[6] M. Kobilarov, G. Sukhatme, J. Hyams, and P. Batavia, "People tracking and following with mobile robot using an omnidirectional camera and a laser," in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pp. 557–562, IEEE, 2006.

[7] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 7, pp. 1409–1422, 2012.

[8] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.

[9] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *British Machine Vision Conference, Nottingham, September 1-5, 2014*, BMVA Press, 2014.

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

[11] M. Wang, Y. Liu, and R. Xiong, "Robust object tracking with a hierarchical ensemble framework," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pp. 438–445, IEEE, 2016.

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

[13] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3074–3082, 2015.

[14] Y. Wu, J. Lim, and M.-H. Yang, "Object tracking benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, 2015.

[15] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernandez, T. Vojir, G. Hager, G. Nebehay, and R. Pflugfelder, "The visual object tracking vot2015 challenge results," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 1–23, 2015.

[16] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. Van de Weijer, "Adaptive color attributes for real-time visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1090–1097, 2014.

[17] S. Liu, T. Zhang, X. Cao, and C. Xu, "Structural correlation filter for robust visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4312–4320, 2016.

[18] T. Liu, G. Wang, and Q. Yang, "Real-time part-based visual tracking via adaptive correlation filters," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4902–4912, 2015.

[19] R. Mahapatra, K. V. Kumar, G. Khurana, and R. Mahajan, "Ultra sonic sensor based blind spot accident prevention system," in *2008 International Conference on Advanced Computer Theory and Engineering*, pp. 992–995, IEEE, 2008.

[20] I. Ullah, Q. Ullah, F. Ullah, and S. Shin, "Integrated collision avoidance and tracking system for mobile robot," in *Robotics and Artificial Intelligence (ICRAI), 2012 International Conference on*, pp. 68–74, IEEE, 2012.

[21] D. Su and J. V. Miro, "An ultrasonic/rf gp-based sensor model robotic solution for indoors/outdoors person tracking," in *Control Automation Robotics & Vision (ICARCV), 2014 13th International Conference on*, pp. 1662–1667, IEEE, 2014.

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

[23] C. Dondrup, N. Bellotto, F. Jovan, M. Hanheide, *et al.*, "Real-time multisensor people tracking for human-robot spatial interaction," 2015.

[24] M. Munaro and E. Menegatti, "Fast rgb-d people tracking for service robots," *Autonomous Robots*, vol. 37, no. 3, pp. 227–242, 2014.

[25] H. Durrant-Whyte and T. C. Henderson, "Multisensor data fusion," in *Springer Handbook of Robotics*, pp. 585–610, Springer, 2008.

[26] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893, IEEE, 2005.