Unifying Obstacle Detection, Recognition and Fusion Based on Millimeter Wave Radar and RGB-Depth Sensors for the Visually Impaired

Ningbo Long, Kaiwei Wang,[a)] Ruiqi Cheng, Weijian Hu and Kailun Yang

State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, Hangzhou, China

(Received XXXXX; accepted XXXXX; published online XXXXX)
(Dates appearing here are provided by the Editorial Office)

Abstract:
It is very difficult for visually impaired people to perceive and avoid obstacles at a distance. To address this problem, the unified framework of multiple target detection, recognition and fusion is proposed based on the sensor fusion system comprised of a low-power MMW radar and an RGB-D sensor. In this paper, Mask R-CNN and SSD network are utilized to detect and recognize the objects from color images. The obstacles depth information is obtained from the depth images using the MeanShift algorithm. The position and velocity information of the multiple target are detected by the millimeter wave radar based on the principle of frequency modulated continuous wave. The data fusion based on the Particle Filter obtains more accurate state estimation and richer information by fusing the detection results from the color images, depth images and radar data compared with using only one sensor. The experiment results show that the data fusion enriches the detection results. Meanwhile, the effective detection range is expanded compared to using only the RGB-Depth sensor. Moreover, the data fusion results keep high accuracy and stability under diverse range and illumination conditions. As a wearable system, the sensor fusion system has the characteristics of versatility, portability and cost-effectiveness.

1. Introduction

According to the data from the World Health Organization (WHO), 253 million people are estimated to be visually impaired worldwide, and 36 million are blind [1]. It is very difficult for visually impaired people (VIP) to perceive and avoid obstacles at a distance. To address this problem, we propose a sensor fusion system which unifies the multiple target detection, recognition and fusion functions based on the cost-effective RGB-Depth (RGB-D) sensor and low-power millimeter wave (MMW) radar sensor.

Over the past years, computer vision (CV) has undergone a striking improvement especially because of the development of the deep learning, which has been an enormous benefit for the VIP to access, understand and explore surrounding environments [2-4]. The development of the CV has a close relationship with the stereo vision sensor. The RGB-D sensor has also received rising attention and been used widely because of its outstanding performance [5-7]. They provide much more information compared the traditional assistive tools, which are able to acquire color information and perceive the environment in three dimensions at video frame rates. Auxiliary approaches based on RGB-D sensors have been investigated to help VIP to avoid obstacles [2,8-11]. However, the RGB-D sensors, including light-coding sensors, time-of-flight sensors (TOF camera), and stereo cameras, could not solve the problems of remote obstacle detection and velocity detection perfectly. The detection range of low power light-coding sensors is too small in outdoor environment, especially in sunny environment [5]. The measurement results of TOF camera are sensitive to ambient light and show poor performance in outdoor environments [12]. The ranging results of remote objects derived from stereo cameras are not accurate, and the remote objects without texture are not robustly detected [10]. Based on these observations, it is difficult for one to measure the velocity of object using all these kinds of RGB-D sensors.

In contrast, the range and velocity of the obstacles could be calculated at the same time with the help of the MMW radar, and the accuracy of the range is very high, e.g., several centimeters. The radar detection results are rarely influenced by the varying illuminance and severe weather [13]. Meanwhile, the detection range could be very large compared with the RGB-D sensor. Thanks to the technological development [14], the MMW radar sensors have become small, low-cost and accurate, which makes them especially suitable for portable low-power applications. Moreover, the single-chip radar sensor has already appeared [14]. However, the MMW radar has its own drawbacks, such as, the azimuth beam with of the MMW radar always covers more than several degrees due to the limited antenna distributions, which results in a low directional resolution compared with the camera.

[a)]Author to whom correspondence should be addressed. Electronic mail: wangkaiwei@zju.edu.cn

There are various advantages of fusing RGB-D sensor and radar sensor [14-18]. The fusion mutually complements the drawbacks of each sensor and maximizes the capability of object detection location and recognition in varying environments. The MMW radar provides relatively high distance resolution and velocity information of every detected object in the scene. Meanwhile, the RGB-D sensor provides relatively high spatial resolution, the depth images that although suffer from ineffectiveness in some conditions, and the color images which is used to achieve the object recognition by using a single deep neural network, namely Single Shot MultiBox Detector (SSD) [19] or the object instance segmentation based on the Mask R-CNN [20]. At the same time, the sensor fusion system increases the overall system robustness to varying lighting conditions.

In this paper, we propose a unified target detection, recognition and fusion framework based on the sensor fusion system which is comprised of a low-power MMW radar and an RGB-D sensor, as Fig.1 shows. Obstacles in the scene are detected by the radar and RGB-D sensor simultaneously. The range, velocity and angle information of objects are obtained by the MMW radar based on the principle of frequency-modulated continuous wave (FMCW) [21]. The MeanShift [22] algorithm is applied onto the depth images to achieve feature extraction, then the depth and position information of the obstacles are achieved. The objects in the color images are recognized by the method of deep learning, herein we adopt the SSD or the Mask R-CNN neural network. The effective information coming from the radar, color images and depth images are fused by the means of Particle Filter [23]. After that, we have the detailed information about the current scene including the objects' class, position and velocity. The non-semantic stereophonic interface is leveraged to transfer the detection information to VIP. The sensors hardware system will be elaborated in the section 3.1.

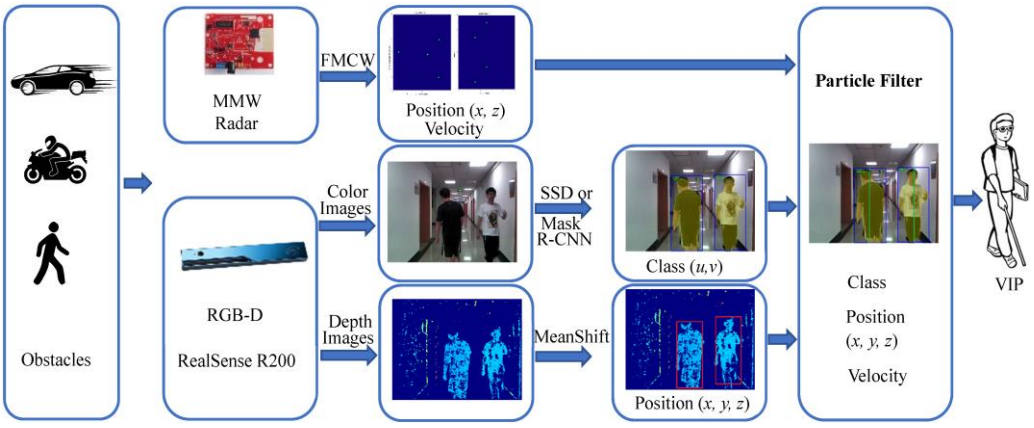


FIG. 1. The proposed multiple target detection, recognition and fusion framework for VIP assistance.

In addition to technical aspects, some other factors should be considered in the VIP assistance domain, such as the price, dimension, weight and energy consumption of the assistive system. The main purpose of our research in this paper is to design a robust and cost-effective multiple object detection recognition and fusion system to help VIP perceive and avoid obstacles.

The remainder of this paper is structured as follows. In section 2, the related work is reviewed. Section 3 describes the specific methodology, including the sensors hardware, the objects detection and recognition principle, the calibration between the different sensors and the data fusion method. Section 4 presents the experiments results. Last, section 5 draws the conclusions and gives an outlook to future work.

2. Related work

In the literature, some approaches have been proposed to help VIP perceive and avoid obstacles at a distance.

Jindal proposed an obstacle detection method [24] for VIP by ground plane removal using speed-up robust features and gray level co-occurrence matrix. He presented the design of a smart phone based cost-effective system to guide VIP to walk safely on the roads by detecting obstacles in real-time scenarios. Monocular vision was used to capture the video which was pre-processed by removing motion blurriness. Then, SURF feature is used for matching localized points. Different ROI areas from the image are segmented out using an active contour model. Finally, the classification of these ROIs as obstacles and non-obstacles was done by calculating texture features.

Kaur [25] presented the development of a real-time system based on detection, classification and position estimation of objects in an outdoor environment to provide the VIP with a voice output-based scene perception. An odroid board integrated with an USB camera and laser was utilized and the system was low-cost, light weight, simple and easily wearable. In this paper, the object detection framework was exploited which used a multimodal feature fusion-based deep learning architecture using edge, multiscale as well as optical flow information. At last, the experiments were carried out using PASCAL VOC 2007 dataset, Caltech dataset as well as captured real-time data.

Kiuru [26] presented a clinical investigation results of an assistive device that uses radar technology. The radar device detected objects in the environment and conveyed this information to the users by the way of sound or vibration feedback. There were 25 VIP participated the two-week period of the investigation which included a training session as well as opening and closing interviews. At last, the results indicated the device improved the ability to perceive environment and increased their confidence in independent mobility.

Yang put forward seizing pixel-wise semantic [27] segmentation to cover navigation related perception needs

in a unified way. They had integrated the approach in a wearable navigation system by incorporating robust depth segmentation. They also presented a closed-loop field test involving real visually-impaired users, demonstrating the effectivity and versatility of the assistive framework.

Aladren [28] proposed a new system for NAVI (Navigation assistance for visually impaired). A consumer RGB-D camera was chosen which the range and visual information were utilized. The depth information was enhanced with the long-range visual information. The proposed system had been tested on a wide variety of scenarios and data sets, giving successful results and showing that the system was robust and worked in challenging indoor environments.

There is almost none specialized radar and stereo vision fusion system for VIP assistance. Although plenty of related work [16,18,29-35] have been done to achieve objects detection by fusing the radar and the stereo vision system, most of them are part of automobile ADAS.

Ćesić [16] presented a paper addressed detection and tracking of moving objects within the context of ADAS. He used a multi-sensor setup consisting of a radar and a stereo camera mounted on top of a vehicle. The extended Kalman filter on Lie groups was employed to solve the problem of data fusion. To solve the multitarget tracking problem, the JIPDAF (Joint Integrated Probabilistic Data Association Filter) was used. At last, the proposed approach was tested on a real-world dataset collected with the described multi-sensor setup in urban traffic scenarios.

Kim [18] developed a firefighting robot that used sensor fusion between stereo thermal infrared (IR) vision and FMCW radar to locate objects through zero visibility smoke in real-time. The stereo IR vision was used to obtain 3D information about the scene while the radar provided more accurate distances of objects in the field of view. The system was sufficiently fast to provide real-time matching of objects in the scene allowing for dynamic reaction object tracking and locating.

Kim presented a multiple-object tracking system [31] whose design was based on multiple Kalman filters dealing with observations from a CCD camera and a cheap radar module. The integrated probability data association (IPDA) was used to achieve the multi-object tracking. At last, the proposed complementary system was experimentally evaluated through a multi-person tracking scenario.

Compared with these works, the main advantages of our system are summarized as follows:

- Our system unifies the multiple target detection, recognition and fusion to provide navigation assistance for the VIP.
- The framework is specially optimized for the VIP and achieved based on the low-power MMW radar and the RGB-D sensor, which have the characteristics of small size, low energy consumption and cost-effectiveness.
- The sensor fusion enhances the overall system robustness to varying conditions.

3. Methodology

In this section, the system hardware configuration is described in detail firstly. Then the MMW radar detection principle based on the FMCW, the feature extraction on the depth images and the object recognition on the color images are introduced. After that, the calibration between the radar and RGB-D sensor is accomplished. At last, the data fusion is presented.

3.1. Sensors

In our sensor fusion system, the Intel RealSense R200 stereo vision system [2] and the TI short range MMW radar [14] are utilized, as shown in Fig. 2(a). They are mounted on a frame fabricated by 3D printing and their positions are fixed. The sensors are closely spaced at about the same plane, while the sensor fusion is performed at the object detection level [15,36]. The current fixed mode is sufficiently precise to achieve the data fusion.

The RealSense R200 is composed of two infrared cameras (right and left), an infrared laser projector, a color camera and an image processor [5], as illustrated in Fig. 2(b). The color camera image resolution is 1920×1080 pixels with rolling shutter, which is used to acquire color images. The infrared laser projector projects static non-visible near-infrared patterns on the scene, then the patterns are acquired by the right and left infrared cameras. The depth images are generated by the image processor through the embedded stereo-matching algorithm. When the active projecting has no effect, the depth images could also be generated through the passive stereo matching. With the principle of active projecting and passive stereo matching, the performance of RealSense R200 is excellent under indoor and outdoor circumstances [5]. The R200 is quite suitable for VIP navigation because of its environment adaptability, small size and cost-effectiveness.

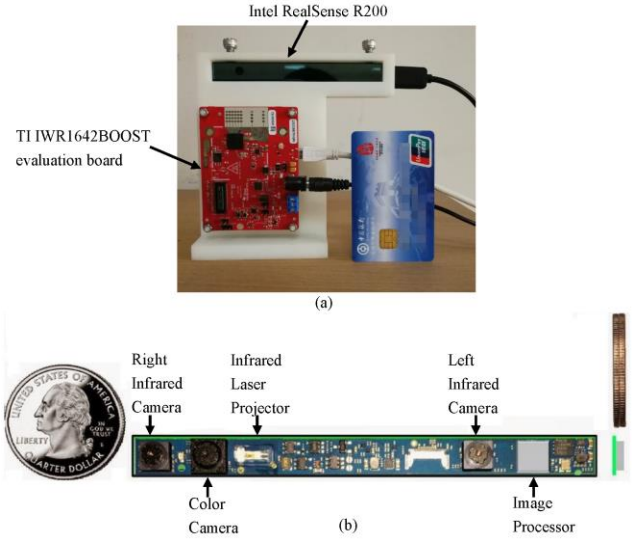


FIG. 2. The hardware configuration. (a) Experimental platform equipped with the RealSense R200 and the TI short range radar evaluation board. The size of the R200 and radar evaluation board is so small that they are portable, and a credit

card is placed nearby for comparison. (b) The RealSense R200 module is shown without reinforcement frame, it includes two infrared cameras (right and left), an infrared laser projector, a color camera and an image processor.

However, the R200 ranging accuracy is reduced when the depth exceeds the general detectable range of 650-2100 mm [5] and the measurement results are influenced by the varying environments and weather conditions. On the contrary, the range accuracy of the MMW radar based on the principle of FMCW is high and the measurements results are stable. The TI short range MMW radar based on the single radar chip IWR1642 is employed in our assistance system. The IWR1642 is an integrated single-chip MMW sensor based on FMCW radar technology capable of operating in the 76 to 81 GHz band with up to 4 GHz continuous chirp. It is an ideal solution for low-power, self-monitored, ultra-accurate radar systems in the industrial and consumer electronics applications.

As the VIP generally belong to low income group, it is necessary to consider the cost. The price of the R200 developer kit in the Intel official website is 79USD, and the price of R200 module without reinforcement frame may be cheaper. The price of the IWR1642 MMW radar chip is about 19.97USD in the TI official website. When the usage amount is bigger, the price is a possibility of decline. Added some other material costs, the total cost of the sensor fusion system is about 110USD, which is very cheap for that low-income group.

Before the data fusion, the sensor features and the field of view (FOV) also need to be considered. The horizontal and the vertical FOV of the color camera is 70° and 43° respectively. The experiments are carried out at the resolution of 640×480 pixels in order to improve the efficiency. Meanwhile, in order to fit the color image and project the depth information into the corresponding color image, the resolution of the depth image is set to be 640×480. The depth images are obtained using the two infrared cameras. And the horizontal and the vertical FOV of the infrared camera is 59° and 46° respectively. The color and depth images are recorded in the auto-exposure mode of the camera. In contrast, the FOV of the MMW radar is ±60° with angular resolution of approximately 15°.

3.2. Objects Detection and Recognition

In this subsection, we mainly introduce the radar detection principle based on the FMCW, the feature extraction on the depth images using the MeanShift algorithm and the object recognition on the color images through the SSD network or the instance segmentation based on Mask R-CNN.

3.2.1. Radar detection principle

FMCW is a technique that obtains range and velocity information from a radar by the way of frequency modulating a continuous signal [21]. The frequency modulation takes many forms, and the linear frequency modulation is the most commonly used. The basic principle of the sawtooth modulation is illustrated in Fig. 3.

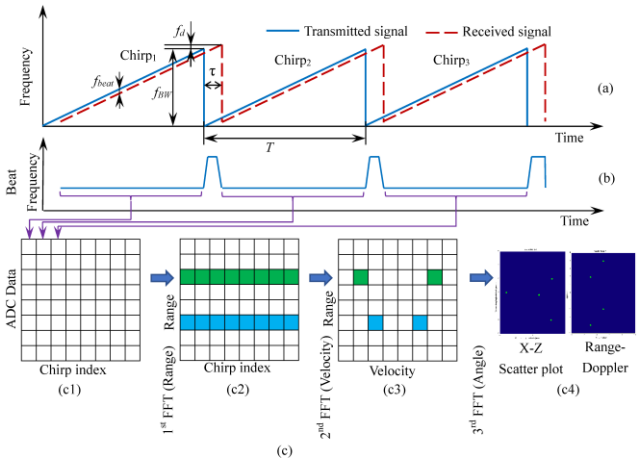


FIG. 3. The basic principle of the FMCW radar with the sawtooth shape modulation. (a) The transmitted and received signal. (b) The corresponding beat frequency. (c) The beat signal processing flow.

The transmitted signal is frequency modulated by a periodic saw-wave [37]. The received signal is similar to the transmitted one but subject to the frequency shift (i.e., Doppler shift, f_d) and the time delay (i.e., τ), as illustrated in Fig.3(a). The f_{BW} is the modulation bandwidth, and T is the modulation period. The frequency difference between the transmitted and received signal is called "beat frequency", which carries the range and velocity information, as shown in Fig.3(b). For saw-wave modulation, the frequency shift and beat frequency are coupled and are difficult to separate for multiple objects [34]. We obtain the range, velocity and angle information of different objects by a special processing flow, as shown in Fig.3(c).

We acquire a number of chirps' beat signal, organize them in a matrix, where each column contains a single sweep beat signal. This single chirp beat signal is processed using a Fast Fourier Transform (FFT) in order to separate the different range objects. Fourier transform processing results in a frequency spectrum that has separate peaks and each peak denotes the presence of an object at a specific distance. This processing is called the range-FFT. Then, a FFT on the sequence of phasors corresponding to the range-FFT peaks outputs the velocity information, which is called the Doppler-FFT. Angle estimation is based on the phase change in the peak of the range-FFT or Doppler-FFT because of differential distance from the object to each of the antennas, which requires at least 2RX antennas. Similarly, a FFT on the sequence of phasors corresponding to the 2D-FFT (range-FFT and Doppler-FFT) peaks resolves the angle estimation problem. This is called angle-FFT.

After these processing, the range, velocity and angle information of objects are obtained. In addition to these above issues, some other problems need to be considered, such as, the pre-processing of the raw ADC data, the constant false alarm rate (CFAR) [38,39], and so on. However, these topics are out of scope of this paper and not discussed.

3.2.2. Objects detection on depth images

Compared with the ordinary digital image processing on the color image, this paper achieves the objects detection function on the depth image produced by the RealSense R200. The obstacles are detected in indoor and outside environments, the color images are shown in Fig.4(a1-a4), and the depth images and the detection results are presented in Fig.4(b1-b4). The detection results are indicated using the red bounding box. Herein, we use MeanShift algorithm [22] to detect objects with the help of depth differences in the depth images and achieve it using OpenCV.

The distance of the detection object is decided by the average depth in the red bounding box. The detection object coordinates in the pixel coordinate system are the center of the red bounding box. Then, we can get the specific coordinates in the camera coordinate with the help of the camera intrinsic.

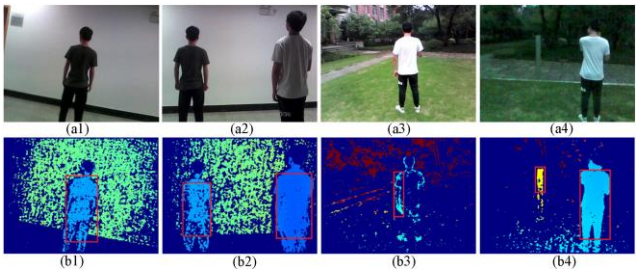


FIG. 4. The object detection on the depth images. (a1-a4) The color images, the objects are detected in indoor and outdoor environments. (b1-b4) The depth images and the detection results, the objects are indicated using the red bounding box.

3.2.3. Object recognition on color images

In addition to the information from the MMW radar and the depth images, we can also achieve the object recognition on the color images. As Fig. 5 shows, the objects are detected in indoor and outdoor environments, while the color images are shown in (a1-a4), and the objects recognition results based on the SSD network are presented in (b1-b4). The detection results are labeled using the bounding box. Similarly, the objects recognition results based on the instance segmentation using the Mask R-CNN are shown in (c1-c4).

SSD [19] is simple compared to methods that require object proposals because it completely eliminates proposal generation and subsequent pixel or feature resampling stages and encapsulates all computation in a single network. The base network is VGG16 which is used for high quality image classification. The convolutional feature layers are added to the end of the base network. The SSD model is trained on the COCO dataset [40] and is able to detect and recognize about 80 categories objects which includes the person, bicycle, car, motorcycle. These objects are frequently encountered in our daily lives, which is a great help to VIP for obstacles perceiving and avoiding. The SSD network could achieve the near real-time object detection without the GPU, for instance, about 100ms per frame running on a portable laptop (with I5-6300@2.4GHz, 8G RAM) for 640×480 input.

Compared with the SSD, instance segmentation based on the Mask R-CNN [20] has emerged as an extremely powerful approach to detect and identify multiple classes of scenes and objects simultaneously. Mask R-CNN [20] extends Faster R-CNN [41] by adding a branch for predicting an object mask in parallel with the existing branch for bounding box detection. The Mask R-CNN is also trained on the COCO dataset, and it has a better detection result compared with the SSD, for instance, in the Fig. 5(b2) and (c2), the Mask R-CNN could discover three people in the image, but the SSD could only find two. However, the Mask R-CNN takes more time than the SSD (about 2000ms) to process an image on the laptop. The research topic of designing pixel-wise instance segmentation to assist the visually impaired has not been widely investigated. In this paper, we achieve the object recognition using the SSD and Mask R-CNN respectively.

3.3. Calibration

In multiple sensor system, each sensor obtains data in its own coordinate system, which needs to be transformed into a unified coordinate system. The depth images are generated by the image processor through stereo matching between the right and left infrared cameras. Then the R200 takes the depth images corresponding to the left infrared camera. Therefore, we need to calibrate the left infrared camera and the radar coordinate, the left infrared camera and the color camera respectively.

In this paper, the standard pinhole model can be used for the infrared camera. As shown in Fig. 6, the (x_c, y_c, z_c) and (u, v) are the camera coordinate and the image plane coordinate respectively. The relationship between them is described as equation (1). The f_x, f_y, c_x, and c_y are the x, y direction focal lengths and principal point coordinates respectively, and K is the matrix of intrinsic parameters.

$$z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} \quad (1)$$

By contrast, a target is detected by the radar, its coordinate is (x_r, y_r, z_r), as illustrated in Fig. 6. Because the radar is unable to give the pitch angle in the 3D plane, the y_r coordinate has no meaningful value. The calibration matrix M^{RT}, which includes the rotation R and the

translation T , between the RGB-D coordinate and the

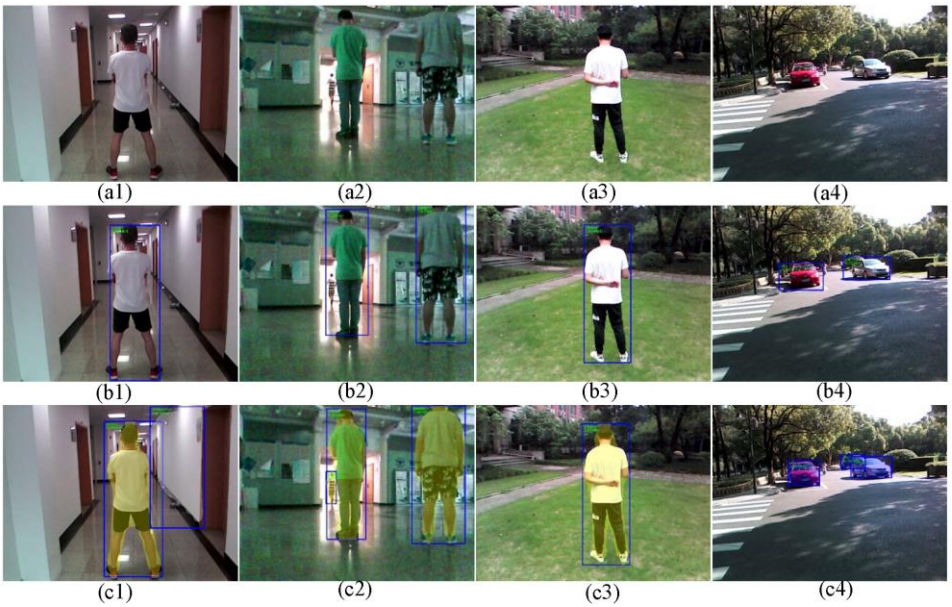


FIG. 5. The object recognition based on the color images in both indoor and outdoor environments. (a1-a4) The color images. (b1-b4) The object recognition using the SSD network. (c1-c4) The object recognition and instance segmentation using the Mask R-CNN network.

radar coordinate, is obtained through:

$$\begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = M^{RT} \begin{bmatrix} x_r \\ z_r \\ 1 \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{bmatrix} \begin{bmatrix} x_r \\ z_r \\ 1 \end{bmatrix} \quad (2)$$

where the calibration matrix M^{RT} is made up of 9 elements.

To estimate the matrix M^{RT} , we concurrently observed the 3D coordinates of the objects with both sensors and estimated the best-fit transformation between them using the linear least squares procedure.

Rotation and translation also exist between the left infrared camera and the color camera, as shown in Fig. 2(b). It is necessary to calibrate them. We calibrate them using the MATLAB stereo camera calibrator toolbox, as shown in Fig. 7. We acquire a number of checkerboard images using the left infrared camera and the color camera simultaneously. These two datasets are fed into the stereo camera calibrator toolbox, and the corner detection and the calibration are automatically completed. Then, the calibration results are obtained.

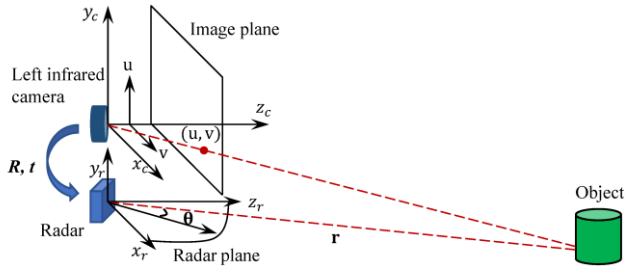


FIG. 6. The RGB-D sensor coordinate and MMW radar coordinate

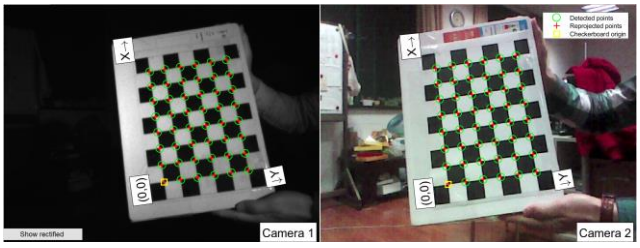


FIG. 7. The calibration between the left infrared camera and the color camera using the MATLAB stereo camera calibrator toolbox. Camera 1 is the left infrared camera; Camera 2 is the color camera. The checkerboard images are acquired concurrently, fed into the toolbox and the corner detection is successfully accomplished.

3.4. Data fusion

The target fusion task from different sensors is generally solved by the method of Kalman Filter [31]. However, the Kalman Filter is unimodal, while the Joint Integrated Probabilistic Data Association (JIPDA) algorithm [16,31] is needed to track and label this multiple target before performing the Kalman Filter, which makes the full process very complex and time-consuming. Following this rationale, the Particle Filter based on the principle of Monte Carlo sampling is used in our application to accomplish the multiple object data fusion. The Particle Filter is multimodal which is able to track and fuse more than one object simultaneously. It is also suitable when the multivariate, nonlinear behavior and non-Gaussian noise situation appear.

A particle filter [23,42,43] is a recursive Bayesian state estimator that uses discrete particles to approximate the posterior distribution of the estimated state. As presented in Fig. 8, the particle filter algorithm computes the state estimate recursively and involves two main iteration steps for continuous estimating state: prediction and correction.

The prediction uses the previous state to predict the current state based on a given system model. And the correction uses the current sensor measurement to correct the state estimate. The particle filter also periodically resamples which makes the particles in the state space match the posterior distribution of the estimated state. The estimated state consists of all the state variables. Each particle represents a discrete state hypothesis. The set of all particles is used to help determine the final state estimate.

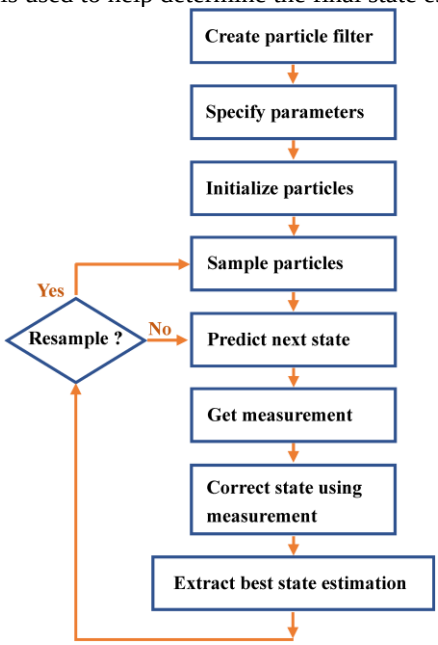


FIG. 8. The Particle Filter workflow

In our sensor fusion system, we use 1000 particles to estimate the state. The spatial position and velocity of the obstacles are regarded as state variables. Covariance matrix is set reasonably according to the radar and RGB-D sensor characteristics. We take the first frame radar, depth image and color image detection results as the initial particle location. Then the next state is evolved by the state transition function which is developed and implemented based on our system motion model. The weight for the state hypotheses based on a given measurement from the detection results is given through measurement likelihood function. Then the predicted particles are corrected and the best estimation is got according to the weight. The resampling of the particles is a vital step for continuous tracking objects, which selects particles based on the current state instead of particle distribution given at initialization. We trigger resampling based on the number of effective particles when a minimum effective particle ratio is reached. At last, the best estimation of the actual state is decided by the sum of all particle weights.

4. Experiments

In order to test and verify the performance of our objects detection, recognition and fusion system, the experiments are designed and performed. The sensors, which are introduced in the sub-section 2.1, are connected to a portable PC (with I5-6300@2.4GHz, 8G RAM) by the USB port. The portable PC is mainly responsible for the objects detection and recognition on the color images, the objects detection on the depth images, accepting the MMW radar detection results, the data fusion based on the Particle Filter and creating the non-semantic stereophonic sound and ordinary semantic speech. We achieve the running speed of around 8 FPS when the SSD network is taken. Nevertheless, the time consumed on per frame is taken more than 2000ms when the Mask R-CNN is used. The bone-conducting headphone that does not block VIP's ears from hearing environmental sounds is applied, which is connected to the PC by the Bluetooth. The non-semantic stereophonic interface [5] is utilized to achieve the multiple obstacle warning simultaneously, which makes VIP perceive the surrounding environments quickly. In the actual assistance process, the RGB-D and MMW radar sensors are hung from the user's neck, the Bluetooth bone conduction headphones are worn by the users and the portable PC is put in the backpack, as the Fig. 9 shown.

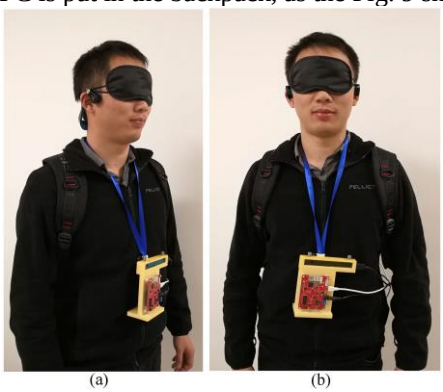


FIG. 9. The sensors are hung from the user's neck. (a) The side view. (b) The front view. The Bluetooth bone conduction headphones are worn on the head, and the portable PC is put in the backpack. The device is light and easy to wear by the user.

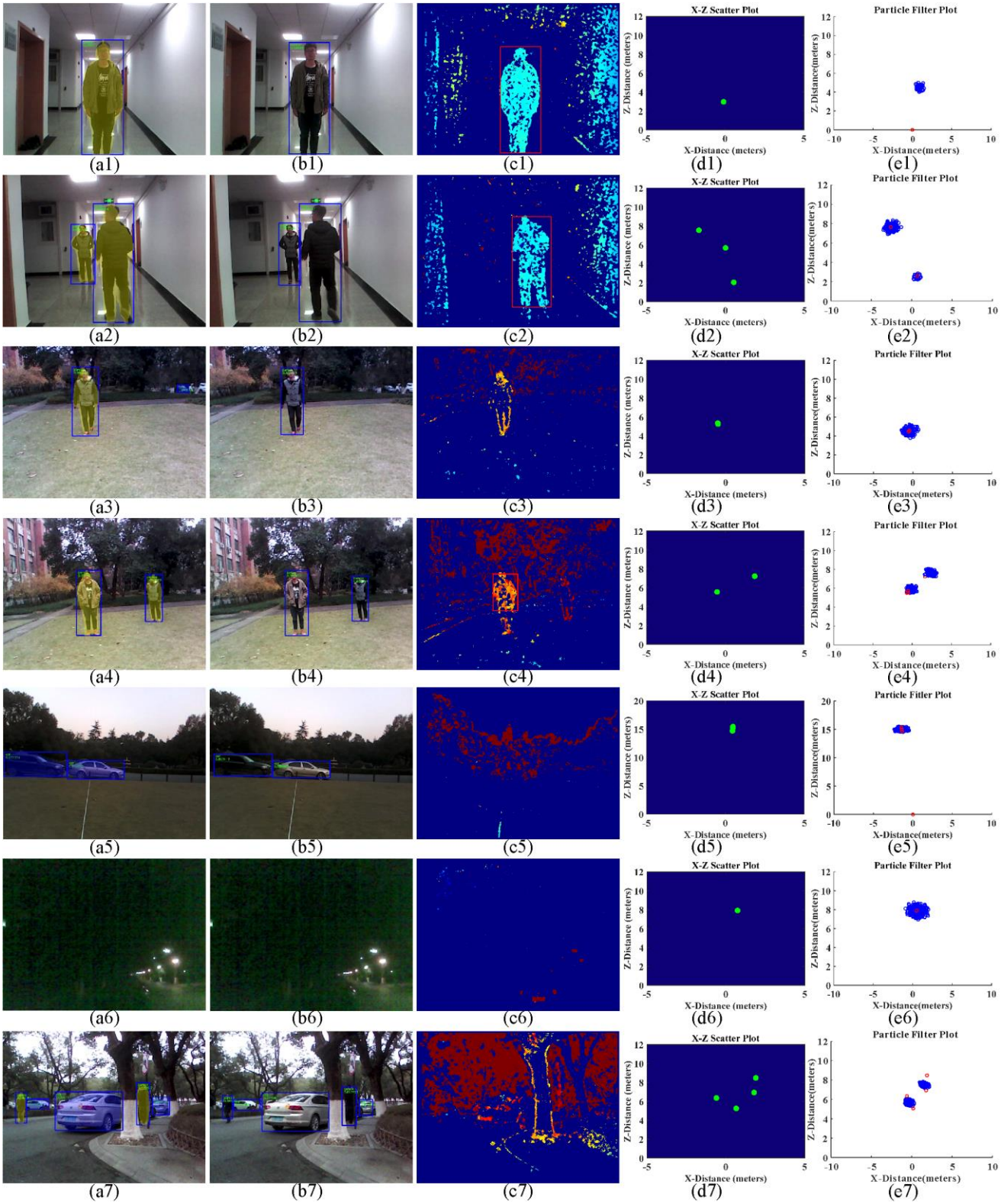
4.1. Field tests

The field tests are designed and performed with different surroundings, as shown in Fig. 10. The color images and the recognition results using the Mask R-CNN are presented in a1-a7, the targets are detected with different surroundings. The detection and recognition results by the SSD network are shown in b1-b7 simultaneously. The depth images and the detection results based on the MeanShift algorithm are described in c1-c7. The detection results are represented by the red bounding boxes. The MMW radar detection results are shown in d1-d7. At last, the data fusion results, which is based on the method of the Particle Filter, are described in e1-e7. The position information of the detected objects is represented by the mean value of these particles.

In the scenario 1, a person stands in the corridor, which could be detected and recognized in the color images using the Mask R-CNN and SSD, as the Fig. 10 (a1) and (b1) illustrated. The obstacles at most cases could be detected in the depth images when the illumination and distance are suitable, as shown in (c1). Meanwhile, the

person is detected by the MMW radar, as (d1) said. At last, these detected results, which come from the color images, the depth images and the radar, are put into the Particle Filter. After several iterations, the position of the observed person is confirmed, as shown in (e1). Then we could know the position, velocity and category of the detected object. Compared with the single RGB-D sensor or the single MMW radar, the data fusion enriches the detection results.

FIG. 10. The field tests are performed in indoor and outdoor. (a1-a7) The objects are detected and recognized in the color images using the Mask R-CNN. (b1-b7) The objects are detected and recognized in the color images using the SSD network. (c1-c7) The depth images and the detection results, the detection results are represented by the red bounding box. (d1-d7) The radar detection results. (e1-e7) The data fusion using the Particle Filter, the position information is represented by the mean value of these particles.



Similarly, two people appear in the corridor, one near and the other far. They are both detected and recognized in color image, as shown in (a2) and (b2). Since the effective detection range of the RealSense R200 is only about 0.6m to several meters, the near one is detected in depth image, but the other one is not, as presented in (c2). By contrast, the detection range of this low-power MMW radar is longer, about 15m for a person and 80m for a car, and the results are stable. Then they are both detected by the radar, as shown in (d2). At last, the data fusion results are presented in (e2). In this regard, the effective detection range of the fusion system has been significantly expanded.

Meanwhile, similar experiments are also conducted in outdoor environments. In the scenario 3 and 4, one or multiple person stands in a courtyard, as Fig. 10 (a3), (b3), (a4) and (b4) show. They are all detected and recognized successfully on the color images based on the method of Mask R-CNN or SSD network. In the scenario 3, the person appears at about 5.5 meters away. He is not detected in the depth image because of the limited perception range of the RGB-D sensor, as described in Fig. 10 (c3). In contrast, the nearer person in the scenario 4 is found in the depth image, as said in (c4). But the farther one disappears in the depth image. However, all these targets are correctly detected by the radar, as shown in (d3) and (d4). The data fusion results based on the Particle Filter are presented in the (e3) and (e4) respectively. Compared with the single RGB-D sensor, the data fusion improves the robustness of the prototype.

In the scenario 5, two cars are parked on the roadside at afternoon, and they are successfully detected and recognized on the color images, as shown in (a5) and (b5). They are not found in the depth image because of the limited perception range of the RGB-D sensor, as described in (c5). Nevertheless, the MMW radar detection result is stable, as shown in (d5). However, the angle resolution of the MMW radar is limited which makes these two cars not distinguished. With the help of high-resolution color images, we could know the targets' number and their categories. Compared with the single RGB-D sensor or the single MMW radar, the data fusion enriches the detection results, expand the effective detection range and improves the robustness of the prototype.

In some special extreme environments, the MMW radar provides the last security guarantees. For instance, a car is parked on the roadside at nightfall, as shown in Fig. 10 (a6) and (b6). The car is not found in the color images because of the illumination. It is also not found in the depth image because the quality of the depth image declines when the lights dim or the perception distance exceeds the effective detection range of the RealSense R200. The MMW radar detection results, by contrast, are stable. While the last data fusion results are shown in (e6), which reveals the effectiveness of our approach even with low illumination. In this sense, our fusion system enhances the robustness of obstacle detection across different illumination conditions.

In some complex environments, three or more objects are detected. For example, in the scenario 7, three cars and two people appear near the roadside, as Fig. 10 (a7) and (b7) depicted. All of the cars and people are completely detected on the color images using the Mask R-CNN or SSD network. Because all the targets are out of the effective perception range of the RGB-D sensor, they are not detected on the depth image, as shown in Fig. 10 (c7). The MMW radar only detects four objects because of the low directional resolution. And the detection results are presented in (d7). In contrast, the targets' number and their categories could be confirmed on the high-resolution color images. At last, the data fusion results based on the Particle Filter are described in (e7). There are two pairs of objects, and the two objects of each pair are close to each other, which makes only two cluster particles generated in the data fusion results. Although the performance of the sensor fusion system has a certain degradation in some complex environments, it is still able to provide some assistance and help the VIP perceive and avoid obstacles at a distance.

The field tests show the effective detection range is significantly expanded with the help of the sensor fusion. Meanwhile, the more accurate state estimation, the richer information and more robust perform under diverse illumination conditions are obtained compared with the single sensor.

4.2. The performance evaluation at different ranges

In order to verify the performance of our sensor fusion system at different ranges, the experiments are designed and performed. As shown in Fig. 11, the car is placed at roadside. We measure the different distances between the car and our system, and the distance of 2m, 4m, 8m, 15m, 20m, 30m, 40m and 50m are selected. The 50-meter leather tape box ruler is placed on the ground to get the accurate ground truth range information.

The car is successfully detected and recognized on the color images when the distance is 2m, 4m, 8m, 15m and 20m, as shown in Fig. 11 (a1-a5) and (b1-b5). However, the car is detected on the depth images only the distance is about 2m and 4m. In spite of being detectable, only a small part of this car is correctly detected on the depth images. The depth images lack effective detection information when the distance exceeds 4m. By contrast, the MMW radar still keeps accurate range perception ability, as shown in c1-c5. Because of the limited angle resolution, the MMW radar has no ability to distinguish the two close cars. The color images detected results exactly provide this information, i.e., the number and the category. Because the two cars are too close, they are not separated in the last data fusion results, which are described in e1-e5.

As the distance increases, the car imaging area becomes small. The error of object detection and recognition in the color images starts to appear. As shown in b6 and b7, the car in the middle position is lost because of the SSD network lacking robustness. In contrast, the

object detection and recognition based on the Mask R-CNN solves this problem very well in spite of taking more time, as described in a6 and a7. Compared to the object detection on color images, the radar detection results keep still stable when the distance varies. In summary, we obtain the objects category, position and velocity information through the data fusion based on the RGB-D sensor and the MMW radar.

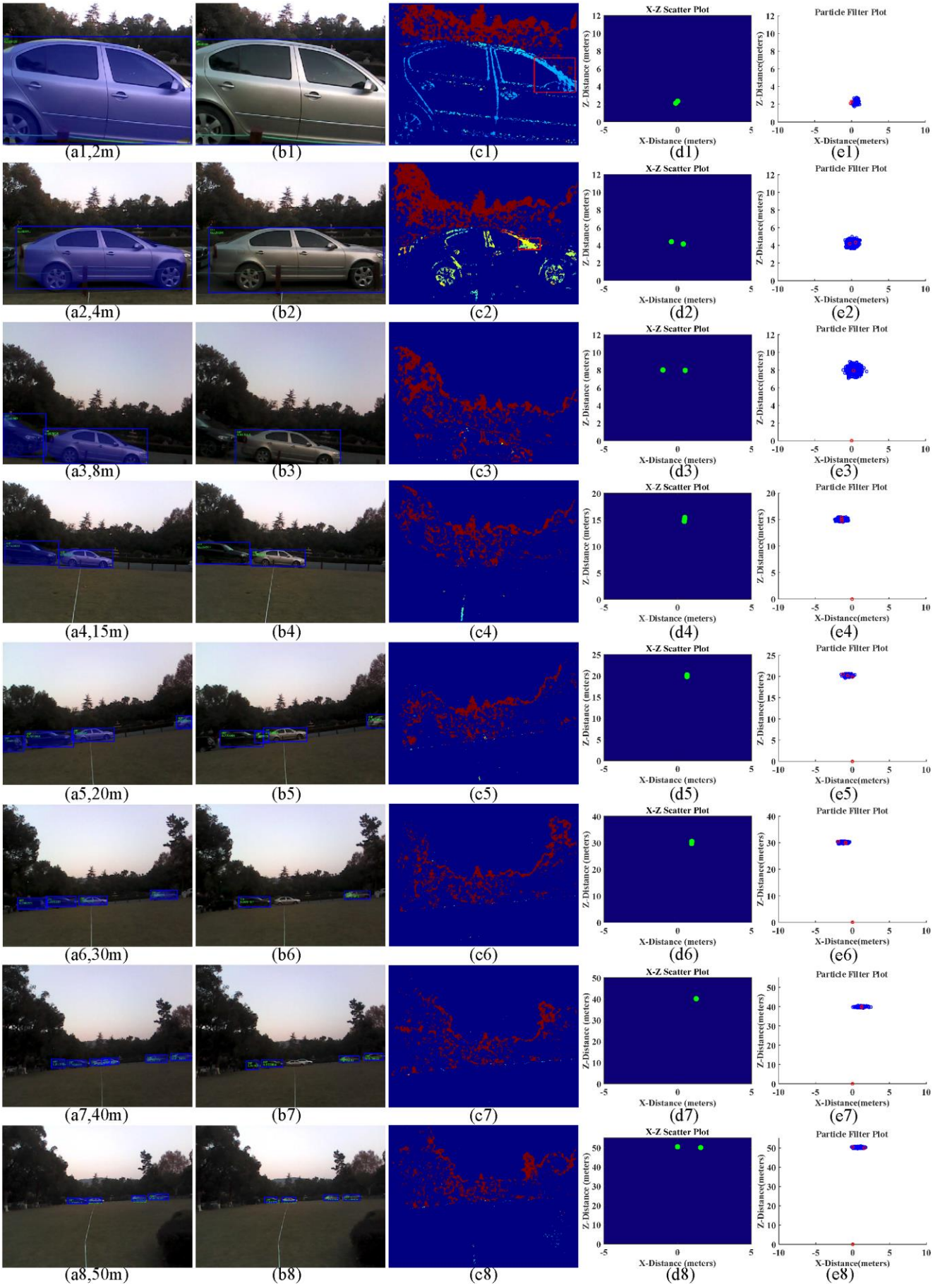

(a1,2m)
(b1)
(c1)
X-Z Scatter Plot
Z-Distance (meters)
X-Distance (meters)
(d1)
Particle Filter Plot
Z-Distance(meters)
X-Distance(meters)
(e1)
(a2,4m)
(b2)
(c2)
(d2)
(e2)
(a3,8m)
(b3)
(c3)
(d3)
(e3)
(a4,15m)
(b4)
(c4)
(d4)
(e4)
(a5,20m)
(b5)
(c5)
(d5)
(e5)
(a6,30m)
(b6)
(c6)
(d6)
(e6)
(a7,40m)
(b7)
(c7)
(d7)
(e7)
(a8,50m)
(b8)
(c8)
(d8)
(e8)

FIG. 11. The data fusion performance evaluation at different ranges using the Particle Filter. (a1-a8) The color images, the objects are detected and recognized using the Mask R-CNN. (b1-b8) The color images, the objects are detected and recognized using the SSD network. (c1-c8) The depth images and the detection results, the detection results are represented by the red bounding box. (d1-d8) The radar detection results. (e1-e8) The data fusion using the Particle Filter, the position information is represented by the mean value of these particles.

At last, the RGB-D sensor detection results, the MMW radar detection results, the data fusion results and the corresponding ground truth at different ranges are listed in the Table 1. It can be seen that the distance measurement accuracy is high and keeps stable at different ranges with the help of RGB-D sensor and MMW radar. Moreover, the objects at a long distance could be detected accurately with the help of this sensor fusion system.

TABLE I. The detection and fusion results at different ranges

Scene	a1	a2	a3	a4	a5	a6	a7	a8
RGB-D(m)	1.978	3.859	-	-	-	-	-	-
Radar (m)	2.093	4.042	7.935	15.015	20.142	30.029	39.917	49.797
Fusion Data (m)	2.018	4.023	7.958	14.987	20.208	29.832	40.124	49.985
Ground Truth (m)	2	4	8	15	20	30	40	50
Fusion Deviation (m)	0.018	0.023	-0.042	-0.013	0.208	-0.168	0.124	-0.015
Relative Error (%)	0.9	0.575	0.525	0.087	1.04	0.56	0.31	0.03

5. Conclusion

In this paper, we present a unified target detection, recognition and fusion framework based on the sensor fusion system which is comprised of a low-power MMW radar and an RGB-D sensor. The experiment results show the object detection and recognition on the color images is achieved based on the Mask R-CNN or the SSD network. The feature extraction on the depth image is achieved using the MeanShift algorithm, and the depth and position information of the obstacles is obtained. The field tests show the different ranges and angles of the objects are calculated by the MMW radar based on the principle of FMCW. With the help of the data fusion, we have achieved more accurate state estimation and obtained richer information of the detected targets. Moreover, the measurement results are stable under diverse illumination conditions. As a wearable system, the sensor fusion system has the characteristics of versatility, portability and cost-effectiveness, which is very suitable for blind navigation application. Simultaneously, this system could be flexibly applied in the field of self-driving, unmanned aerial vehicle (UAV), robotics, surveillance and defence.

For future work, we plan to achieve the data fusion algorithm on the Field Programmable Gate Array (FPGA) chip. This greatly reduces the size and weight of the system, which is more portable during navigation.

REFERENCES

[1]Bourne, R.R., Flaxman, S.R., Braithwaite, T., Cicinelli, M.V., Das, A., Jonas, J.B., Keeffe, J., Kempen, J.H., Leasher, J., Limburg, H. and Naidoo, K., The Lancet Global Health, 5, e888 (2017).

[2]Yang, K., Cheng, R., Bergasa, L. M., Romera, E., Wang, K., and Long, N., In 2018 IEEE International Conference on Robotics and Biomimetics (ROBIO), 1034 (2018).

[3]A. Bhowmick and S. M. Hazarika, Journal on Multimodal User Interfaces, 11, 149 (2017).

[4]L. Shao, J. Han, P. Kohli and Z. Zhang, Computer Vision and Machine Learning with RGB-D Sensors, Springer International Publishing, (2014).

[5]K. Yang, K. Wang, W. Hu and J. Bai, Sensors 16, 1954 (2016).

[6]K. Yang, K. Wang, R. Cheng, W. Hu, X. Huang and J. Bai, Sensors 17, 1890 (2017).

[7]K. Yang, K. Wang, X. Zhao, R. Cheng, J. Bai, Y. Yang and D. Liu, Journal of Ambient Intelligence and Smart Environments 9, 743 (2017).

[8]R. Cheng, K. Wang, K. Yang, N. Long, J. Bai and D. Liu, Multimedia Tools and Applications, 1 (2017).

[9]R. Cheng, K. Wang, K. Yang, N. Long, W. Hu, H. Chen, J. Bai and D. Liu, Journal of Electronic Imaging 26, 053025 (2017).

[10]K. W. Lin, T. K. Lau, C. M. Cheuk and Y. Liu, IEEE International Conference on Mechatronics and Automation, 1423 (2012).

[11]X. Zhao, K. Wang, K. Yang and W. Hu, International Conference on Graphics and Image Processing (ICGIP2016), 1022509 (2017).

[12]A. Tamjidi, C. Ye and S. Hong, 2013 IEEE International Symposium on Robotic and Sensors Environments (ROSE), 178 (2013).

[13]G. L. Charvat, Small and Short-Range Radar Systems, CRC Press, (2014).

[14]N. Long, K. Wang, R. Cheng, K. Yang and J. Bai, Proceedings of SPIE Security + Defence, 1080006 (2018).

[15]M. Bertozzi, L. Bombini, P. Cerri, P. Medici, P. C. Antonello and M. Miglietta, Proceedings of 2008 IEEE Intelligent Vehicles Symposium, 608 (2008).

[16]Ćesić, J., Marković, I., Cvišić, I., and Petrović, I., Robotics and Autonomous Systems 83, 338 (2016).

[17]Cippitelli, E., Fioranelli, F., Gambi, E., and Spinsante, S., IEEE Sensors Journal, 17, 3585 (2017).

[18]Kim, J. H., Starr, J. W., and Lattimer, B. Y., Fire Technology, 51, 82 (2015).

[19]Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., and Berg, A. C., In European conference on computer vision, 21 (2016).

[20]He, K., Gkioxari, G., Dollár, P., and Girshick, R., In Proceedings of the IEEE international conference on computer vision, 2961 (2017).

[21]Stove, A. G., In IEE Proceedings F (Radar and Signal Processing),139, 343 (1992).

[22]D. Comaniciu and P. Meer., IEEE Transactions on Pattern Analysis & Machine Intelligence, 24, 603 (2002).

[23]F. Gustafsson, IEEE Aerospace and Electronic Systems Magazine 25, 53 (2010).

[24]Jindal, A., N. Aggarwal, and S. Gupta. Pattern Recognition and Image Analysis 28, 288 (2018).

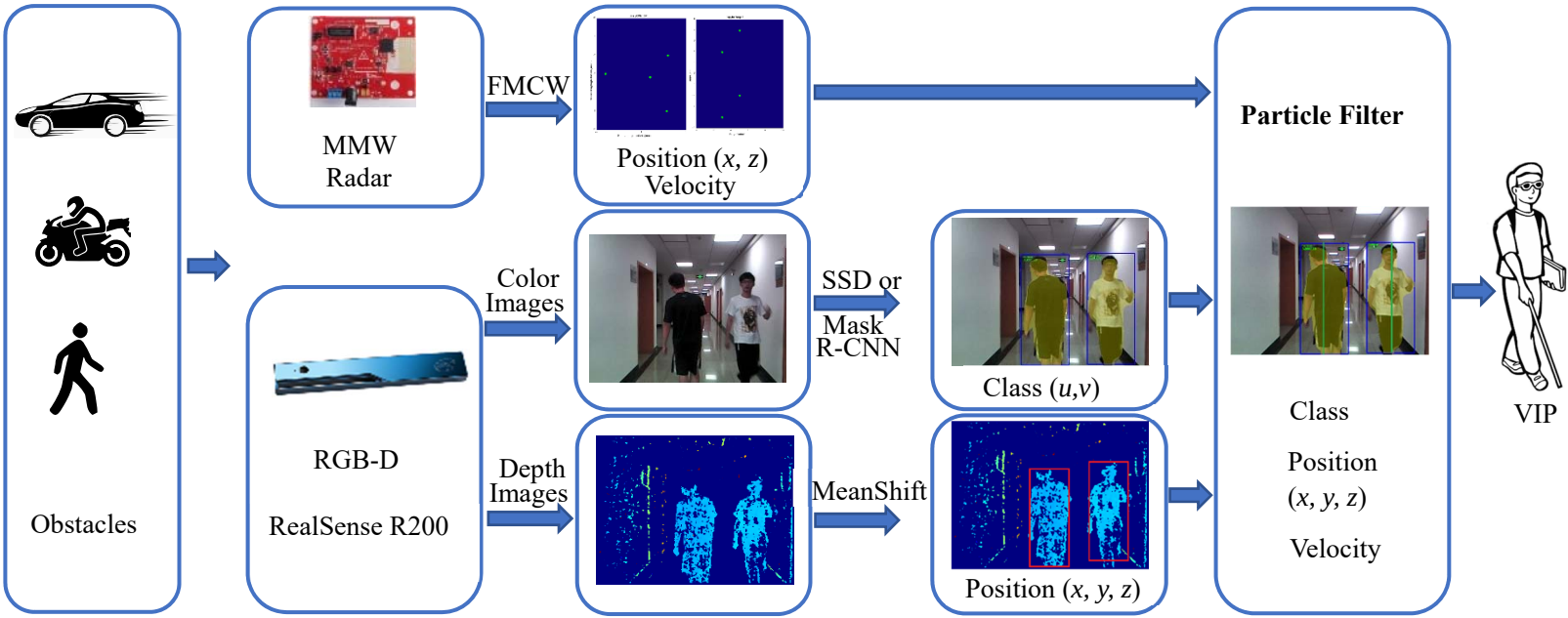
[25]Kaur, B., Bhattacharya, J., Journal of Electronic Imaging, 28, 013031 (2019).

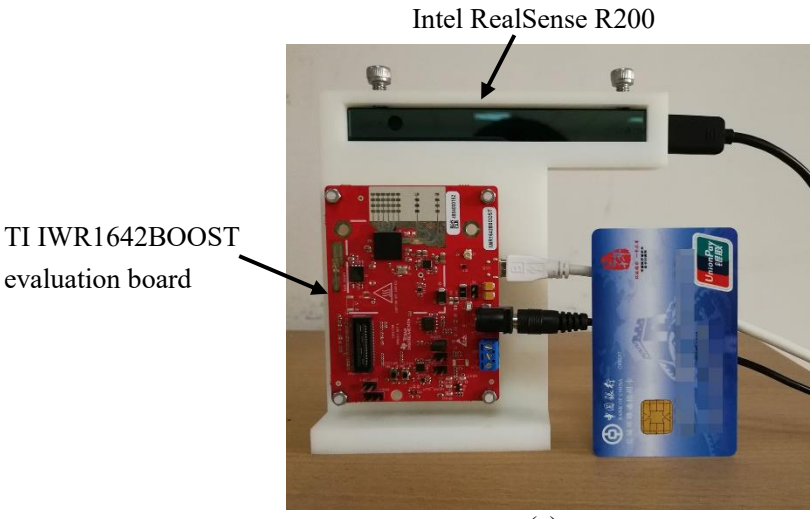
[26]Kiuru T, Metso M, Utriainen M, Metsävainio K, Jauhonen H. H., Rajala R, Savenius R, Ström M, Jylhä T.N., Juntunen R. and Sylberg J., Cogent Engineering, 5, 1450322 (2018).

[27]K. Yang, K. Wang, L.M. Bergasa, Romera E., W. Hu, D. Sun, J. Sun, R. Cheng, T. Chen and E. Lopez, Sensors, 18, 1506 (2018).

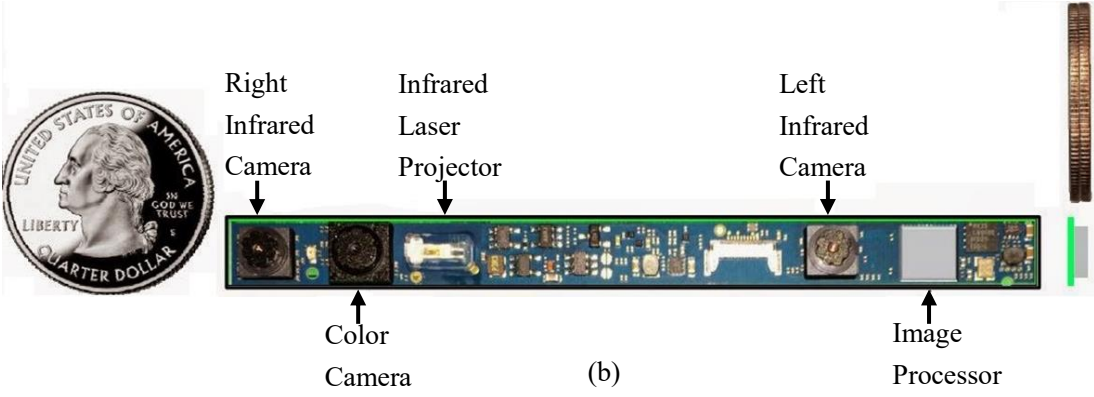
[28]Aladren A, López-Nicolás G, Puig L, Puig L and Josechu J. Guerrero, IEEE Systems Journal, 10, 922 (2016).

[29]S. Sugimoto, H. Tateda, H. Takahashi and M. Okutomi, Proceedings of Proceedings of the 17th International Conference on Pattern Recognition, 3, 342 (2004).
[30]T. Wang, N. Zheng, J. Xin and Z. Ma, Sensors 11, 8992 (2011).
[31]D. Y. Kim and M. Jeon., Information Sciences 278, 641 (2014).
[32]Y. Fang, I. Masaki and B. Horn IEEE Transactions on Intelligent Transportation Systems 3, 196 (2002).
[33]S. Wu, S. Decker, P. Chang, T. Camus and J. Eledath IEEE Transactions on Intelligent Transportation Systems 10, 606 (2009).
[34]P. Molchanov, S. Gupta, K. Kim and K. Pulli, Proceedings of 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 1 1(2015).
[35]M. Obrvan, J. Ćesić and I. Petrović, In Robot 2015: Second Iberian Robotics Conference, 437 (2015).
[36]R. O. Chavez-Garcia, J. Burlet, T. D. Vu and O. Aycard, In 2012 IEEE Intelligent Vehicles Symposium, 159 (2012).
[37]P. Molchanov, S. Gupta, K. Kim and K. Pulli, In 2015 IEEE Radar Conference, 1491 (2015).
[38]R. Rouveure, P. Faure and M.-O. Monod, Journal of Field Robotics, 35, 678 (2018).
[39]H. Rohling, In 2011 12th International Radar Symposium (IRS), 631 (2011).
[40]Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C.L., In European conference on computer vision, 740 (2014).
[41]Ren, S., He, K., Girshick, R., and Sun, J. In Advances in neural information processing systems, 91 (2015).
[42]Jaward, M., Mihaylova, L., Canagarajah, N., and Bull, D., In 2006 IEEE Aerospace Conference, 8 (2006).
[43]Jinan, R., and Raveendran, T., Procedia Technology, 24, 980 (2016).

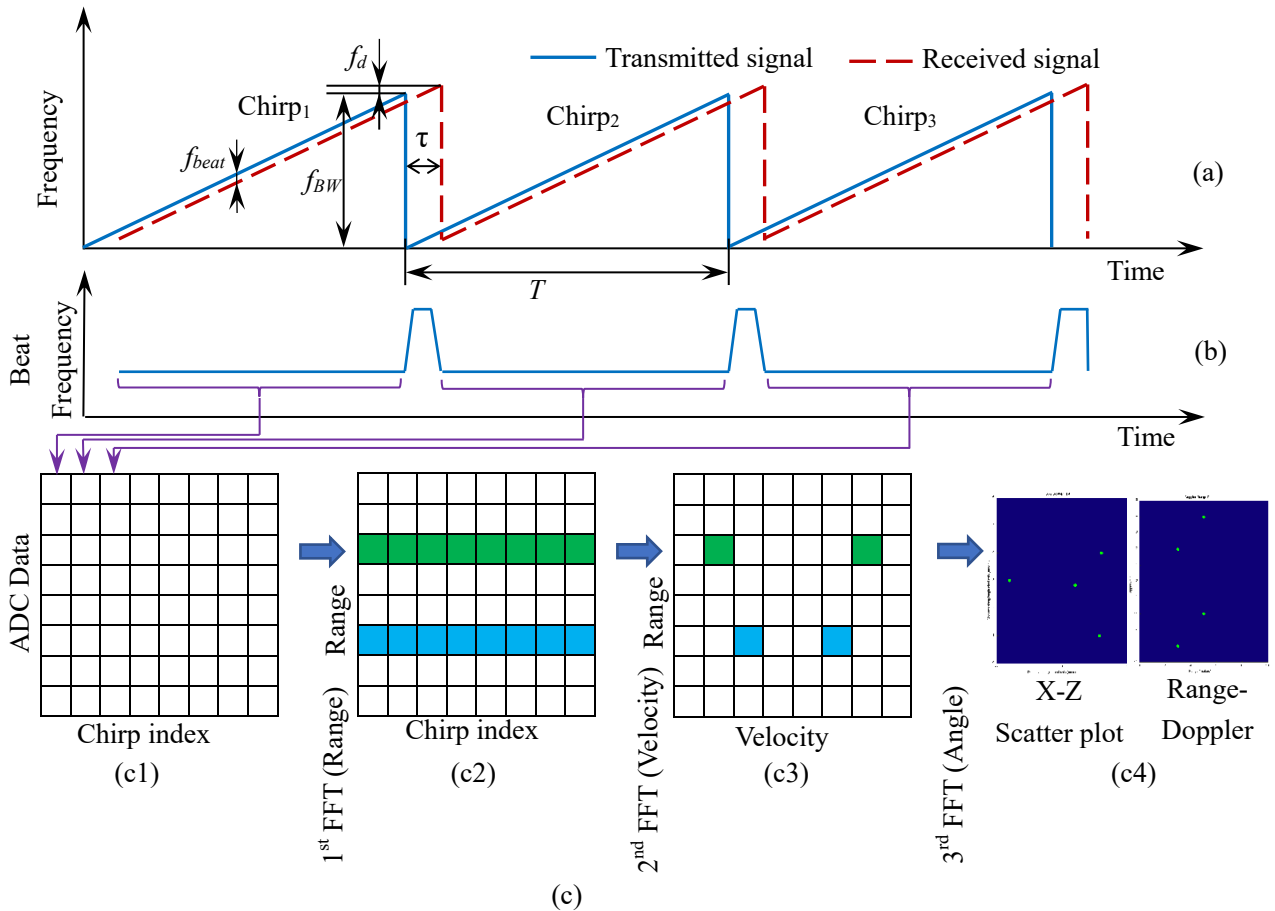

Obstacles
MMW
Radar
FMCW
Position (x, z)
Velocity
RGB-D
RealSense R200
Color
Images
Depth
Images
SSD or
Mask
R-CNN
MeanShift
Class (u,v)
Position (x, y, z)
Particle Filter
Class
Position
(x, y, z)
Velocity
VIP

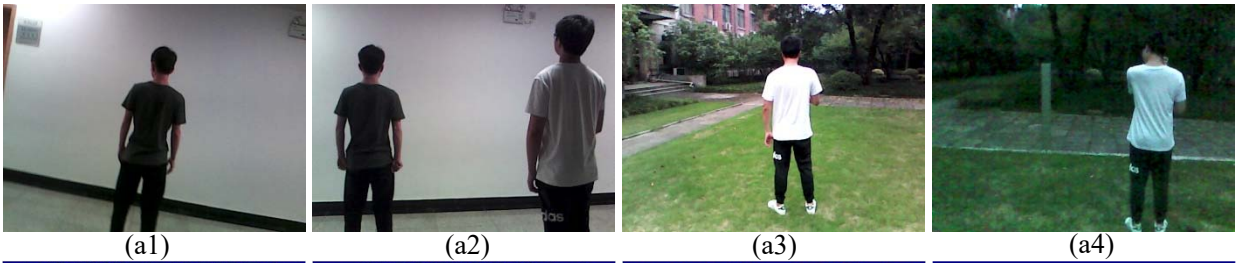

Intel RealSense R200
TI IWR1642BOOST
evaluation board

(a)


Right
Infrared
Camera
Infrared
Laser
Projector
Left
Infrared
Camera
Color
Camera
Image
Processor

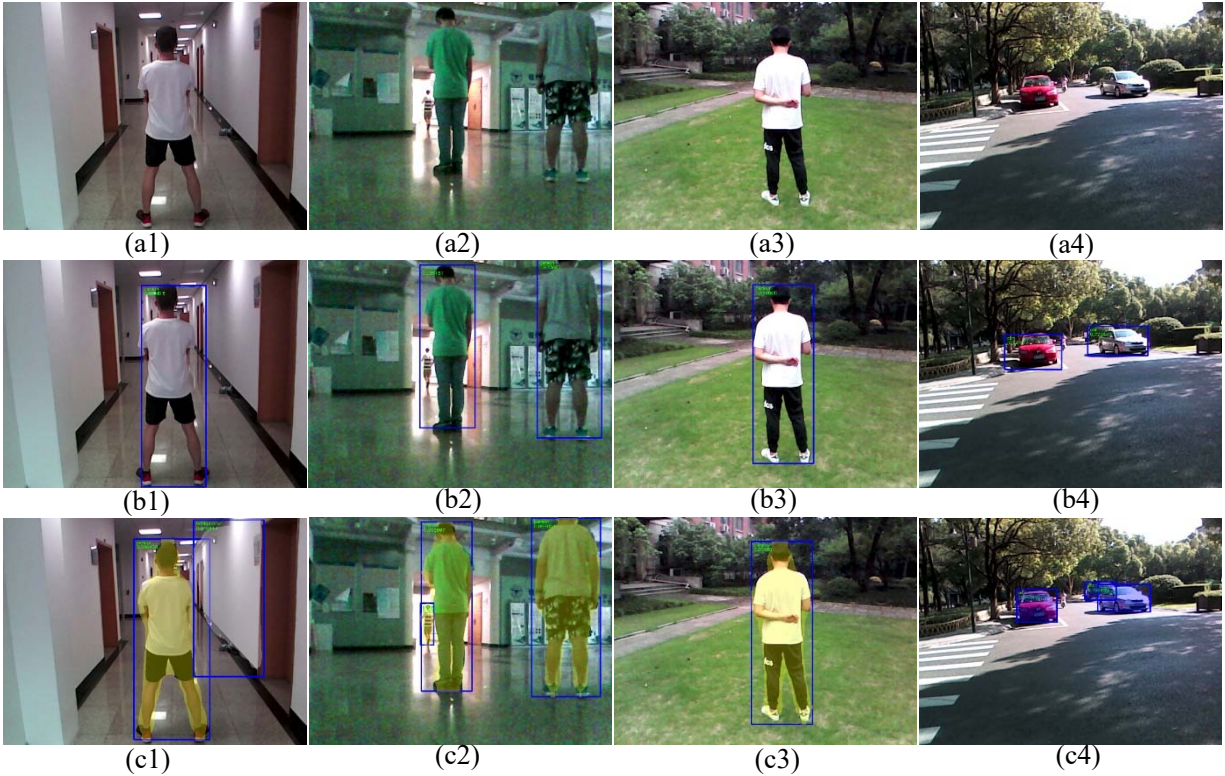
(b)

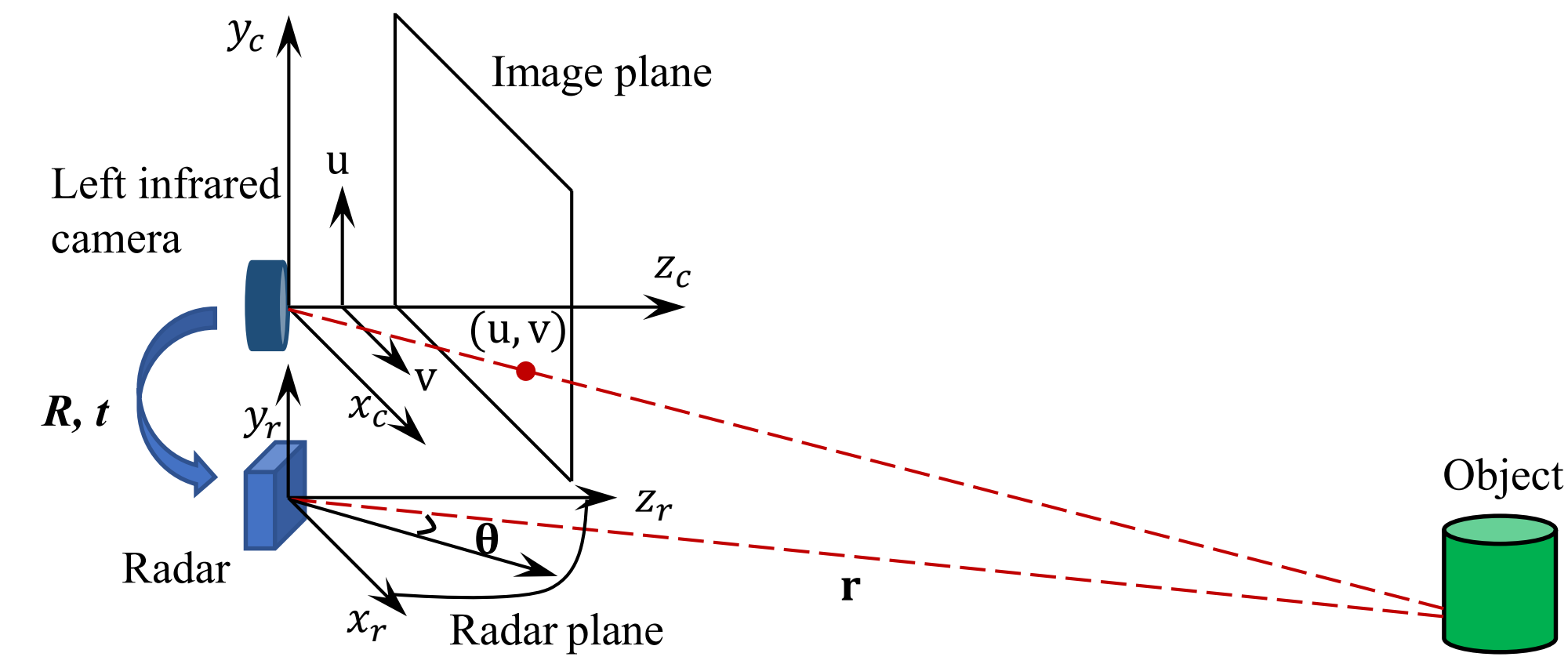

Frequency
Transmitted signal
Received signal
f_d
Chirp1
Chirp2
Chirp3
τ
f_{beat}
f_{BW}
(a)
Time
T
Beat
Frequency
(b)
Time
ADC Data
Chirp index
(c1)
1st FFT (Range)
Range
Chirp index
(c2)
2nd FFT (Velocity)
Range
Velocity
(c3)
3rd FFT (Angle)
X-Z
Scatter plot
Range-
Doppler
(c4)
(c)

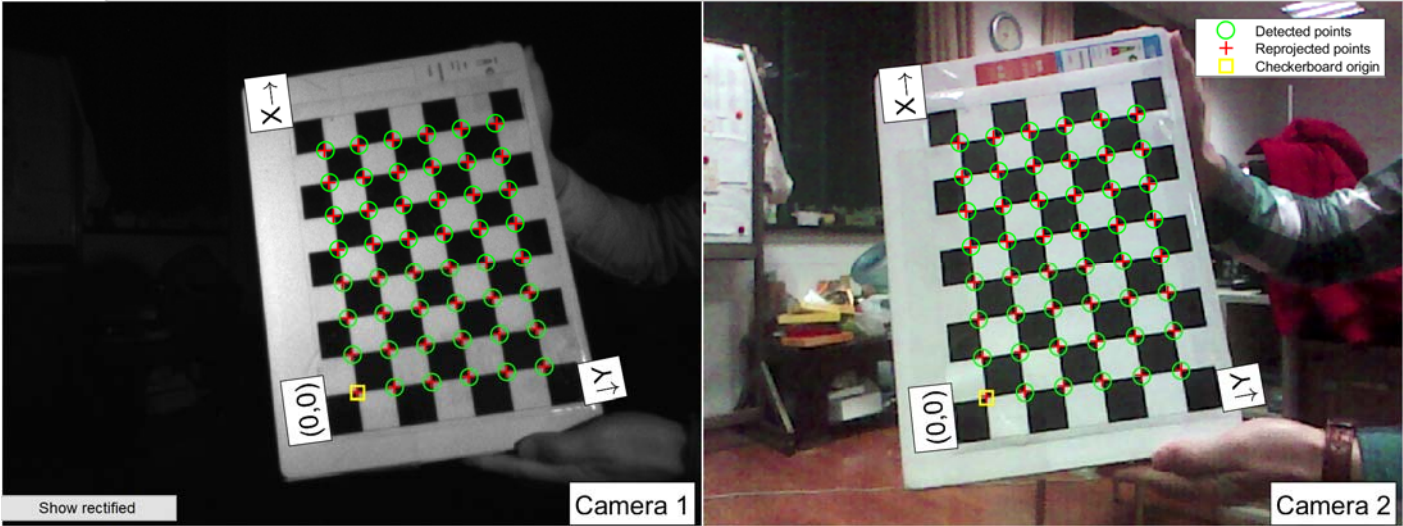


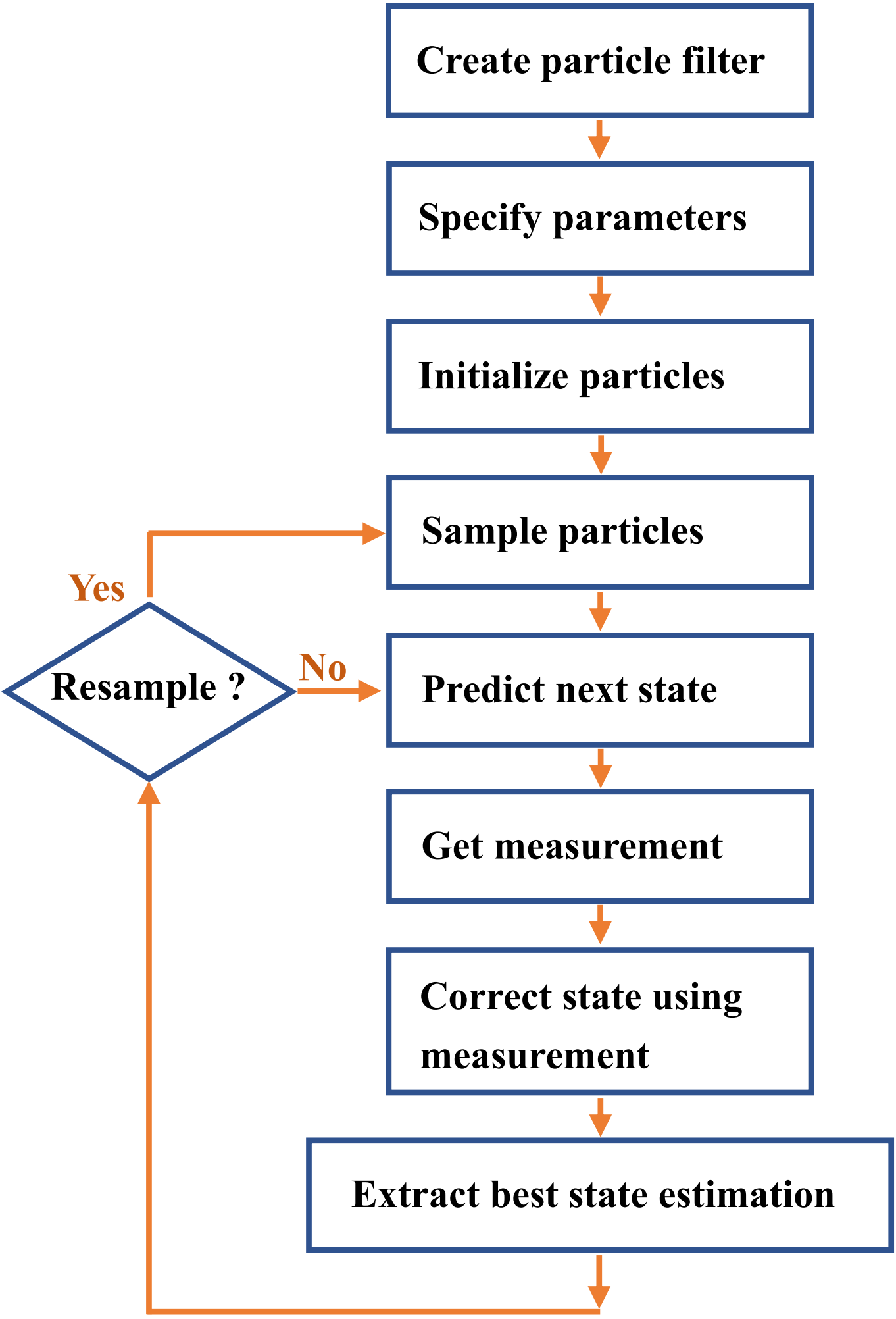
(a1) (a2) (a3) (a4)

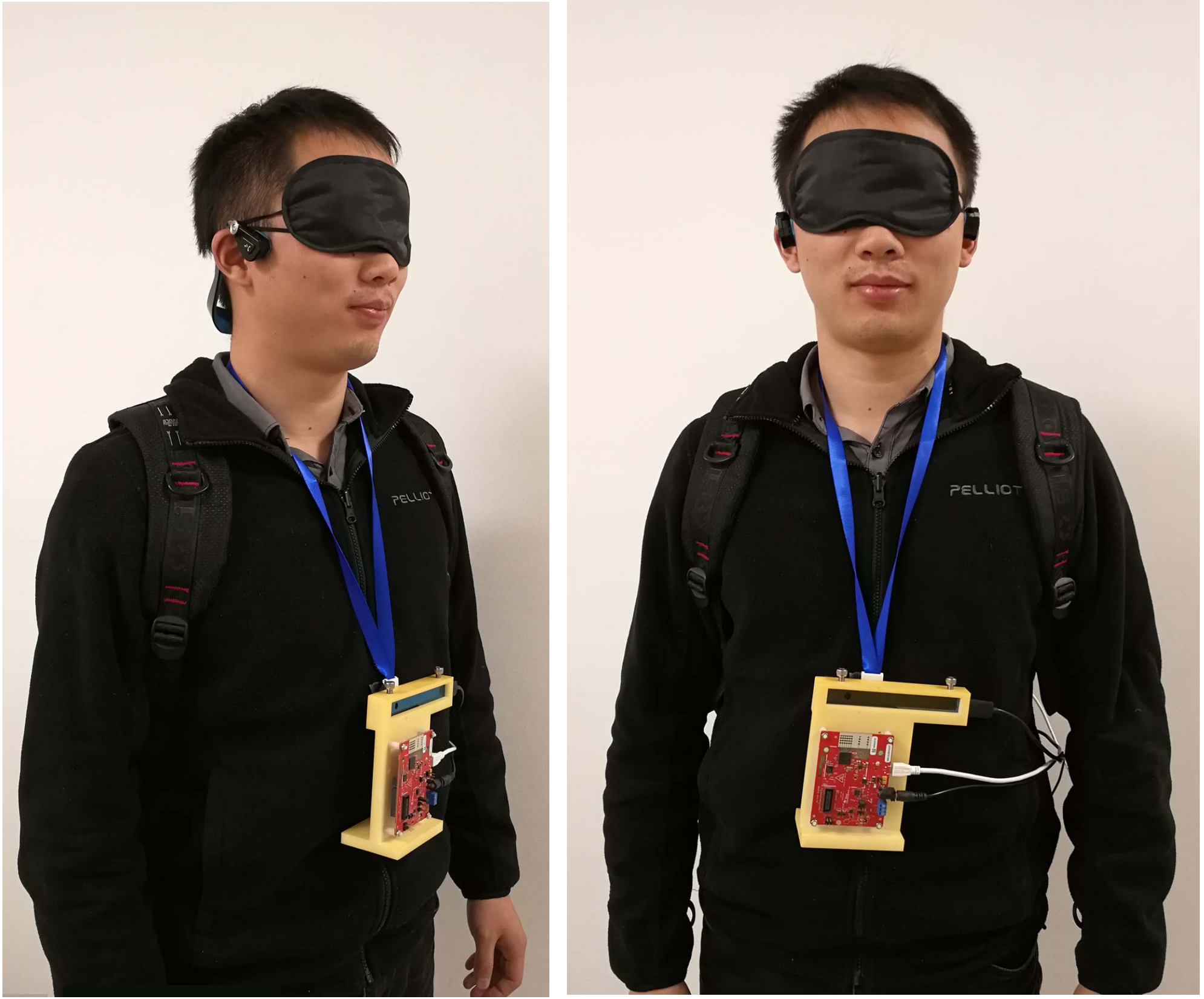
(b1) (b2) (b3) (b4)


(a1)
(a2)
(a3)
(a4)
(b1)
(b2)
(b3)
(b4)
(c1)
(c2)
(c3)
(c4)

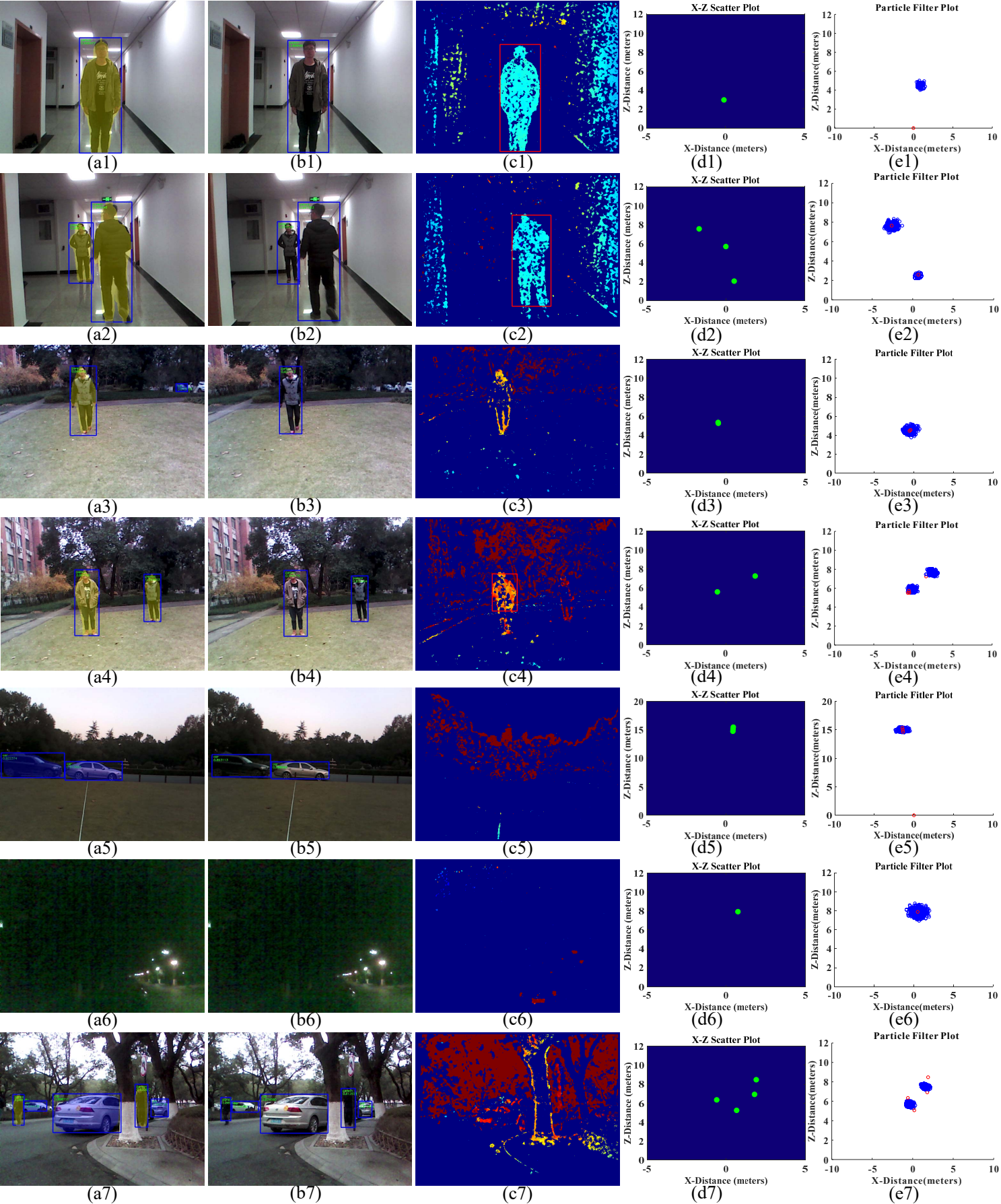

y_c
Image plane
u
Left infrared
camera
z_c
(u, v)
v
x_c
R, t
y_r
z_r
θ
Radar
r
Object
x_r
Radar plane

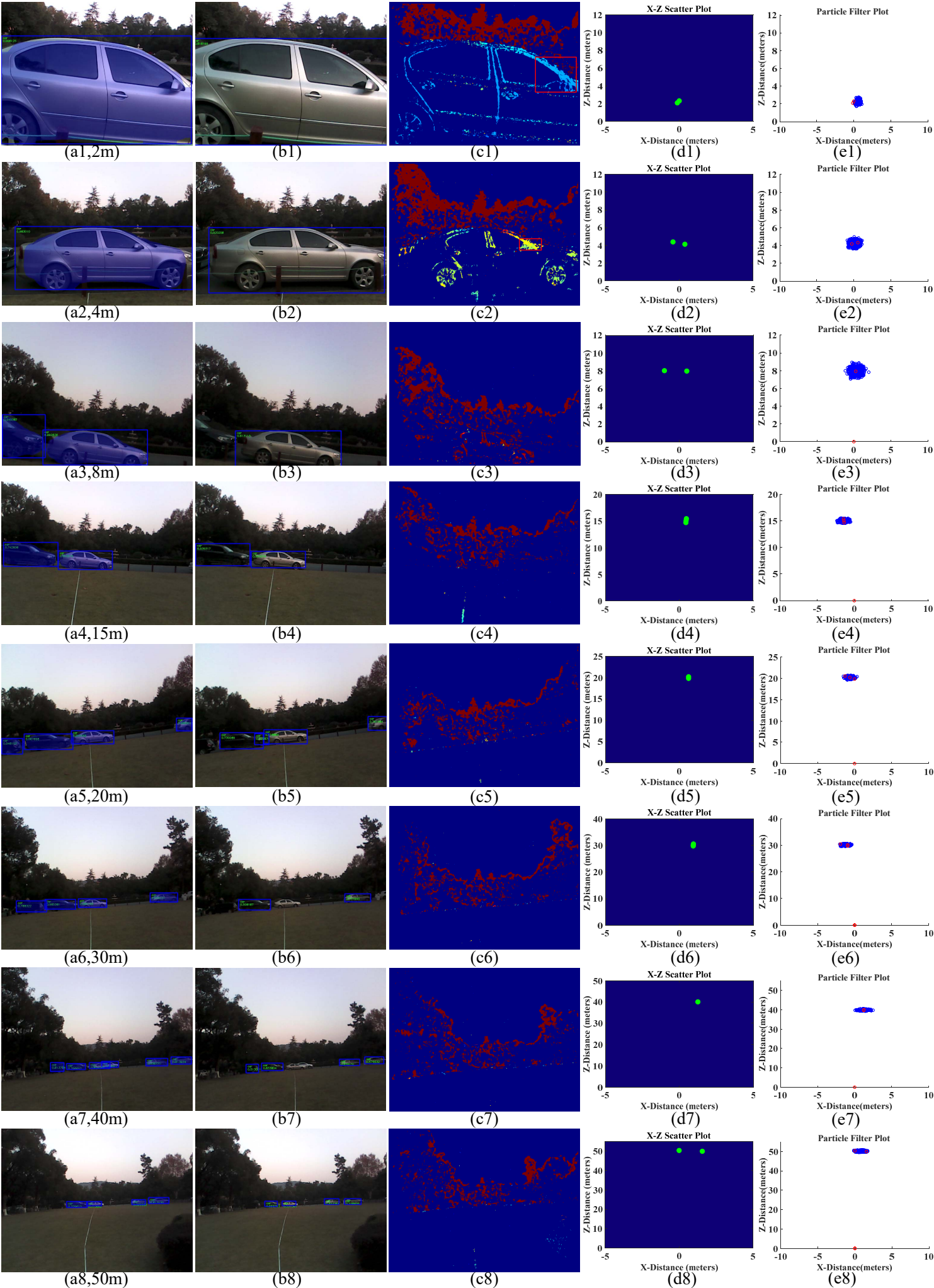

X
Y
(0,0)
Show rectified
Camera 1
Detected points
Reprojected points
Checkerboard origin
X
Y
(0,0)
Camera 2


Create particle filter
Specify parameters
Initialize particles
Sample particles
Yes
Resample ?
No
Predict next state
Get measurement
Correct state using measurement
Extract best state estimation



(a) (b)


(a1)
(b1)
(c1)
X-Z Scatter Plot
Z-Distance (meters)
X-Distance (meters)
(d1)
Particle Filter Plot
Z-Distance(meters)
X-Distance(meters)
(e1)
(a2)
(b2)
(c2)
(d2)
(e2)
(a3)
(b3)
(c3)
(d3)
(e3)
(a4)
(b4)
(c4)
(d4)
(e4)
(a5)
(b5)
(c5)
Particle Fitler Plot
(d5)
(e5)
(a6)
(b6)
(c6)
(d6)
(e6)
(a7)
(b7)
(c7)
(d7)
(e7)


X-Z Scatter Plot
Particle Filter Plot
Z-Distance (meters)
X-Distance (meters)
(a1,2m)
(b1)
(c1)
(d1)
(e1)
(a2,4m)
(b2)
(c2)
(d2)
(e2)
(a3,8m)
(b3)
(c3)
(d3)
(e3)
(a4,15m)
(b4)
(c4)
(d4)
(e4)
(a5,20m)
(b5)
(c5)
(d5)
(e5)
(a6,30m)
(b6)
(c6)
(d6)
(e6)
(a7,40m)
(b7)
(c7)
(d7)
(e7)
(a8,50m)
(b8)
(c8)
(d8)
(e8)