

## 2D Laser and 3D Camera Data Integration and Filtering for Human Trajectory Tracking

Hamed Bozorgi<sup>1</sup>, Xuan Tung Truong<sup>1</sup>, Hung Manh La<sup>2</sup> and Trung Dung Ngo<sup>1</sup>

**Abstract**—This paper addresses a robust human trajectory tracking method through the data integration of 2D laser scanner and 3D camera. Mapping the deep learning-based 3D camera human detection onto the point cloud of the depth information to build up the 3D bounding box-represented human and using the state-of-the-art 2D laser-based leg detection are the main data streams of the human tracking system. The human-oriented global nearest neighbour (HOGNN) data association, inspired from the Hall's proxemics, was developed to improve both the 3D camera-based and 2D laser-based human detection techniques. The dual Kalman filters are employed to track the human trajectory in parallel. The integration of the 3D camera-based and 2D laser-based human tracking is the key function of the system providing the real-time feedback for both the HOGNN to reduce false-positives of the camera-based and laser-based human detection and the Kalman filter to enhance the quality of the human trajectory tracking under uncertain environmental conditions. We implemented the sensor integration on ROS and validated it through real-world experiments.

### I. INTRODUCTION

Robust and accurate human detection and tracking is essential for a wide range of applications in robotics. This functionality is even important in robot navigation and manipulation when interacting and collaborating with humans in the shared workplaces [1], [2]. In recent years, human detection and tracking methods have been developed on a variety of sensors. In general, the sensors often used for human detection and tracking are Laser Range Finder (LRF), 2D and RGB-D cameras. Consequently, human detection and tracking methods can be roughly classified according to the used sensors including LRF based, 2D-image based, and RGB-D based techniques. Each human detection and tracking technique has advantages and disadvantages compared to the others and their robustness and accuracy is highly dependent on the features and characteristics of the sensor system and their appearance in a certain working environment.

Vision-based approaches for detecting humans and objects are very popular in recent years. Dalal et al. proposed Histogram of Oriented Gradients (HOG) method training a SVM classifier based on gradients which are densely pooled

into overlapping orientation bins [3]. With the growth of deep learning methods, object detection based on 2D images has been further improved. "You Only Look Once" (YOLO) method, proposed by Redmon et al. [4], is a real-time deep-learning-based object detection and bounding box generation around the ROI. This method has been demonstrated that it is comparatively faster and less computationally intensive than other deep learning-based object detection while its accuracy is almost double than the handcrafted feature-based object detection techniques [5].

While these vision-based techniques were successfully employed to detect humans and objects, they do not provide accurate ranging information of the detected objects, which is extremely important for mobile robot navigation and manipulation. The development of RGB-D cameras has helped to overcome this shortcoming and a number of human detection and tracking methods were recently presented [6], [7], [8]. Jafari et al. [7] proposed a RGB-D based real-time human detection and tracking technique through the switching between the depth-based upper body detector and the Histogram-of-Oriented-Gradient (HOG) full-body detector corresponding to the range of the object and the robot. Although these techniques are robust in detecting people in many cases, uncertain light conditions can affect their performance and range estimation because depth information of this technique is not reliable [9], [10].

LRFs are widely used for human detection and localization based on their remarkable distance measurement accuracy. The laser scanner provides ranging information below human knees due to their mounting position on the robot platform, e.g. 30cm, thus LRF-based human detection techniques are often referred as leg detection techniques. Leg detection techniques are designed by extracting geometrical features related to human legs and the supervised AdaBoost classifier is applied learn the shape of leg feature clustering based on 14 geometrical features [11], [12]. Lu et al. [12] extended the existing leg detection method and trained a random forest classifier to distinguish human from non-humans. Recently, Leigh et al. [13] proposed a joint-leg-detection method in which the behavior of each person was estimated by a Kalman filter and the scan-to-scan data association problem was addressed using a Global Nearest Neighbour technique.

In contrast to the single sensor-based methods mentioned above, a few studies incorporate multi-sensor data to detect people in a real environment. Truong et al. [14] proposed the RGB-D and LRF fusion for human detection using a particle filter and demonstrated this method with the socially aware robot navigation. Aguirre et al. [10] proposed a method using

\*We acknowledge the support of the Natural Sciences and Engineering Council of Canada, NSERC-RGPIN-2017-05446 and NSERC-CREATE 528099-2019.

<sup>1</sup>Hamed Bozorgi, Xuan Tung Truong, and Trung Dung Ngo are with The More-Than-One Robotics Lab (www.morelab.org), Faculty of Sustainable Design Engineering, University of Prince Edward Island, 550 University Ave, Charlottetown Prince Edward Island, Canada hbozorgi@upe.ca, xtruong@upe.ca, tngo@upe.ca

<sup>2</sup>Hung Manh La is with the Advanced Robotics and Automation (ARA) Laboratory, University of Nevada, Reno, USA. hla@unr.edu

the Mask R-CNN-based human detection learning algorithm to detect humans and highlight feature of leg regions on the relative LRF frame which is used to train a classifier capable of detecting human legs.

In this paper, we focus on developing a robust human trajectory tracking with the persistence of the identity of tracked people in cluttered environments. The key contribution of this study is the integration of 2D laser-based leg detection and 3D camera-based YOLO bounding box with the new techniques of *the human-oriented global nearest neighbour* and *the Kalman filter-based human tracking*. Specifically, the human-oriented global nearest neighbour plays the role as the data association and the Kalman filter is applied to track the human trajectory with both the LRF and 3D camera data. To take advantage of both the LRF and the 3D camera in order to deal with the environmental uncertainties, we integrated both the LRF-based and 3D camera-based human tracking and use it as the real-time feedback to improve the quality of both *the human-oriented global nearest neighbour* and *the Kalman filter-based human tracking*. Our experiments proved that the human tracking system is reliable and robust, thus it can be applied for real-time human trajectory tracking applications.

The remainder of the paper is organized as follows. Section II describes the existing human detection techniques based on RGB-D and LRF which set preliminaries of this study. Section III addresses the integration of multiple sensor channels for human trajectory tracking. The experimental results and analysis are discussed in Section IV. Finally, the conclusion of this paper is drawn in Section V.

## II. PRELIMINARIES ON HUMAN DETECTION

Human detection and tracking is a complex but crucial task, especially for unmanned vehicles when deployed in human-populated environments. While important, it has been proved that variation in human's appearance in the complex nature of dynamic and unstructured environments, such as illumination, clothing, and viewpoint, makes human detection and tracking a challenging task [15]. As each sensor mounted on the unmanned vehicle can only take a limited number of environmental characteristics into account, a multi-sensor system is required to increase robustness and reliability of the sensing and perceiving systems. In this study, we address a robust human detection and trajectory tracking by taking advantage of the state-of-the-art of the LRF and RGB-D methods.

### A. Human Detection using RGB-D

The RGB camera is the *vertical* sensor used to detect human in our system. Taking 2D image and depth information into account, a number of human detection techniques have been extensively studied and reviewed [16]. In this work, we chose to detect people on 2D images using the current state-of-the-art deep-learning technique, called YOLO (You Only Look Once), due to its higher accuracy but lighter computation in comparison with the traditional methods using handcrafted features [5], [4]. The method was used

to detect objects including people on the image plane and generates a bounding box around the identified objects.

In our work, detected people using YOLO detector at time step  $t$  are denoted as  $YO_t^C = [yo_t^{C1}, yo_t^{C2}, \dots, yo_t^{CO}]$  where  $O$  is the total number of people detected by YOLO within the field of camera view and each person has set of parameters representing the bounding box coordinate and class  $yo_t^{Ci} = (x_{min_t}^{Ci}, x_{max_t}^{Ci}, z_{min_t}^{Ci}, z_{max_t}^{Ci})$ . As the coordinate represents the person on the 2D vertical image plane, bounding box was mapped onto point-cloud data generated by the depth sensor to find person's location on the horizontal plane. Each person is then localized on the 2D plane at time  $t$  by:

$$\left( x_t^{Ci}, y_t^{Ci} \right) = \left( \frac{x_{min_t}^{Ci} + x_{max_t}^{Ci}}{2}, \frac{y_{min_t}^{Ci} + y_{max_t}^{Ci}}{2} \right) \quad (1)$$

where  $y_{min_t}^{Ci} = \min(\text{depth}(x_i, z_i))$  and  $y_{max_t}^{Ci} = \max(\text{depth}(x_i, z_i))$  within the acceptable range from depth of the center of the image bounding box, for  $x_i \in [x_{min}, x_{max}]$  and  $z_i \in [z_{min}, z_{max}]$ . The set of parameters for a person is determined by  $d_t^{Ci} = (x_t^{Ci}, y_t^{Ci})$  and the final matrix of  $O$  number of people localized on the RGB-D sensor at time  $t$  is denoted by  $D_t^C = [d_t^{C1}, d_t^{C2}, \dots, d_t^{CO}]$ .

Although YOLO is a very robust image object detector, the results of our real-time experiments in complex environments highlighted a couple of shortcomings. On the one hand, during experiments, we observed that YOLO may detect a single person multiple times in some frames and create bounding boxes with different sizes for the detected person, thus the bounding box size of the detected person is not reliable. On the other hand, it occasionally missed people presenting in front of the camera during the pattern recognition process. Last but not least, most RGB cameras have the limited field of view leading to insufficiency of using single RGB camera to detect and track human in complex environments.

### B. Human Detection using Laser Range Finder

In addition to the RGB-D, we take advantage of accuracy of LRF in measuring the distance to human and objects. To detect humans using LRF, we used the leg detection technique developed in [11], [12]. The technique was selected based on its low computational complexity due to the low number of geometrical handcrafted features for high localization accuracy in a simple environment. On the other hand, the algorithm suffers from false detections in crowded and complex environments and the performance of the algorithm is highly dependent on the density of laser beam of LRF leading to losing people tracking while eliminating the false detections using the confidence threshold criteria.

Detected people using the leg detector at time step  $t$  are denoted as  $D_t^L = [d_t^{L1}, d_t^{L2}, \dots, d_t^{LK}]$  where  $K$  is the total number of people detected by the leg detector in the surrounding area of the LRF and each person has a parameter set of  $d_t^{Li} = (x_t^{Li}, y_t^{Li})$ .

### III. INTEGRATION AND FILTERING FOR HUMAN TRAJECTORY TRACKING

Our proposed system is composed of the current state-of-the-art components and our developed functions as depicted in Fig 1. Using the people detection by the RGB-D-based YOLO to correct missing and incorrect information of the LRF-based leg detector where false positives occur and vice versa is the key methodology in this study. In addition, we embed the human proxemic model [17] to improve the global nearest neighbour and enhance the human trajectory tracking by the dual Kalman filters on both RGB-D and LRF channels with the observation from the sensor integration.

#### A. Human-Oriented Global Nearest Neighbour Data Association

In order to track all people appearing in the vicinity, the data association system should be considered to match all people that have been previously tracked,  $H_t^F$ , with new observations from the camera and LRF. At each scanning, the matrix of all the tracked people is denoted by  $H_t^F = [\hat{h}_{t|t}^{F1} \ \hat{h}_{t|t}^{F2} \ \dots \ \hat{h}_{t|t}^{FN}]$ , where  $N$  is the number of tracked people over time and  $\hat{h}_{t|t}^{Fi} = (x_t^{Fi}, y_t^{Fi}, v_t^{Fi}, \theta_t^{Fi})$  represents the estimated position, velocity and motion direction of the  $i$ th person at time  $t$ .

To address the observation-to-track matching problem, we used the global nearest neighbour (GNN) data association method with the Euclidean distance cost function which were used by [12] and showed promising results on a real system. The cost matrix is individually formed for each sensor detection between every new detection and tracking estimates  $H_t^F$ . The cost matrix of people observed by the camera at time  $t$  is denoted by:

$$Cost_t^C = \begin{bmatrix} d^{C1} \hat{h}^{F1} & d^{C1} \hat{h}^{F2} & \dots & d^{C1} \hat{h}^{FN} \\ d^{C2} \hat{h}^{F1} & d^{C2} \hat{h}^{F2} & \dots & d^{C2} \hat{h}^{FN} \\ \vdots & \vdots & \ddots & \vdots \\ d^{CM} \hat{h}^{F1} & d^{CM} \hat{h}^{F2} & \dots & d^{CO} \hat{h}^{FN} \end{bmatrix} \quad (2)$$

where each element of the matrix represents the cost between the corresponding camera observation and tracking estimate. Similarly, the cost matrix of people observed by LRF, denoted  $Cost_t^L$ , is formed.

Based on the GNN each observation is assigned to a track with the minimum cost. In this study, as the system is designed to track people, we embedded the proxemics model [17] into GNN to improve the human tracking. According to the Hall's model, the space around the centroid of a person is divided into four zones of intimate, personal, social, and public. Due to the fact that people rarely enter in each other's personal zone, we decided to take advantage of the personal zone formed around each person's state to restrict the GNN estimation in order to deal with unpredictable nature of the human movement. If an observation is located within an intimate zone of a previously matched track, it would be deleted, while if it is in the personal space, it is rechecked for the next best matching track. The human-oriented GNN

(HOGNN) with the observation of the leg detection filters out faulty of multiple bounding boxes generated by YOLO when detecting single person. Alternatively, the HOGNN also filters out the leg detector's false-positives inside the camera field of view based on the observations from the RGB-D camera. Finally, a track is deleted if it is not updated by any observations from either of the sensors for a given number of steps. After matching and filtering the false detections of both the camera and LRF, the matrix of camera-based human detection is identified as  $M_t^C = [m_t^{C1}, m_t^{C2}, \dots, m_t^{CO}]$  where  $O$  is the total number of validated people and the matrix of LRF-based human detection is identified as  $M_t^L = [m_t^{L1}, m_t^{L2}, \dots, m_t^{LK}]$ .

#### B. Human Trajectory Tracking

As shown in Fig 1, positioning of all the detected people is tracked individually using two parallel Kalman filters corresponding to the source of observation available for each person [18]. The tracking matrix of newly detected people in the camera view region would be initialized based on YOLO observation with zero initial velocity while the leg detector is responsible for initializing new detections for the non-camera view angle around the vicinity of the sensor system.

In this study, a constant velocity motion model was used to estimate each person's movement. The state estimate vector of each tracked person on the camera is presented as  $h_t^{Ci} = [x \ y \ \dot{x} \ \dot{y}]$ , where  $x$  and  $y$  is the position and  $\dot{x}$  and  $\dot{y}$  is the velocity in a 2D motion model. Based on the state matrix,  $h_{t-1|t-1}^{Fi}$ , as our best estimate of the position of the person calculated by the sensor integration at time  $t-1$ , the Kalman filter for the camera using a linear motion at time  $t$  is presented by:

$$\hat{h}_{t|t-1}^{Ci} = A_t \cdot \hat{h}_{t-1|t-1}^{Fi} \quad (3)$$

$$p_{t|t-1}^{Ci} = A_t \cdot p_{t-1|t-1}^{Ci} \cdot A_t^T + Q_t \quad (4)$$

where  $\hat{h}_{t|t-1}^{Ci}$  and  $p_{t|t-1}^{Ci}$  represent the posteriori state estimate and the covariance matrix on the camera tracker at time  $t$  based on previous observations, respectively.  $A$  denotes the state transition matrix modelling a linear 2D motion and  $Q = qI$  is the process noise covariance considered as a zero mean gaussian white noise. State of the system and covariance matrix is then updated on Kalman filter based on the associated observation on the camera,  $m_t^{Ci}$ , presented by:

$$\hat{h}_{t|t}^{Ci} = \hat{h}_{t|t-1}^{Ci} + K_t^{Ci} (m_t^{Ci} - H_t \cdot \hat{h}_{t|t-1}^{Ci}) \quad (5)$$

$$p_{t|t}^{Ci} = I - (K_t^{Ci} H_t) p_{t|t-1}^{Ci} \quad (6)$$

where  $H_t$  and  $K_t^{Ci}$  denote the observation matrix and Kalman gain, respectively. In this work  $H_t$  only includes position information from the corresponding observation source. In (5),  $m_t^{Ci} - H_t \cdot \hat{h}_{t|t-1}^{Ci}$  is called *innovation*. The Kalman gain matrix,  $K_t^{Ci}$ , and innovation matrix,  $S_t^{Ci}$ , for Kalman filter of the camera are defined as:

$$K_t^{Ci} = p_{t|t-1}^{Ci} H_t^T S_t^{-1} \quad (7)$$

$$S_t^{Ci} = H_t p_{t|t-1}^{Ci} H_t^T + R_t^{Ci} \quad (8)$$

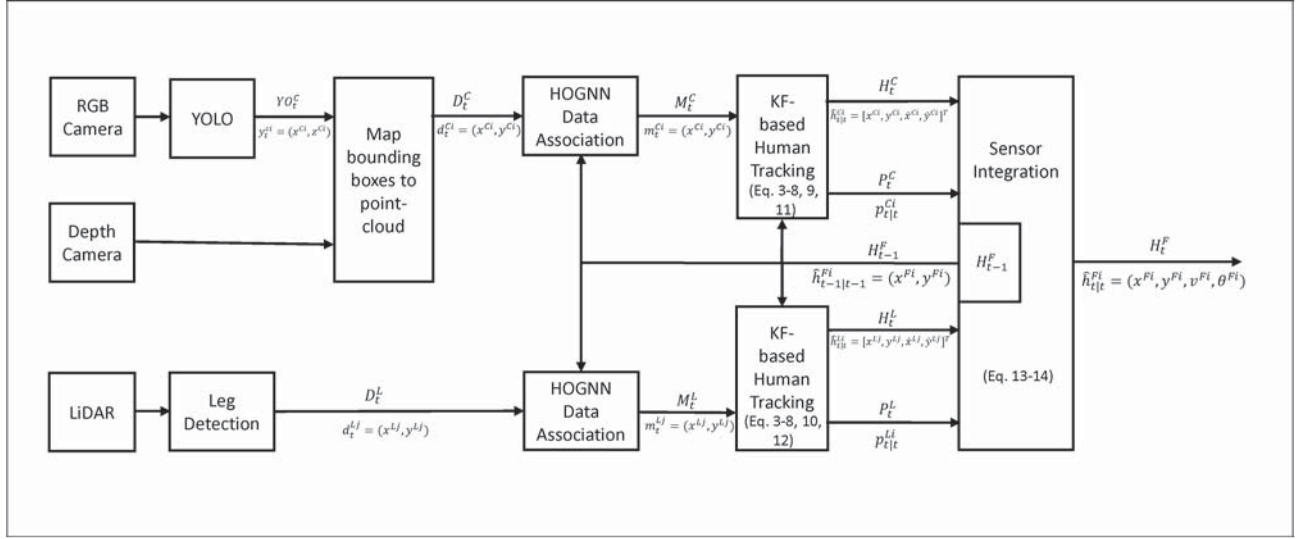


Fig. 1. Block diagram of the 2D Laser and 3D camera-based human trajectory tracking system.

where  $R_t^{C^i}$  is the covariance matrix of the observation noise. In this study, the measurement noise for each Kalman filter is considered as zero-mean Gaussian noise with the covariance of  $R = rI$ . Measurement noise covariance of people detected on each sensor is defined as an adaptive function composing both the initial noise of the sensor and the distance to the corresponding tracked position based on the observation. This adaptive noise on each of the RGB-D and LRF sensors is defined as:

$$r_t^{C^i} = \text{Initial Depth noise} + \frac{1}{1 + e^{-10(\text{dist}_t^{C^i} - 0.6)}} \quad (9)$$

$$r_t^{L^i} = \text{Initial LRF noise} + \frac{1}{1 + e^{-10(\text{dist}_t^{L^i} - 0.6)}} \quad (10)$$

This error function is constructed from two main sections of initial noise and adaptive part. The initial noise represents the intrinsic error of the sensor and the relevant positioning method. The adaptive part formulated by a normalized sigmoid function increasing gradually within the personal space ( $0.6 * 2 = 1.2m$ ) of the Hall's proxemics model [17] based on the distance between the observation and filter estimate. In this error function,  $\text{dist}$  is calculated for each sensor and observation separately as

$$\text{dist}_t^{C^i} = \sqrt{(x_t^{C^i} - x_{t-1|t-1}^{F^i})^2 + (y_t^{C^i} - y_{t-1|t-1}^{F^i})^2} \quad (11)$$

$$\text{dist}_t^{L^i} = \sqrt{(x_t^{L^i} - x_{t-1|t-1}^{F^i})^2 + (y_t^{L^i} - y_{t-1|t-1}^{F^i})^2} \quad (12)$$

Note that, while the variance of this error function is small for observations relatively close to the last information of the human position, it grows in a normalized form when  $\text{dist}$  increases. This, in particular, compensates in cases where the faulty data association may occur.

Similar to the Kalman filter for camera, the people detected by LRF are tracked using the Kalman filter for the LRF and the equations 3-8 are also applicable for the LRF-based tracking system.



Fig. 2. Experiment platform equipped with Microsoft Kinect and RPLIDAR-A3 laser range finder.

### C. Sensor Integration

As illustrated in Fig 1, if a person is detected by both the RGB-D camera and LRF, two distinct trackers would be formed. This would result in two different state estimations of  $h_{t|t}^{L^i}$  and  $h_{t|t}^{C^i}$  generated by the laser and camera tracker, respectively. Corresponding to each state estimate, the state covariance estimation matrixes of  $p_{t|t}^{L^i}$  and  $p_{t|t}^{C^i}$  are propagated, representing the uncertainty in each estimate.

In order to reach a final state estimate of the  $i$ th person, the two state estimations of  $h_{t|t}^{L^i}$  and  $h_{t|t}^{C^i}$  are combined through the following equation:

$$\hat{h}_{t|t}^{F^i} = w_t^i \cdot \hat{h}_{t|t}^{L^i} + (1 - w_t^i) \cdot \hat{h}_{t|t}^{C^i} \quad (13)$$

where  $w_t^i$  is a  $4 \times 4$  weight matrix for the  $i$ th person at time  $t$ . Taking advantage of the uncertainty generated by each tracker,  $w_t^i$  is optimally defined as

$$w_t^i = \frac{p_{t|t}^{L^i}}{p_{t|t}^{L^i} + p_{t|t}^{C^i}} \quad (14)$$

Doing so, the human state estimate of the  $i$ th person trajectory obtained by the 2D laser and 3D camera-based human tracking system at time  $t$  can be represented as

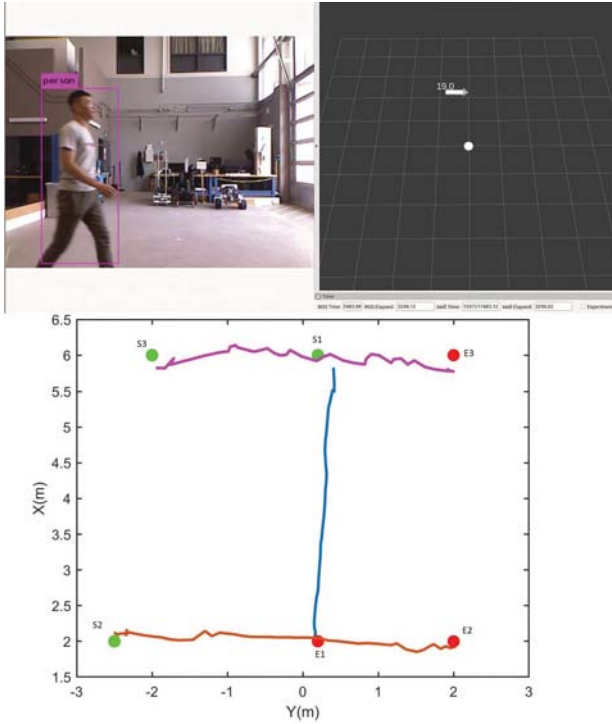


Fig. 3. One person walking scenario: Three trajectories related to three experiments are shown in blue, orange, and magenta. Green points show start points where goals are depicted in red. A frame from a person passing in close-range in front of the robot is depicted in the top image where the location of the robot is shown by a white circle. The arrow shows the person with its relative position, velocity, and movement orientation.

$\hat{h}_{t|t}^{F^i} = (x_t^{F^i}, y_t^{F^i}, v_t^{F^i}, \theta_t^{F^i})$  including position  $(x_t^{F^i}, y_t^{F^i})$ , velocity  $v_t^{F^i}$ , and motion direction  $\theta_t^{F^i}$ , respectively.

#### IV. IMPLEMENTATION & EXPERIMENT RESULTS

##### A. Experimental Platform

To demonstrate and verify our proposed method, we have implemented and tested the system on our Eddie robot platform equipped with the first-generation Microsoft Kinect sensor and the laser range finder as depicted in Fig 2. The standard Kinect sensor includes an RGB camera, an infrared light projector, a depth sensor and multi-array microphone positioned at the height of  $1.35m$  from the ground. The range of the depth sensor is  $0.8m$  to  $6.0m$  with the vertical and horizontal viewing angles of  $43^\circ$  and  $57^\circ$  respectively. This sensor provides image and depth information with the resolution of  $640 \times 480$  pixels at the maximum frame rate of 30 frames per second. Experimental results showed that the error of depth measurement on the Kinect sensor increases quadratically in proportion to the distance from the sensor and reaches  $4cm$  at the maximum range of  $5m$  [19]. RPLIDAR-A3 was used as the laser range finder in this study. This LiDAR system can perform a 2D scan within a  $0.2-25m$  range and  $360^\circ$  field of view with a dense spatial resolution ( $0.3375^\circ$  angular resolution) at  $15Hz$  typical scan rate. In all the experiments, camera frame was set at the default and laser measurements were calibrated accordingly. We implemented the proposed method using Python programming language

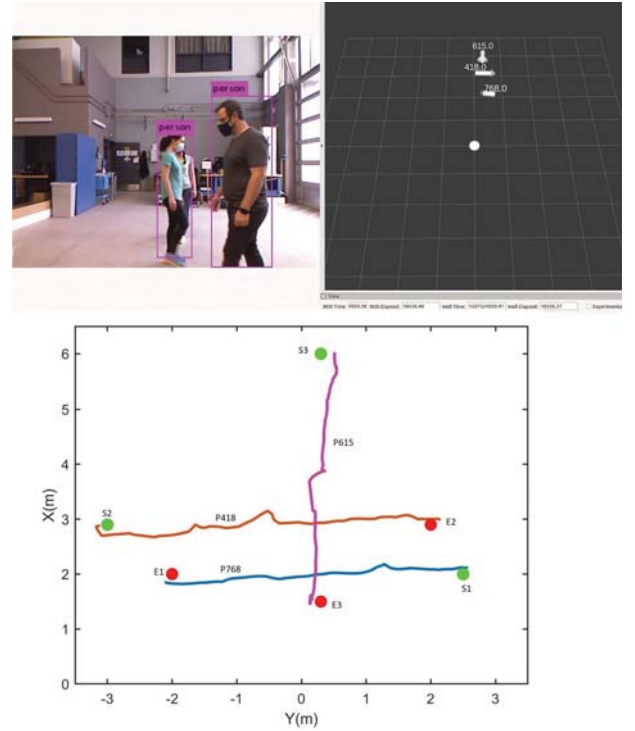


Fig. 4. Three people walking scenario: Three trajectories related to each person are shown in blue, orange, and magenta. Green points show start points where goals are depicted in red. A frame from a people passing each other while sharing the personal zone and obstructing the observation of each other is depicted in the top image where the location of the robot is shown by a white circle. The arrow shows each person's position, velocity, and movement orientation generated by the proposed tracking method.

with Robot Operating System (ROS) [20] on the intel Core i7 2.6 GHz laptop with NVIDIA Geforce RTX 2060 graphics card. The overall tracking process ran in realtime with 10 FPS. The video demonstration can be seen at the link <sup>1</sup>

##### B. Experiments and Discussion

1) *One Person Walking*: The set of experiments of one person walking was designed to validate the accuracy of human trajectory tracking according to both the  $x$  and  $y$  axis: 1) a person walks toward the robot, and 2) a person walks in front of the robot. Due to the fact that the noise of sensors proportional to the distance between the sensor and detected objects, the second experiment were separated into two experiments to validate the tracking at the close range and at the maximum range of the Kinect sensor. The start and end point of these experiments are denoted by  $(S1, E1)$ ,  $(S2, E2)$ , and  $(S3, E3)$  as illustrated in Fig 3.

In the first experiment, a person walked toward the robot from  $S1 = (6.0, 0.2)$  to  $E1 = (2.0, 0.2)$ . As the person moved along the  $x$  axis, the variation of the localization along the  $y$  axis were recorded. The results showed that the average and standard deviation of the human localization along the  $y$  axis is  $0.2479m$  and  $0.0809m$ , respectively, and the  $RSME$  is  $0.093$ . In the second experiment, the person

<sup>1</sup><https://youtu.be/XeQtOLLjuK4>

walked in front of the robot from  $S2 = (2.0, -2.5)$  to  $E1 = (2.0, 2.0)$ . As the person moved in the  $y$  axis, the level of alteration was recorded in the  $x$  axis. In this case, average and standard deviation of the human localization along the  $x$  axis was  $2.059m$  and  $0.073m$ , respectively and the  $RSME$  is  $0.0941m$ . In the last experiment, the person walked in front of the robot from  $S3 = (6.0, -2.0)$  to  $E3 = (6.0, 2.0)$ . The average and standard deviation along  $x$  axis were measured as  $5.922m$  and  $0.1082m$  while the  $RSME$  showed  $0.1321m$ . As this experiment is performed on the edge of the Kinect sensor range, in 52% of the overall 48 frames, reliable depth measurement was not available in which the system completely depended on the LRF observations.

### C. Three People Walking

As shown in Fig 4, the experiment scenario was designed to evaluate the performance of the human-oriented GNN data association in a cluttered environment. This experiment includes three people with the ID of P768, P418 and P615 walking simultaneously from  $S1 = (2.0, 2.5)$ ,  $S2 = (2.9, -3.0)$ , and  $S3 = (6.0, 0.2)$  toward goals at  $E1 = (2.0, -2.0)$ ,  $E2 = (2.9, 2.0)$ , and  $E3 = (1.5, 0.2)$ , respectively. One challenge in this scenario was about people sharing of their personal space when they passed by each other, e.g., the relative distance between P768 and P418 dropped down to  $0.8m$ . Moreover, as people passed each other, they obstructed the observations of the person behind them. We found that out of the overall 96 frames, 10.4% of the camera and 8.3% of LRF observations, were blocked by P768 and P418 and no observation was available at 3.1% of the frames. We also identified that 4.1% of camera and 1.04% of LRF observations of P418 were obstructed by P768.

We validated the experimental result to confirm that the developed system maintained the ID of people throughout the tracking process. In addition, the human trajectory tracking was accomplished when one source of the observation was not available by switching to other observation, or even estimating the trajectory based on the previous steps when no observations were recorded. Last but not least, the results shows that false positives generated by YOLO algorithm at 4.1% of 96 frames, were filtered out thanks to the human-oriented GNN data association.

## V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a robust human trajectory tracking by integrating the 3D camera and 2D LRF. Human-oriented GNN plays the role as an observation-to-track data association on each sensor separately to enhance the robustness of the matching thanks to the personal space of the proxemics model. The human trajectory of detected people is estimated by integrating two parallel Kalman filters with the observation noise adaptive to conditions of the two separate observations of the 3D camera and LRF. The system integration was successfully tested under different scenarios in a real-life environment. The experiments showed the system is able to robustly maintain human trajectory tracking in both single and multi-people cluttered environment. In future

works, we have planned to identify a confidence factor for each tracked person to further increase the system robustness and accuracy. Moreover, we will develop an algorithm to further enhance the detection and tracking of human groups in the real world.

## REFERENCES

- [1] T. Kruse, A. K. Pandey, R. Alami, and A. Kirsch, "Human-aware robot navigation: A survey," *Robotics and Autonomous Systems*, vol. 61, no. 12, pp. 1726 – 1743, 2013.
- [2] X.-T. Truong and T.-D. Ngo, "Dynamic social zone based mobile robot navigation for human comfortable safety in social environments," *International Journal of Social Robotics*, vol. 8, no. 5, pp. 663–684, Nov 2016.
- [3] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *International Conference on Computer Vision & Pattern Recognition (CVPR '05)*, C. Schmid, S. Soatto, and C. Tomasi, Eds., vol. 1. San Diego, United States: IEEE Computer Society, June 2005, pp. 886–893.
- [4] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018.
- [5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 779–788.
- [6] M. Munaro and E. Menegatti, "Fast rgb-d people tracking for service robots," *Auton. Robots*, vol. 37, no. 3, pp. 227–242, Oct. 2014.
- [7] O. H. Jafari, D. Mitzel, and B. Leibe, "Real-time rgb-d based people detection and tracking for mobile robots and head-worn cameras," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 5636–5643.
- [8] T. Linder and K. O. Arras, "Multi-model hypothesis tracking of groups of people in rgb-d data," in *17th International Conference on Information Fusion (FUSION)*, 2014, pp. 1–7.
- [9] H. Song, W. Choi, and H. Kim, "Robust vision-based relative-localization approach using an rgb-depth camera and lidar sensor fusion," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 6, pp. 3725–3736, 2016.
- [10] E. Aguirre and M. García-Silvente, "Using a deep learning model on images to obtain a 2d laser people detector for a mobile robot," *International Journal of Computational Intelligence Systems*, vol. 12, pp. 476–484, 2019.
- [11] K. O. Arras, B. Lau, S. Grzonka, M. Luber, O. M. Mozos, D. Meyer-Delius, and W. Burgard, *Range-Based People Detection and Tracking for Socially Enabled Service Robots*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 235–280.
- [12] D. V. Lu and W. D. Smart, "Towards more efficient navigation for robots and humans," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 1707–1713.
- [13] A. Leigh, J. Pineau, N. Olmedo, and H. Zhang, "Person tracking and following with 2d laser scanners," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 726–733.
- [14] X. Truong, V. N. Yoong, and T. Ngo, "Rgb-d and laser data fusion-based human detection and tracking for socially aware robot navigation framework," in *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Dec 2015, pp. 608–613.
- [15] L. Spinello and R. Siegwart, "Human detection using multimodal and multidimensional features," in *2008 IEEE International Conference on Robotics and Automation*, 2008, pp. 3264–3269.
- [16] Z. Zhao, P. Zheng, S. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, Nov 2019.
- [17] E. Hall, *The Hidden Dimension*. Anchor Books, 1992. [Online]. Available: <https://books.google.ca/books?id=p3g0ngEACAAJ>
- [18] R. E. Kalman, "A new approach to linear filtering and prediction problems," *ASME Journal of Basic Engineering*, 1960.
- [19] K. Khoshelham and S. O. Elberink, "Accuracy and resolution of kinect depth data for indoor mapping applications," *Sensors*, vol. 12, no. 2, pp. 1437–1454, Feb 2012.
- [20] M. Quigley, K. Conley, B. P. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "Ros: an open-source robot operating system," in *ICRA Workshop on Open Source Software*, 2009.