Deep Learning-based Multiple Objects Detection and Tracking System for Socially Aware Mobile Robot Navigation Framework

Do Nam Thang[1], Lan Anh Nguyen[1], Pham Trung Dung[1],
Truong Dang Khoa[1], Nguyen Huu Son[1], Nguyen Tran Hiep[1], Pham Van Nguyen[1],
Vu Duc Truong[1], Dinh Hong Toan[1], Nguyen Manh Hung[1], Trung-Dung Ngo[2], Xuan-Tung Truong[1]

***Abstract*— Multiple objects (including humans) detection and tracking system plays an essential role in socially aware mobile robot navigation framework. Because, it provides an important input for the remaining modules of the framework. In this paper, we propose an efficient multiple objects detection and tracking system for mobile service robots in dynamic social environments using deep learning techniques. The proposed system consists of two steps: (1) multiple objects detection, and (2) multiple objects tracking. In the first step, the RGB image-based multiple objects detection is made use of to detect objects in the mobile robot's vicinity using a convolutional neural network. In the second stage of system, the detected objects are tracked using a deep simple online and realtime tracking technique. The experimental results indicate that, the proposed system is capable of detecting and tracking multiple objects including humans, providing significant information for the socially aware mobile robot navigation framework.**

I. INTRODUCTION

Mobile service robots are becoming more and more popular in dynamic human environments, both in public and private places, such as museums, airports, offices and the home, shopping malls, and urban environments. However, the widespread acceptability of the mobile service robots in our daily life is hindered by robot's inability to navigate in the dynamic human environments in socially acceptable manners that would guarantee the human safety and comfort. Thus, developing socially aware mobile robot navigation frameworks are necessary in recent years. These frameworks enable the mobile service robots to navigate autonomously, safely and socially in the dynamic social environments, providing the safety and comfort for the humans, and the socially acceptable behaviors for the mobile robots in two essential tasks: (1) avoiding humans, and (2) approaching humans.

To solve that problem, several human-aware mobile robot navigation systems have been proposed in recent years in order to make sure the safety and comfort of the human and generate socially acceptable behaviors for the mobile robots [1], [2] and [3]. In such social navigation frameworks, the human detection and tracking system is the essential component, because it provides an important information for the rest sub-systems of the socially aware navigation framework of the mobile robot [4]. That is, in dynamic social environment the mobile robot should be able to detect people, and extract the their states including the human's pose and the human's motion. To accomplish that, researchers have been recently proposed several human detection and tracking systems. These human detection and tracking systems can be divided into three categories corresponding to the data utilized as the input of the detection system: (i) human detection system based on laser data, (i) RGB images-based methods and (ii) RGB and depth images-based techniques. In the first category of the system, the author solely utilized the data from the laser range finder as the input of the human detection frameworks [5]. In contrast, in the second category of methods, RGB images from webcams and cameras are utilized for detecting existence of the people in the images [6]. Although these techniques have achieved considerable success, they do not make use of one important piece of information that is now available-depth. In order to address these shortcomings, in the third group, researchers utilize the RGB-D data as the input of their algorithms [7]. In contrast to the aforementioned human detection and tracking approaches, some methods utilize a combination of techniques to detect the existence of the people in human-centered environments, such as laser rangefinder data and visual image information-based fusion algorithm [8].

However, in order to make sure the safety and comfort of the human in the dynamic social environments the mobile service robot should recognize not only humans but also the human–object interaction information in its surrounding environment, then embeds those information into the robot's navigation system [9]. Therefore, beside detecting humans, it is necessary to detect interesting objects in the robot's surrounding environment such as televisions, refrigerators, telephones, screens, and paintings, etc. To the best of our knowledge, none of existing mobile robot perception system is able to detect and track multiple objects including humans for mobile service robot navigation in human-centered environments. Thus, in this study, we propose a multiple objects detection and tracking system for the socially aware navigation framework of the autonomous mobile robot based on deep learning techniques.

The rest of this paper is organized as follows. Section II describes block diagram of the proposed socially

[1] Xuan-Tung Truong, Do Nam Thang, Lan Anh Nguyen, Pham Trung Dung, Truong Dang Khoa, Nguyen Huu Son, Nguyen Tran Hiep, Pham Van Nguyen, Vu Duc Truong, Dinh Hong Toan and Nguyen Manh Hung are with Le Quy Don Technical University, Hanoi, Vietnam xuantung.truong@gmail.com

[2] Trung-Dung Ngo is with The More-Than-One Robotics Laboratory, School of Sustainable Design Engineering, University of Prince Edward Island, Canada dungnt@ieee.org

978-1-5386-7983-8/18/$31.00 ©2018 IEEE

aware navigation framework for mobile service robots in dynamic human-centered environments. Section III presents the proposed multiple objects detection and tracking system. Section IV discuss the experimental results of a conducted experiment, and Section V presents the conclusions of this study.

II. Socially Aware Mobile Robot Navigation Framework

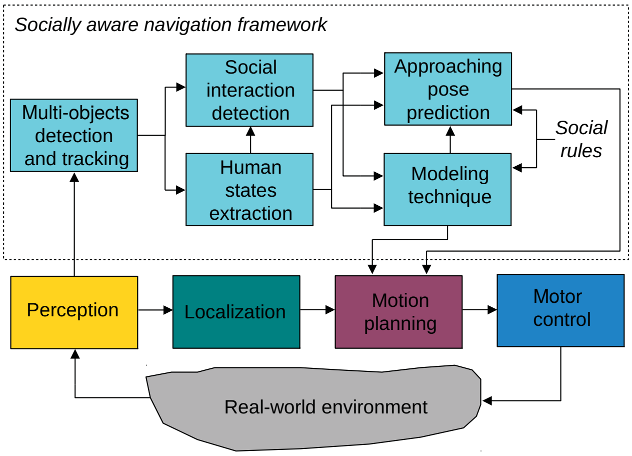

Socially aware navigation framework
Multi-objects detection and tracking
Social interaction detection
Human states extraction
Approaching pose prediction
Modeling technique
Social rules
Perception
Localization
Motion planning
Motor control
Real-world environment

Fig. 1. The socially aware navigation scheme for mobile service robots in dynamic human environments.

It has been known that, the human-centered environments are dynamic, unstructured and uncertain with existence of interesting objects and people. Thus, in order to make sure the safety and comfort of humans in these environments the mobile service robot should be able to detect multiple objects including humans, extract features from their characteristics, and then embed those information into mobile robot navigation system. In order to do that, in this paper, we propose an efficient mobile robot navigation framework based on the conventional navigation system presented in [10] by adding an additional socially aware navigation framework, as shown in Fig. 1. As can be seen in Fig. 1, the proposed social navigation framework is composed of two essential parts: (i) the conventional navigation framework, and (ii) the proposed socially aware navigation framework (in the dash line box). In the first part, the navigation system consists of four basic functional blocks include, perception, localization, motion planing and motor control. In the second part of the proposed framework, the socially aware navigation framework is utilized to firstly detect multiple objects including people and interesting objects. The framework then extracts social characteristics of the humans including human pose, human motion, human group and human–object interaction in the surrounding environment of the mobile robot. Those characteristics are then made use of for the development of the modeling technique, and the approaching pose prediction of a person, a group of people and a human–object interaction. The proposed socially aware mobile robot navigation framework is presented in detail in the next paragraphs.

Multi-objects detection and tracking system: This functional block is used to detect and track multiple objects including humans in surrounding environment of the mobile robot. The outputs of this block is the human position (x_i^p, y_i^p), interesting object position (x_j^o, y_j^o), the human movement orientation θ_i^{vp}, the object movement orientation θ_j^{vo}, the human velocity v_i^p, and the object velocity v_j^o. Thus, the human state extracted by this system is $p_i = (x_i^p, y_i^p, \theta_i^{vp}, v_i^p)$, and the object state is $o_j = (x_j^o, y_j^o, \theta_j^{vo}, v_j^o)$.

Human states extraction system: This functional block is utilized to detect and track 3D human pose in the robot's vicinity, such as the left hand position (x_i^{lh}, y_i^{lh}), the right hand position (x_i^{rh}, y_i^{rh}), and the head orientation θ_i^{hp} of the person p_i in the $xy - plane$. Thus, the human state could be written as $p_i = (x_i^p, y_i^p, \theta_i^{vp}, v_i^p, x_i^{lh}, y_i^{lh}, x_i^{rh}, y_i^{rh}, \theta_i^{hp})$. This information is then used to identify the social interactions in the next paragraph.

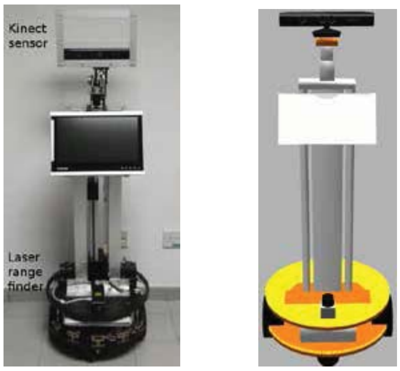
Social interaction detection system: It has been known that, in dynamic social environments, the behaviour of human is usually impacted by other humans and surrounding objects, especially the interesting objects, such as paintings, advertising screens, televisions, refrigerators, etc. In addition, in [11] the authors found out that 70% of people in dynamic social environments wish to form groups to interact with each others. Therefore, it is not efficient enough if the mobile service robot does not consider these information. To this end, to make sure the safety and comfort of the humans, it is necessary for the mobile service robots to perceive these social interaction contexts, extract their features, and incorporate them into the mobile robot's navigation system. The social interaction in dynamic human-centered environment can be divided into two categories including human group interaction or human–object interaction [4].

Modelling technique block: Several techniques have been used to model the human and social interaction features [1], [2]. In this paper, we utilize dynamic social zone model proposed by Truong et al. [9].

Approaching pose prediction system: This functional block is utilized to estimate the proper approaching pose including the approaching position, orientation and velocity and of the mobile service robot to a person, a people–object interaction, or a people group [4].

Social rules: Every country has a system of manners, that is, social rules for behaviour. Each particular culture has a history of acceptable behaviour such as passing on the right or left hand side. In the social environment, the people teach these manners to their children. Thus, when you visit a new place, you need to learn the manners for that place. To this end, in order for the accepted behaviour of the mobile robots, we should incorporate the social rules into the navigation system of the robots.

System integration: The outputs of the modeling technique and the approaching pose prediction blocks are then incorporated into the conventional navigation scheme of the mobile robot, particularly the motion planning system. Therefore, the entire navigation system is able to guide the



(a) (b)

Fig. 2. Our mobile robot platform: (a) Real world mobile robot platform equipped with Microsoft Kinect sensor and laser rangefinder, (b) 3D model of our mobile robot.

mobile robot to navigate towards the predicted approaching goal pose, while providing human comfort and safety for the people around the mobile robot.

III. PROPOSED MULTIPLE OBJECTS DETECTION AND TRACKING SYSTEM

In human environment, mobile robots need to perceive not only humans and human groups but also surrounding objects, such as televisions, refrigerators, telephones, screens, and paintings. Because, to ensure the human comfort, it is necessary for the mobile robot to identify human-object interaction, which is utilized to define an interaction space between humans and interesting objects. Thus, multiple objects (including humans) detection and tracking system plays an essential role in mobile robots, which are deployed in dynamic human environments.

A. The Used Mobile Robot Platform

We designed a mobile robot platform, which is equipped with a laser scanner and a Microsoft Kinect sensor, as shown in Fig. 2(a). The laser scanner is UGR-04LX-UG01, and positioned at the height of 0.4[m]. It can measure the distance up to 6.0[m], and provide the angular field of view of 240^o. Whereas, the Microsoft Kinect sensor can provide booth the RGB image and depth image, and it is positioned at the height of 1.35 [m] from the ground. The resolution of the RGB image is 640 x 480 pixels, and the maximal frame rate is 30 frames per second. This low-cost hardware can provide the depth image range from 0.8[m] to 6.0[m] with the horizontal viewing angle of 57^o and the vertical viewing angle of 43^o. In this paper, we utilize the RGB images from Microsoft Kinect sensor as the input of the proposed multiple objects detection and tracking system. Whereas, the depth images are used to define the distance from the robot to detected objects.

B. Multiple Objects Detection System

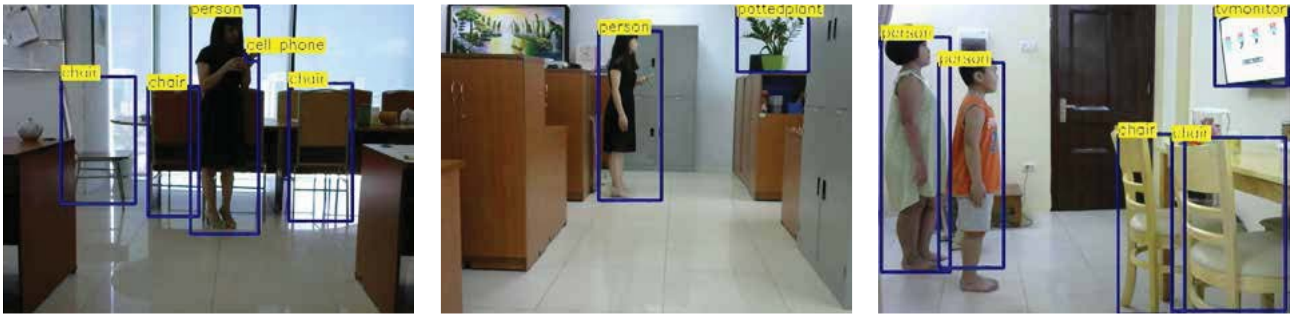
The aim of the multiple objects detection system is to detect multiple objects including humans in the surrounding environment of the mobile robot. This information is very important for the robot to ensure the safe and social navigation in dynamic social environments. Therefore, detection system need to be fast and accurate. To accomplish that, in this paper, the you only look once (YOLO) object detection algorithm proposed by Redmon et al. [12] is adopted. Because YOLO technique balances between the speed and the accuracy. To do that, instead of applying the model to an image at multiple locations and scales, like conventional region-based convolutional neural network [13], YOLO model applies a single convolutional neural network to the full image for both classification and object localization tasks.

The YOLO algorithm firstly divides an input image into *SxS* grid cells. Thus we should convert the input image to square image. Each cell in the grid has responsibility to recognize the objects, which their center positioned in the cell in the image coordinates. To accomplish that, each cell is utilized for predicting the position of *B* bounding boxes and their corresponding confidence scores, and the probability of object class conditioned on the presence of the object in the bounding box.

Each bounding box is determined by a tuple of $(x, y, w, h, score)$, where the (x, y) coordinates is the relative center of the bounding box to the corresponding cell; w and h are the relative width and height of the bounding box to the width and height of the image, respectively; and *score* is the confidence score. To accomplish that, the x and y values are derived by normalizing the center position of the bounding box by the coordinates of the top-left corner of the corresponding cell. And the w and h values are obtained by normalizing the width and height of the bounding box to the width and the height of the image, respectively. Therefore, their values are range from 0 to 1. The confidence score indicates the accuracy of the bounding box and whether the bounding box actually contains an object.

Suppose that the YOLO algorithm can identify C object classes. In this case, if there is an object in a grid cell, then the cell can estimates a probability of this object belonging to class C_i, where $i = 1, 2, \dots, C$. In addition, the YOLO technique solely estimates one set of class probabilities per grid cell. Therefore, in total, there are *SxSxB* bounding boxes in an input image. Each bounding box contains 2 relative center positions, 2 relative width and height, 1 confidence score, and C conditional probability for identifying the object class. Finally, there are $SxSx(5 * B + C)$ prediction values in total for an input image. However, in the input image, there are only some objects. Therefore, YOLO has to reduce number of the bounding boxes. To do that, YOLO utilizes the threshold for the object confidence score and the non-maximum suppression technique for the redundant bounding boxes. A detailed description of the YOLO technique was given in [12].

As a result, the output of the multiple objects detection system is the detected objects with corresponding bounding boxes. These information are then fed into the multiple objects tracking in the next section.



(a) (b) (c)

Fig. 3. The example results of the multiple objects detection and tracking system.

C. Multiple Objects Tracking System

Multiple objects tracking is usually utilized in many applications, such as action detection and recognition, surveillance and security systems [14], and especially in social robot in recent years [9]. The objective of a tracking system is that, it is be able to connect detections across video frames into trajectories. To accomplish this, the tracking system consists of two main phases, including prediction and correction. In the prediction phase of the system, the tracker predicts new locations of existing tracks in the current video frame. In the second phase, the tracking system should (1) assign new detections in the current frame to existing tracks and update motion model of these tracks, (2) decide which new detections should be considered as the first detections of new tracks, and (3) delete existing tracks that have been invisible for some consecutive frames, using association algorithm.

In recent years, deep learning algorithm has made inroads into the field of multiple objects tracking to improve the performance of the tracker [15]. To exploit this advantage, in this paper, we adopt the multiple objects tracking algorithm named simple online and realtime tracking with a deep association metric (Deep-SORT) proposed by Wojke et al. [16]. The Deep-SORT technique is an extension of the simple online and realtime tracking (SORT) algorithm proposed in [17] by integrating the appearance information, which is based on a deep appearance descriptor to reduce the large number of identity switches in the SORT method.

The Deep-SORT algorithm firstly utilizes the bounding box information detected from the multiple objects detection system as a input. Then, an association algorithm is performed to link these detections to existing tracks using motion and appearance information. To accomplish that, the state space of each object is defined by a tuple of $(u,v,\gamma,h,\dot{u},\dot{v},\dot{\gamma},\dot{h})$ where, the position of the bounding box center (u,v), the aspect ration γ, the height of the bounding box h, and the overdots mean their corresponding velocities in the image coordinates. Then, a linear Kalman filter is utilized to predict the aforementioned state. In order to assign new detections to existing tracks, initialize new tracks and delete lost tracks, the authors of Deep-SORT proposes a matching cascade algorithm, in which the motion and appearance information are used. In particular, to incorporate the motion information of the object, the authors utilize the Mahalanobis distance to compute the distance between newly arrived detections and predicted Kalman states. In addition, for each bounding box of the existing tracks, Deep-SORT calculates a feature vector using a pretrained appearance descriptor named wide residual network [18] and keeps a gallery of the last associated appearance descriptors for each track. Then the distance between the existing tracks and new detections in visual appearance space is computed using simple nearest neighbor queries. Finally to solve the sudden visual changes of the objects, the intersection over union association is utilized on the set of unmatched and unconfirmed tracks of age 1. A detailed presentation of the Deep-SORT technique was given in [16].

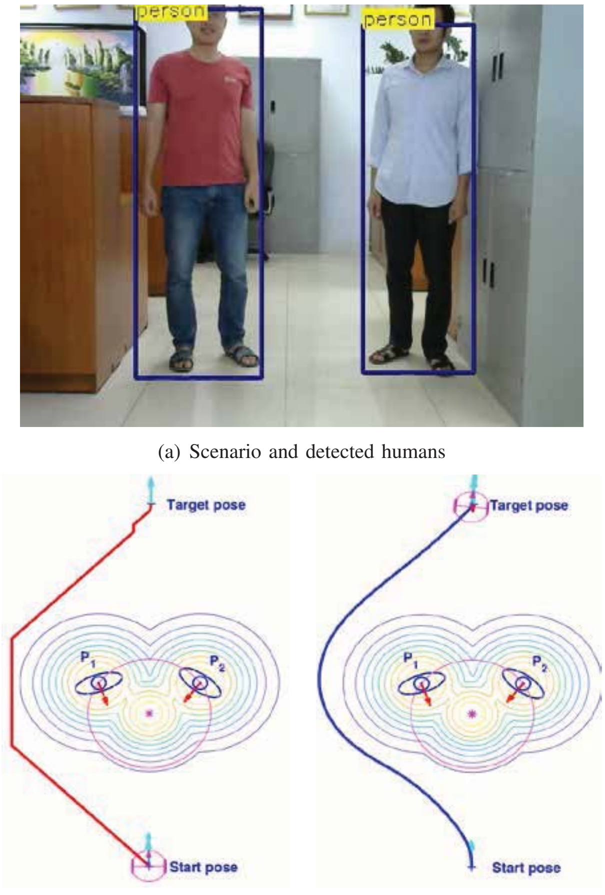
It has been known that, the motion of objects in the social environments is very dynamic, especially humans. Leading to their motion model is not always linear. Therefore, nonlinear motion model is made use of in this paper. In other words, instead of using linear Kalman filter like the original Deep-SORT algorithm, we utilize the extended Kalman filter to handle the motion prediction of the objects.

Figure 3 shows example results of the proposed multiple objects detection and tracking system using deep learning techniques. As can be seen in Fig. 3(a), the proposed system can detect and track chairs, a cellphone and a standing person. Whereas, a standing person and a potted plant is detected and tracked by the proposed system in Fig. 3(b). The proposed multiple objects detection and tracking system can detect chairs, a television, and two people watching the television, as shown in Fig. 3(c).

IV. EXPERIMENTAL RESULTS

A. Experiment Setup

In this paper, to verify and demonstrate the usefulness and feasibility of the proposed multiple objects detection and tracking system, we have implemented the proposed system on our robot hardware platform, as seen in Fig. 2(a). We also implement the entire navigation system in Fig. 1 on this platform. However, in this paper, we only show the output of the motion planing block, and do not control the mobile robot platform. In other words, we only show the legible trajectory of the mobile robot in this experiment. To



(a) Scenario and detected humans

(b) Legible robot's path

(c) Legible robot's trajectory

Fig. 4. The example result of the proposed multiple objects detection and tracking system and the output of the motion planing system.

accomplish that, the astar path planing algorithm [19] and the dynamic window approach technique [20] are made used of in the motion planing block in Fig. 1.

We implemented the proposed system utilizing MATLAB, C/C++ and Python programming languages. We also used the Robot Operating System (ROS) [21], the OpenCV Library [22]. The entire system are tested on an Intel core i7 2.2 GHz laptop.

B. Experimental Results

An experiment was conducted to verify the proposed multiple objects detection and tracking framework. The results are depicted in Fig. 4, in which the first row shows the image scenario and the output of the proposed multiple objects detection and tracking system, whereas the second row shows the output of the motion planing system.

In this experiment, the robot makes use of the Kinect sensor to perceive the surrounding environment. The RGB images are used as input of the proposed multiple objects detection and tracking system. Whereas, the depth images are utilized to measure the distance between the robot and the detected objects. As can be seen in Fig. 4(a), the proposed multiple objects detection and tracking system can detect and track two standing people. The output of the system is the position and the velocity of the people. These information are then modeled in the modeling technique block in Fig. 1 using the dynamic social zone presented in [9]. The output of the modeling technique block is then incorporated into the motion planing block. Then, the motion planing system generates the legible path, as shown in Fig. 4(b), and the legible trajectory of the mobile robot, as shown in Fig. 4(c). The legible trajectory is then used to guide the mobile robot to navigate socially and safely around human.

V. CONCLUSION

In this paper, we have presented an efficient multiple objects detection and tracking system for mobile service robots in dynamic social environments using deep learning techniques. The proposed system is comprised of two steps. In the first step, the RGB image-based multiple objects detection is used to detect objects in the surrounding environment of the mobile robot using a convolutional neural network. In the second stage of system, the detected objects are tracked using a deep learning technique. The experimental results illustrate that, the proposed framework is capable of detecting and tracking multiple objects including humans, providing significant information for the socially aware mobile robot navigation framework.

ACKNOWLEDGMENT

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01-2018.10.

REFERENCES

[1] T. Kruse, A. K. Pandey, R. Alami, and A. Kirsch, "Human-aware robot navigation: A survey," *Robotics and Autonomous Systems*, vol. 61, pp. 1726–1743, 2013.

[2] J. Rios-Martinez, A. Spalanzani, and C. Laugier, "From proxemics theory to socially-aware navigation: A survey," *International Journal of Social Robotics*, September 2014.

[3] X. T. Truong and T. D. Ngo, "Toward socially aware robot navigation in dynamic and crowded environments: A proactive social motion model," *IEEE Transactions on Automation Science and Engineering*, vol. 14, pp. 1743–1760, October 2017.

[4] X. T. Truong and T. D. Ngo, "To Approach Humans?: a unified framework for approaching pose prediction and socially aware robot navigation," *IEEE Transactions on Cognitive and Developmental Systems*, pp. 1–27, 2017.

[5] K. O. Arras, O. M. Mozos, and W. Burgard, "Using boosted features for the detection of people in 2d range data," in *IEEE International Conference on Robotics and Automation*, pp. 3402–3407, April 2007.

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

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

[8] 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, pp. 167–181, Feb 2009.

[9] 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*, pp. 1–22, 2016.

[10] R. Siegwart, I. R. Nourbakhsh, and D. Scaramuzza, *Introduction to Autonomous Mobile Robots*. The MIT Press, February 2011.

[11] M. Moussaid, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz, "The walking behaviour of pedestrian social groups and its impact on crowd dynamics," *PLoS ONE*, vol. 5, no. 4, p. e10047, 2010.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, June 2016.

[13] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, June 2017.

[14] A. Yilmaz, O. Javed, and M. Shah, “Object tracking: A survey,” *ACM Computing Surveys*, vol. 38, pp. 1–45, Dec. 2006.

[15] A. Milan, S. H. Rezatofighi, A. Dick, I. Reid, and K. Schindler, “Online multi-target tracking using recurrent neural networks,” in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 4225–4232, 2017.

[16] N. Wojke, A. Bewley, and D. Paulus, “Simple online and realtime tracking with a deep association metric,” in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3645–3649, IEEE, 2017.

[17] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple online and realtime tracking,” in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3464–3468, 2016.

[18] S. Zagoruyko and N. Komodakis, “Wide residual networks,” *arXiv:1605.07146*, vol. abs/1605.07146, 2016.

[19] P. E. Hart, N. J. Nilsson, and B. Raphael, “A formal basis for the heuristic determination of minimum cost paths,” *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, pp. 100–107, July 1968.

[20] D. Fox, W. Burgard, and S. Thrun, “The dynamic window approach to collision avoidance,” *IEEE Robotics Automation Magazine*, vol. 4, pp. 23–33, March 1997.

[21] M. Quigley, B. Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Ng, “ROS: an open-source Robot Operating System,” in *ICRA Workshop on Open Source Software*, vol. 32, pp. 151–170, 2009.

[22] G. Bradski, “The OpenCV Library,” *Dr. Dobb’s Journal of Software Tools*, 2000.