University **Politehnica** of Bucharest
Faculty of Automatic Control and Computers
Department of Computer Science

PhD THESIS - Summary

A Predictive Approach to People Tracking in Autonomous Applications

Author:

Alexandra Ștefania Ghiță

Thesis Advisor:

Prof. Dr. Eng. Adina Magda Florea

PhD Thesis Committee

Chairman	Prof.Dr.Eng. Florin Pop	University Politehnica of Bucharest
Thesis Advisor	Prof.Dr.Eng. Adina Magda Florea	University Politehnica of Bucharest
Member	Prof.Dr.Eng. Rodica Potolea	Technical University of Cluj-Napoca
Member	Prof.Dr.Eng. Costin Bădică	University of Craiova
Member	Prof.Dr.Eng. Ștefan Trăușan-Matu	University Politehnica of Bucharest

BUCHAREST

2022

Abstract

Currently, the importance of autonomous operating devices is rising with the increasing number of applications that can benefit from them. Autonomous applications, such as robotic platforms or self-driving cars, assume that platforms operate without human intervention in environments where people move or perform their daily activities.

A critical component for autonomous devices is the ability to track and re-identify the same people through a sequence of images in a short amount of time, in order to generate safe and reliable behaviours. In this thesis, we introduce a real-time people tracking and re-identification system based on a trajectory prediction method. We combine the output of a trajectory prediction method with a simple tracking technique to create a stable and accurate system. We tackle the problem of trajectory prediction by introducing a system that incorporates semantic information from the environment with social influence from the other participants into the motion of each individual, to predict the most probable trajectories. We evaluate the systems considering different possible case studies, namely social robotics, autonomous driving and self-operating drones. For the context of social robotics, we present a social robotics framework, designed in a modular and robust way for assistive care scenarios. The framework includes robotic services for visual understanding, navigation, multi-lingual natural language interaction and dialogue management, as well as activity recognition and general behaviour composition. For all three case studies, we perform multiple experiments to validate our approach considering both the trajectory prediction component and the person re-identification system. We focus on both qualitative and quantitative evaluation by integrating existing related datasets and self-acquired data.

Contents

Chapter 1

Introduction

Artificial intelligence is constantly developing and improving, reaching new peaks in all domains. The spectrum of applications it can be applied to is expanding, having a significant impact on people and on their quality of life. The progress made in research allows more complex projects to be developed, allowing autonomous operating devices, which can help or replace activities made by humans, to be more compelling and more robust. This thesis tackles multiple problems in the context of autonomous applications by taking into consideration their real-time requirements.

1.1. Motivation

Autonomous devices are a recent trend in the research area considering the multitude of places where they can be deployed. Autonomous vehicles and social robots are the most prevalent applications that are currently being studied, and they are constantly being improved, given the advancements in artificial intelligence. People tend to depend more and more on devices that can make their life easier, by reducing a part of their daily tasks. Social robots, and in particular assistive social robots, have a variety of purposes, as they can be used as assistant for people in need, tourist guides in museums, checkpoint information in shops, data collectors in hospitals, etc. Autonomous vehicles help by reducing the stress and anxiety people feel when driving. Every device that can raise the quality of life of individuals is currently being improved to more stable and robust systems. The development of the research, alongside the wide applicability of such systems, influenced the decision to implement a project that can generate a system appropriate for autonomous devices.

The challenges that arise in the context of autonomous devices come from a variety of factors: unforeseen events, sensors malfunctions, limited information,

user understanding, etc. One of the most challenging factors for an autonomous device is the human himself. Humans have a very complex and unpredictable nature, so such a device must be able to cope with the sudden changes, to be able to react accordingly when an unexpected situation appear. In the context of social assistive robots, besides the complex behavioural nature, humans have a lot of variation in terms of physical aspect (facial features, height, clothing, etc.) or voice (tone, accent, speed), making the interaction far more difficult. In the context of autonomous driving, the visual appearance is changing significantly due to the fast movement of the vehicle. The ability to overcome these challenges to obtain a secure autonomous system was one of the main motivations behind this project.

Given the general usage of an autonomous system, a component for people detection and tracking is essential. Autonomous systems in general are designed for helping people and reducing a part of their responsibilities. As a result, the systems need to map the environment where they operate and understand the behaviours of nearby agents, in order to execute their tasks safely and with minimum errors. The people tracking component is relevant for many reasons: can re-identify the same person in a sequence of images, can analyse the behaviour to extract patterns, can predict the movement to prevent possible complications. The importance and the variety of functions that such a component implies, was a motivation for choosing to design a real-time people tracking system that fits the requirements of an autonomous device.

1.2. Objectives

The thesis is focusing on implementing systems suitable for autonomous applications, in particular for socially assistive robots. Our aim is to integrate a self-designed people tracking system into a framework we designed for robotic devices. The robotic framework integrates multiple capabilities to create a compelling platform. Its objective is to be a general platform, applicable for a variety of social robots in assistive scenarios. It offers capabilities in terms of visual understanding, navigation and vocal interactions, which are combined to define complex behaviours. The people tracking component enhances the functionality of such a platform. Our objective is to design a system which is not limited to only social robotic applications, but can be applied for any autonomous device, including self-driving cars and self-operating drones. We

designed a trajectory prediction component which estimates the future trajectories of the observed people, which we integrated into the people tracking component, to create an accurate and robust system.

Consequently, the main objectives of the project are:

1. Develop a robotic framework which integrates general capabilities for a robotic platform, to create complex behaviours that can be applied in social robotics scenarios.

2. Design and implement a method for real-time people trajectory prediction which integrates the inner behaviour of the people, social influence and environmental information.

3. Design and implement a real-time people tracking and re-identification system based on the people trajectory prediction method.

4. Validate the proposed systems in multiple autonomous contexts: social robotics, self-driving car, drones.

1.3. Thesis Outline

The thesis begins by presenting the problem that the project is approaching, and details the relevant work associated with the problem, together with the related datasets. It follows by presenting the robotic platform introduced for creating complex behaviours for user interactions. It then explains the two systems introduced for analysing people behaviours, namely people trajectory prediction and people re-identification and tracking, validated in the context of social robotics. Finally, the thesis presents the validation of the systems in other autonomous contexts: self-driving cars and drones.

The chapter-by-chapter outline of the thesis is described below.

- In Chapter 2 we define the problems tackled by this thesis in terms of relevance, research and engineering work. The chapter defines the problems of social robotics, people trajectory prediction and people tracking, alongside the associated research challenges.

- Chapter 3 reports an extensive analysis of existing related work on all the problems addressed in this thesis. The chapter analyses different existing systems, by presenting the corresponding advantages and disadvantages in the context of autonomous applications.

- In Chapter 4 we present multiple existing datasets related to autonomous applications. We analyse the datasets from the point of view of quality, resolution and angle view. We also introduce the self-acquired data we collected to evaluate our system in other scenarios.

- Chapter 5 introduces a robotic framework designed for easy development of robotic applications. The framework integrates data from multiple sensors to generate specific capabilities for a social robot. The capabilities can be combined into complex behaviours dedicated for socially assistive scenarios.

- Chapter 6 introduces a new method for people trajectory prediction. We designed an architecture with combines information about the movement of the people with the social influence coming from the other participants and environmental information extracted from the scene. The system is validated on a social robotic context, using an existing dataset and self-acquired data.

- Chapter 7 proposed an architecture for real-time people re-identification and tracking based on the people trajectory prediction method. The system re-identifies people based on their predicted future movement, allowing re-identification after occlusion or camera movement. The system is validated considering a social robotic application, based on an existing dataset and self-acquired data.

- In Chapter 8 we validate our systems for people trajectory prediction and people re-identification and tracking on other autonomous contexts, namely self-driving cars and drones. We evaluate the methods using existing related datasets and on self-acquired data, to include more variation in the analysed scenarios.

- Chapter 9 concludes the work, by outlining the main contributions of the thesis and the perspectives in terms of future development.

Chapter 2

Problem Definition

The objective of this chapter is to define the problems tackled by the thesis and put them in the context of autonomous applications. The goal of the thesis is to develop a people tracking system based on a trajectory prediction approach and validate it in autonomous contexts. For the social robotics context, the thesis introduces a robotic framework designed for creating robotic behaviours suitable for socially assistive scenarios. Considering this structure, this chapter defines the problems of social robotics, trajectory prediction and people tracking, alongside the associated research challenges they involve.

Socially Assistive Robotics [1] refers to robots that are meant to assist people in a manner that focuses on social interactions (e.g., speaking, guiding, reminding, observing, and entertaining). Though physical interaction (e.g., carrying of objects) may be enabled by certain kinds of robot, it is not mandated by the mentioned definition.

One of the most focused domains of application for socially assistive robots (also referred to as *companion* robots) is that of supporting the elderly population, particularly people who are living alone or in care institutions, as well as those who are affected by medical conditions which warrant a closer monitoring of daily habits. The Active and Assisted Living (AAL) domain, which concerns itself with developing technology to support the needs of the aforementioned aging population, is therefore actively sustaining development of the capabilities of companion robots.

People trajectory prediction is the method that estimates the possible followed paths of every tracked person. Every estimated trajectory is computed based on previous observations, which are represented by the positions of the person in the previous moments of time.

The problem formulation behind trajectory prediction is represented by the

ability to estimate X_{gt} based on X_{obs}. X_{obs} represents the observed trajectory of a person, from the moment of time *1* to *o*, where *1* denotes the first element of the sequence and *o* is the number of the observed positions based on which the prediction is made. X_{gt} represents the real trajectory followed by the individual, from the moment of time *o+1* to *o+p*, where p is the number of elements in the prediction horizon. The system we propose aims to generate X_{pred}, an estimation of X_{gt}, such that the errors are minimal.

The predicted future trajectories of the people in the images provide a critical piece of information for autonomous applications. People trajectory prediction is integrated in the architectures of autonomous devices for safety reasons, especially when referring to self-driving cars which operate at higher speeds. The information provided by this system can be a decisive element in preventing undesirable situations.

People re-identification and tracking is a technique that allows systems to identify the same person in a sequence of images. The technique assigns a unique identification number to every person detected in an image. By tracking a person for multiple frames, a system can better understand the behaviour of a person, can associate actions, or plan a better interaction.

From a theoretical point of view, the problem of people re-identification assumes assigning a unique identification number $x_{i:n}$ to every bounding box $y_{i:n}$ representing a person, detected at the moment of time *n*. If at the moment of time *n*, a bounding box $y_{j:n}$ contains the same person as a bounding box $y_{k:n-r}$ from the image acquired at the moment of time *n-r*, where *r* represents a number of frames, then the associated identification number $x_{j:n}$ must match the one that was previously associated, namely $x_{k:n-r}$.

The information provided by a people re-identification system is essential for many autonomous applications. Social robots must differentiate between people to adjust their behaviour according to the behaviour of the person. Self-driving cars must track people to be able to predict their movement and perform safely. Security systems, such as surveillance drones, must track people to analyse possible dangerous situations. The system is fundamental in many scenarios, so accurate results are required.

Chapter 3

Related Work

Given the current improvement in the artificial intelligence domain, the research area is focusing on the development of smart systems alongside autonomous devices that can help people in their lives. One of our goals is to develop a general system suitable for any robotic platform capable of engaging in human-robot interactions. An important part of this research is focusing on computer vision problems which are required for such a system. The analysis of the existing research was done by taking into consideration these areas.

Social robots are designed to interact with people in a natural way, having intention like humans. A number of research projects have developed solutions, comprising a diverse set of functionalities, from more specific ones [2, 3], to general systems [4, 5, 6], to initiatives (e.g. STRANDS [7]) that support development of technologies for long-term autonomy of robots.

Systems such as NAOqi provide rich development capabilities, but are limited to particular robots. Frameworks such as [3, 8] are more focused on the social interactions, but do not provide a means for goal driven development of the robot life cycle or arbitrary behaviour composition. The EnrichMe project [5] targeted aiding the independent living of single older adults via smart-home, robotics and web technologies. While it is overall similar to our project in design principles, functionality modules, testing procedures and qualitative evaluations, it is missing a more comprehensive test of navigation capabilities, as well as means to detect more diverse user actions. The SocialRobot project [6] focused on developing a custom robot for elderly care, but the authors do not report a more systematic performance of individual functionality modules (neither in the live deployment, nor in a lab setting).

As trajectory prediction is a valuable method that improves user experience and safety when talking about autonomous devices, the subject is widely researched in

the field of artificial intelligence. The pioneering work conducted in *Social LSTM* [9] was the starting point for many directions in the research of person trajectory prediction, by integrating a social pooling layer between multiple recurrent neural networks. The idea was later developed and improved in many papers, such as [10, 11, 12], by combining networks, such as generative adversarial networks or deep neural networks, to increase the precision. The main disadvantage of these systems is that the predictions do not consider environmental information, which can alter the motion of an individual.

Papers such as [13, 14, 15, 16] moved a step forward and improved the results of previous systems by introducing scene information into their systems. More recent approaches, such as systems in [17, 18], are using goal-based prediction methods, which are sampling multiple possible goal candidates to chose the best fit from that set of candidates. The main disadvantage of these methods is that they need global scene information from bird-eye view images. Our proposed method overcomes this limitations and generates trajectories using images acquired from an eye-level point of view, allowing dynamic environments with changes in terms of angles and visual cues.

The problem of person tracking and re-identification is a complex problem that has been intensively researched for a significant amount of time [19]. One of the early approaches for person re-identification can be traced back to 1997 [20]. The current approaches propose complex systems to solve the problem of people re-identification. In research such as [21, 22], the architectures integrate temporal data to identify the people, while systems such as [23, 24] perform sophisticated feature matching for re-identification.

The general applications of people re-identification systems usually require real-time processing. One of the popular systems for online people tracking is *DeepSort* [25, 26]. It combines visual information extracted using a convolutional neural network with a simple position estimation technique. *Tracktor* [27] performs online multi-tracking using a simple technique based on a regressor of an object detector. The maine disadvantage of these techniques, especially for the *Tracktor* system, is that they require a high framerate of the images.

In our proposed method we combined the advantages of the *DeepSort* system with the advantages of a trajectory prediction system, to create a more powerful technique. The system overcomes the limitations concerning small framerates, as the introduced trajectory prediction module can adjust the estimation of trajectories based on the speed of the target.

Chapter 4

Relevant Datasets

In order to design an appropriate architecture for the task of human tracking, a preceding phase was required to discover and analyse existing datasets. As one of the goals of the research is to track people based on visual information, the investigated datasets contain images of people, in both indoor and outdoor environments.

The datasets were selected considering the final goal of this thesis, which is people tracking in autonomous applications. The relevant examined characteristics of the datasets are the view angle of the camera and its mobility (fixed or moving). In terms of view angle, the image streams are captured from a bird-eye view angle or an eye-level view angle. In terms of mobility, there are static cameras datasets and moving cameras datasets.

In the thesis we integrated 6 datasets: *MOT17: Multi-Object Tracking* [28], *JRDB* [29], *ETH - BIWI Walking Pedestrians* [30], *UCY - Crowds by example* [31], *Caltech-Pedestrians* [32] and *Stanford Drone* [33].

MOT17: Multi-Object Tracking [28] and *JRDB* [29] are used to evaluate the behaviour of the system in assistive robotic scenarios. *MOT17: Multi-Object Tracking* is a big challenging dataset, with mixed view angles and environments. In some the scenes, the camera moves in correspondence with the carrying robotic platform. The number of people and obstacles in the frames varies depending on the scene. The *JRDB* dataset is acquired from a moving robotic platform with multiple cameras and sensors, and it contains more complex information. The dataset contains images in both indoor and outdoor scenes, with varying number of people in frames. Depending on the scene, the images may include obstacles, such as chairs, tables, trees, poles. The *ETH - BIWI Walking Pedestrians* [30] and *UCY - Crowds by example* [31] datasets contains images collected in different outdoor scenarios, with images acquired from a top fixed point in both scenes. The images present crowded scenarios with sparse,

small obstacles, such as trees, pillars, benches. The characteristic of these datasets is that the coordinates of the people are represented as real values expressed in meters, and do not represent the pixel coordinates.

The *Caltech-Pedestrians* dataset [32] is a large dataset which contains images acquired from a moving car. The purpose of this dataset is to be integrated in applications using autonomous vehicles. The images are oriented towards the lanes of the road, with the tracked pedestrians being mostly on the side of the images. The particularity is that the camera is moving with a variable speed and also the view angle on the scene is changing according to the movement of the car.

Stanford Drone [33] is a wide dataset which contains multiple videos acquired from different outdoor areas in a university campus. The images are acquired from a very high point of view. The characteristic of this dataset is that the bounding boxes associated with the tracked people are small, given the view point. The dataset is used to validate the approach on images acquired using a drone.

In order to evaluate our approach on more particular scenarios, we acquired additional data to include more challenging situations. We acquired multiple videos for two different autonomous contexts: robotic platform and self-driving car. To generate the associated bounding boxes representing the positions of the people in the images, we used an existing system for people detection, namely YOLO [34], which we applied for every image in the data collection.

The videos collected from the point of view of a robot are acquired in both indoor and outdoor environments, in multiple scene settings. We collected 19 videos in total, with a resolution of 640x480 pixels and a framerate of 30 frames per second. The data is acquired with an external camera and includes little to no movement during the recording. For this particular case we also recorded an additional video using our robotic platform in an indoor environment. The video has a resolution of 320x240 pixels and a framerate of 5 frames per second. For the context of autonomous driving we collected 12 different videos. The camera is fixed inside a moving car and its movement is determined by the movement of the vehicle. The speed varies from full stop to 50km/h and the angle of the camera changes accordingly to the direction of the car. The videos have a resolution of 640x480 pixels and a framerate of 30 frames per second.

In addition to the two contexts, we collected 5 videos from a top point of view to simulate the view of a flying drone. The camera is fixed and is recording data from a single point of view. The data is acquired in two scenes, one indoor and one outdoor.

Chapter 5

The AMIRO Robotic Framework for Social Robots

Autonomous robotic platforms are integrated in a variety of applications that require minimum to no human intervention. Given the vast spectrum of possible tasks and functions such a machine can accomplish, it is understandable that there are a multitude of different platforms built upon some distinct and specific systems. Social robots in particular, though they share similar characteristics in terms of roles and objectives, can have very particular implementations. The totality of differences among the existing systems produces two major disadvantages: it complicates the designing process for engineers, as they need to learn and adapt to every new encountered system, and it eliminates the possibility of easily migrating the behaviour or the function of one robotic platform to another.

One of the purposes of this research is to develop a social robotics platform, suitable for the Pepper robot [35], but necessarily applicable to any Robot Operating System (ROS) [36] compatible robot. The AMIRO (Ambient Robotics) system [37] is a robotic framework that aims to provide a platform that can be easily integrated with a variety of robots. The modular design of the platform is allowing the possibility to integrate new behaviours or update existing one with minimal effort. The developed platform operates on an architecture that follows the recent trends of edge and cloud-based robotics. It enables a comprehensive set of functionality modules that facilitate complex behaviour composition.

5.1. Overview and Architecture

The proposed framework addresses the major requirements for a social assistive robot, while maintaining a high degree of modularity in building and integrating

new modules inside the system. The main components are detached, as each component can be run as a separate application. This characteristic ensures the robustness of the system, as it can operate even in the case that one the modules fails. The main components of the system are listed below.

- Visual understanding - integrates multiple computer vision techniques, such as people and object recognition, people re-identification, trajectory prediction and the activity recognition components. When an object detection occurs, the module also computes the 3D position of the object relatively to the robot, which is then forwarded to the *Navigation* component to place it on the map.
- Navigation and obstacle avoidance - is responsible with the movements of the robot inside the environment. This module performs data acquisition and is responsible with the processing of the SLAM algorithms to generate the paths to be followed.
- Speech interpretation and dialogue - is responsible with interpreting the speech of the user and with Text-to-Speech capabilities when required. The module can send new tasks based on the current dialogue with the user.
- Integration with smart systems - gathers data about user health (blood pressure, heart rate, steps, sleep) and environment (room temperature, humidity, luminosity), and provides the necessary instruments to actuate different smart devices (e.g. smart lighting, smart blinds).
- Behaviour composition - is at the centre of the system and publishes commands to all the other modules. It generates behaviours based on the received commands. It is subscribed to the central storage to acquire useful data, such as previously mapped objects. The guiding principle is that of easiness and robustness in bringing together the basic behaviour functionality aspects of the other modules.

The architecture of the system is built on top of the ROS framework. Each main component of the architecture offers a set of ROS publishers and subscribers, which allows the data to be exchanged between the modules continuously and asynchronously. From a module deployment perspective, the AMIRO architecture distributes its services on machines running in the cloud or constituting the cloud-edge. The advantage of such an architecture is the fact that each node can be run on separate machines, allowing the separation of

concerns while facilitating the deployment of the system. As such, the robotic platform is used only for data acquisition and actuation. The proposed structure for the system is justified by the real-time requirements of the project. Using the cloud-edge modules, we eliminate the computational requirements of the integrated robotic platform, and we provide strong computational capabilities, for more complex techniques. We chose to use local data processing instead of cloud processing by taking into consideration the amount of information needed to be transmitted over the internet. For fast responses in the case of the visual information, the bandwidth of the internet connection should be large enough to not add latency.

5.2. Experiments

To evaluate the general functioning of the framework we tested several experimental scenarios, that are plausible in the context of a social robotic application. We deployed the system using the Pepper social robot [35] to validate the AMIRO framework in real legitimate conditions. The robotic platform is responsible with data acquisition, such as visual information and audio data, as well as data output, such as motor movement. We proposed two evaluation scenarios that illustrate a part of the basic functionalities of the integrated framework: voice recognition and comprehension, visual person finding, position estimation of a detected target, human action recognition, environment exploration and navigation, smart environment system interaction, actions planning and execution.

The chosen scenarios prove the applicability of the framework in two separate situations: single person scenario and multi-person scenario. In the first scenario, the robot interacts with a single person that it assists in basic activities. In the second scenario, the robot is used in multi-person environments, requiring more advanced recognition and planning techniques. The scenarios were evaluated using the Pepper robotic platform in laboratory conditions. One important advantage of our proposed framework is that it is platform independent, which allows integration with a variety of robotic platforms. The experiments we performed to assess the performance of each functionality module are based on the input coming directly from the Pepper robot. This allowed us to specifically gauge functionality limitations that are due to the Pepper robotic platform itself, or ones which can be mitigated through future improvements of our modules.

Chapter 6

People Trajectory Prediction System

People trajectory prediction is the ability to analyse the behaviour of a person and deduce their future movements based on the observations. Although it seems straightforward, modelling the complex process behind human behaviour is a difficult task. Trajectory prediction can be integrated in a variety of applications: robotics, autonomous driving, people tracking, surveillance, etc. With the development of autonomous devices, precise results are imperative in realizing a specific goal or preventing undesirable events. Trajectory prediction systems are required to take into consideration multiple factors in order to obtain an accurate result. As the trajectory of a person is influenced by the movement of the other people in the neighbourhood, obstacles in the environment and points of interest in the scene, our proposed technique incorporates social influences between people and environmental context together with the positions of the target to generate the final results.

6.1. Overview and Architecture

The architecture of the system consists mainly of two modules: a module for scene understanding and a module for generating the predicted trajectory. The structure of the architecture is shown in Figure 6.1. The *Scene Understanding* module extracts the visual data associated with the received input image, generating information about the visible scene, such as obstacles or pathways, alongside the corresponding position of the person. The *Trajectory Generation* module integrates the generated visual data to estimate the most plausible trajectory, based on the scene settings, the other participants, and the movement of the person.

To generate the predictions, the proposed architecture integrates three factors that can influence the trajectory of an individual: inner behaviour, social

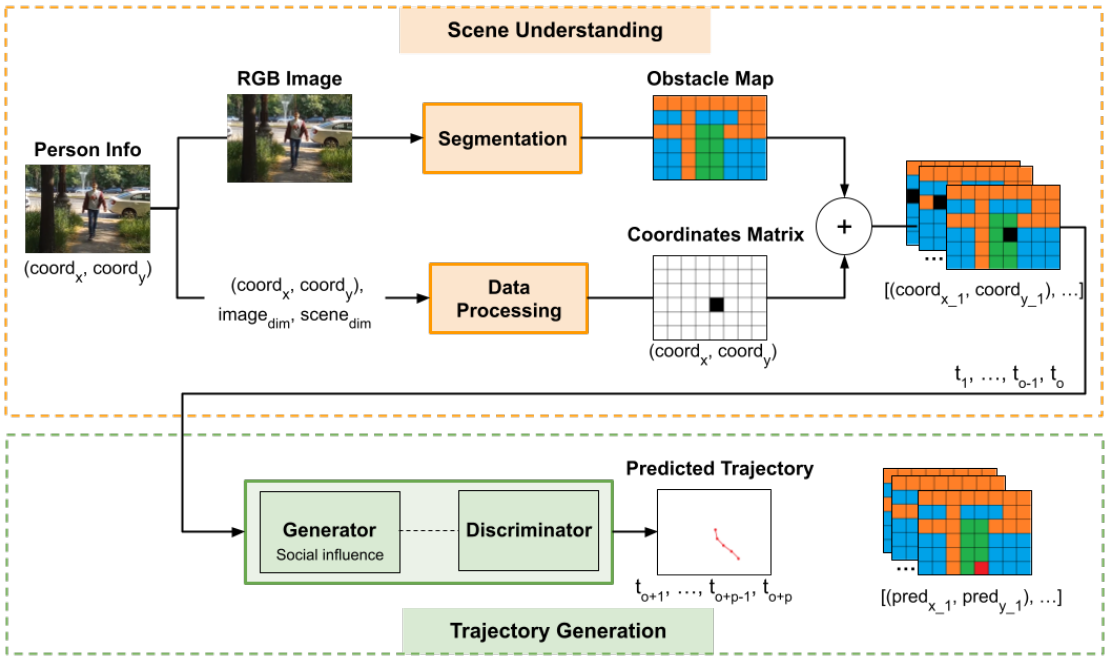

Figure 6.1: Social−GAN with Obstacle Map architecture. The aggregated data is computed for each person in the images.

influence and scene settings. The inner behaviour of a person refers to the previous positions in which the person was observed. In Figure 6.1, the inner behaviour is modelled by the 2D coordinates alongside the corresponding coordinates matrices. Social influence refers to the movement of the other participants in the scene, that can influence the trajectory of a person, by avoiding the intersection of trajectories. In the architecture, the social influence is encoded by a pooling layer in the *Trajectory Generation* module, which transfers information between the people in the neighbourhood. The scene settings are represented by the layout of the environment, in terms of obstacles and possible pathways. This information is encoded by the obstacle map, which is generated based on the segmentation mask of the input image.

The information associated with a person which is passed as input to the system is composed of a mix of the coordinates representing the position of the person at each time step, alongside the RGB images of the scene acquired at that exact time. Based on this information, the system computes two additional pieces of information: the pixel coordinates matrix of the person and the obstacle map of the environment. The system generates a volume of data for each person detected in the image, and is then passing the volumes combined to the generative adversial network.

6.2. Evaluation and Validation

The trajectory prediction system was evaluated by analysing the results obtained on relevant existing datasets and by interpreting the performance of the component when integrated into the AMIRO framework. The evaluation was performed by taking into consideration social robotics scenarios. For the integrated datasets, the errors reported for one subset were obtained by training the network against the rest of the subsets, with prediction sequences of length 8.

The main objective evaluation was performed based on the JRDB dataset, as it is relevant for the autonomous social robotics context. We reported the values in pixels obtained for two standard error metrics, ADE, average displacement error, and FDE, final displacement error. The average error values that we obtained are 23.69 pixels for the ADE metric and 36.34 for the FDE metric, which are small enough for a satisfactory behaviour of the robotic platform. In addition, we evaluated the system on the ETH and UCY datasets, to position the work with respect to other existing systems, even though the method was designed to work on pixel coordinates in images acquired from an eye-level point of view.

For an analysis of the impact of the trajectory prediction system in the context of social robotics applications, we also defined several scenarios which were deployed using the AMIRO framework. We analysed the behaviour of the robot in several possible situations, as follows:

1. The robot is helping a person by providing directions to a specific location.
2. The robot is working as a tourist guide in a museum.
3. The robot is helping people with guidance during an emergency evacuation.

Each described scenario combines existing capabilities from the AMIRO framework with the functions provided by the *People Trajectory Prediction* module. Besides the elementary function of estimating the future trajectories, the framework can detect sooner if a person is approaching a specific target, which can avert unsafe situations or can speed up the interaction process.

Even though it presents some limitations in particular cases, the information provided by the *People Trajectory Prediction* module can boost the understanding of the people and can improve the behaviour of the robot. In terms of processing speed the module is suitable for real-time applications, generating results in around 110 milliseconds regardless of the number of predicted trajectories.

Chapter 7

Real-Time People Tracking and Re-Identification System

A person re-identification system is assigning the same identifiers to the exact same people in a succession of images. The characteristics of such a component are not limited only to social robotics applications, but can be extended to any autonomous device. Whether we are talking about autonomous vehicles, robots, surveillance or smart home environments, being able to detect and recognize the same person through a sequence of images as well as distinguish between multiple individuals is a key constituent.

The problem of person re-identification has a complex nature as the performance of such a system can be strongly impacted by multiple factors. To design a system that performs well in a variety of scenarios implies to create a mechanism that is able to overcome problems like occlusions or appearance variation. These problems arise from natural circumstances such as people passing behind obstacles while moving, similar styles in terms of clothing and hairstyle or interactions between individuals. To handle these types of problems our method integrates information generated by a trajectory prediction system.

7.1. Overview and Architecture

We propose a modular system that tackles the problem of real-time person tracking by combining a standard technique for people tracking with a people trajectory prediction method. The architecture takes as input a stream of images and predicts an identifier number for each identified person, more specifically to each bounding box detected in the input image. Figure 7.1 presents the layout designed for the real-time re-identification system. The proposed architecture can be integrated into a robotic platform system or any autonomous device as it requires only RGB data to generate results.

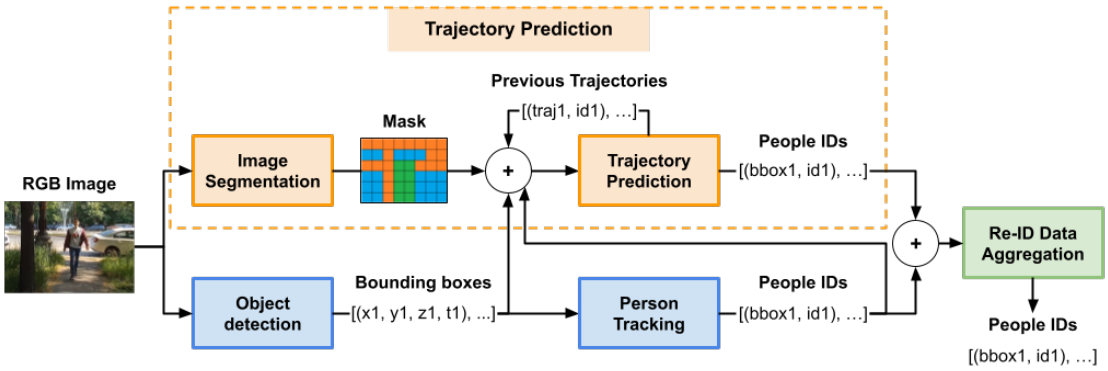

Figure 7.1: Real-time Person Re-identification and Tracking architecture based on a People Trajectory Prediction component.

The architecture for the real-time person re-identification and tracking task combines five vision modules to generate the final result, as follows:

- Object Detection - The object detection module is identifying the people in the images. For each input image the module generates bounding boxes representing the positions of the people. The re-identification system is assigning labels to each bounding box.
- Semantic Segmentation - The semantic segmentation module is part of the people trajectory prediction module responsible with extracting environmental information from the images. The network is run in parallel with the object detection to further reduce the computation time for real-time scenarios.
- Trajectory Prediction - The trajectory prediction module is used to better estimate the movement of the people. It helps the system to assign more precisely the right labels to the generated bounding boxes. By considering the estimated future trajectories
- Person Tracking - The person tracking module is used by the trajectory prediction system to gather the initial observed data before generating a trajectory. The system must be able to distinguish between the existing detections such that the sequence of positions in an observed trajectory belongs to a single person. It uses the system introduced in [25] and [26].
- Re-ID Data Aggregation - The data aggregation module is connecting the information from the person tracking module and from the trajectory prediction module to generate the final identification numbers.

7.2. Evaluation and Validation

The people re-identification and tracking system was tested taking into consideration social robotics scenarios. We performed an evaluation on an existing tracking dataset and an analysis on self-acquired images. Our evaluation compares a standard tracking technique for real-time people tracking [25] with our proposed trajectory prediction-based re-identification system. The evaluation demonstrates the improvement of the performance of our proposed technique. Table 7.1 reports the tracking values obtained on the MOT17 dataset. The proposed system obtains a higher value for the MOTA metric and similar values for IDF and HOTA. These values were obtained by using a trajectory prediction model trained on the JRDB dataset.

Metric	**Standard Tracking**	**Proposed System**
MOTA [38] (%)	55.05	61.04
IDF1 [39] (%)	55.62	52.24
HOTA [40] (%)	45.57	43.79

Table 7.1: Tracking metrics values for the trajectory-based re-identification system on MOT17 dataset.

We performed an analysis of the method on particular scenarios on self-acquired data, in both indoor and outdoor scenes, with one or multiple people in the images. The scenarios vary in terms of environmental conditions, such as background, light, distance or obstacles that may cause full or partial occlusions, but also in terms of movement of the people, such as interactions, occlusions by movement, sudden change in the movement of the people. Our method is able to deal with more complex situations and predict the correct identifiers even after long full occlusions, while the DeepSort method exhibits problems.

The system can integrate dynamic environments through the visual information about the scene, as the segmentation mask is computed for every input image. The continuously processed information also helps in the cases of moving cameras, as the system adjusts the trajectories based more on the recent data. The scaling problem is handled by the 2D coordinates representation of the people. This approach, however, can present some problems with unstable people positions, considering the situations when the bounding boxes are not consistent and significantly vary in terms of detected area.

Chapter 8

Systems Validation in Other Contexts

The context of autonomous robotics is a very popular direction in the research area considering the variety of possible applications and the growing interest of the people. However, the focus is set not only on social robots, but on all existing autonomous devices, such as autonomous cars or self-operating drones.

The principles we applied into the systems for people trajectory prediction and people re-identification and tracking can be applied to other different scenarios. Besides social robotics, we tested the methods on two other contexts, namely self-driving cars and flying drones. Both contexts assume movement of the camera, appearance variation and occlusions, with the main difference being the angle of the camera. The self-driving car applications analyse images from an eye-level point of view, as in the case of social robotics, while drones process images from a bird-eye point of view.

8.1. Validation in the Context of Self-Driving Cars

The context of autonomous driving implies detecting pedestrians on the side of the road and estimating their future positions to prevent possible collisions. The most important differences between the social robotics context and the autonomous driving one are that the speed of an autonomous car is much higher than in the case of a robot and the size of the bounding boxes representing pedestrians are smaller than the ones in socially assistive robotics applications.

For an objective evaluation of the people trajectory prediction component for self-driving cars we integrated the Caltech-Pedestrians [32] dataset, as it includes videos acquired from a moving car in an urban environment. We computed the ADE (average displacement error) and FDE (final displacement error) metrics for each subset, obtaining the average values of 7.00 pixels for the ADE and 12.85

pixels for the FDE. The values obtained are small, proving that the functioning of the system is robust and reliable in autonomous driving applications.

Re-identification in the context of autonomous driving is a more challenging problem than in the context of social robotics. For a good performance the system requires a model that can encode camera movement, scale variation and multiple occlusions. The movement of the camera is an important constraint in the problem of re-identification, as it implies not only changes in the angle views, but also fluctuations in the speed of movement. We analysed several particular scenarios using the self-acquired data, as well as compared the reported values of the standard tracking metrics on Caltech-Pedestrians dataset, to prove the improvement of our system when compared with the standard one. Table 8.1 demonstrates a significant improvement for our system when compared with DeepSort on the standard tracking metrics.

Metric	**Standard Tracking**	**Proposed System**
MOTA [38] (%)	89.22	94.92
IDF1 [39] (%)	78.36	90.10
HOTA [40] (%)	78.96	89.15

Table 8.1: Tracking metrics values for the trajectory-based re-identification system on Caltech-Pedestrians [32] dataset.

8.2. Validation in the Context of Top View Images Applications

The previously presented scenarios assume that images are acquired from an eye-level point of view. Both a social robot and an autonomous vehicle would have the cameras placed on the platform such that the acquired images are seen in perspective. Applications such as drones or surveillance systems use images acquired from a top-level point of view.

To obtain an evaluation of the performance of the systems in conditions where the images have a top view of the scene, we integrated the Stanford Drone dataset [33], which contains a collection of outdoor images acquired from high altitude. We computed the ADE (average displacement error) and FDE (final displacement error), as in the evaluation phases of the other two cases, for each subset. The

reported errors are relatively small, with average values of 13.26 pixels for the ADE and 22.02 pixels for the FDE. The average errors validate that the system can be integrated in an application using bird-eye view images.

The people re-identification and tracking system was tested using self-acquired data from a high point of view. Our experiments proved that the system performs satisfactory for a people tracking application, even though it presents more identity switches when compared with the other two contexts. The results obtained in this scenario are strongly influenced by the semantic segmentation network integrated in the system, as it needs to be trained to segment images with a top view, to fit the requirements of the context of self-operating drones.

These validation scenarios alongside the social robotics one prove that the system can be utilised in a variety of applications. Moreover, the property of being suitable even for bird's-eye view scenarios is important for applications where there are multiple streams of images coming from different angles, which can combine data and obtain a more stable system.

Chapter 9

Conclusions

The thesis presents an approach for a real-time people re-identification and tracking system for autonomous applications based on a people trajectory prediction component. It validates the approach on three different autonomous contexts, namely social robots, self-driving cars and autonomous drones. In addition, it introduces a framework for robotic applications which integrates the people tracking system and the people trajectory prediction component with other robotic capabilities to create a fully autonomous social robotic design for various user interaction scenarios.

The tracking and re-identification system has an essential role in contexts which assume performing in environments that include movement of people or even interactions with them. The autonomous devices require the information about the behaviour of the individuals to prevent unsafe situations. The trajectory prediction component estimates future movement based on the observed data to further improve the stability and safety of the autonomous devices. The data coming from such systems combined with the information extracted by other methods provides a sufficient amount of knowledge to perform as required. The evaluation phase performed on the two systems proves their applicability on various situations and scenarios. The systems were tested on self-acquired data and on relevant datasets to have both a quantitative and qualitative assessment. The systems confirmed their suitability for real-time applications under operating conditions produced by autonomous devices.

The robotic framework combines the above-mentioned information with capabilities generated based on general robotic sensors, to develop a platform suitable for general social applications using robots. It incorporates functionalities for robots in terms of visual understanding, navigation, voice interaction and planning. The framework provides the possibility to combine different functions from each capability to create behaviours suitable for a robotic scenario. The framework was tested under laboratory conditions using an

existing social robot. The evaluation phase includes multiple scenarios, which the robot must be able to execute correctly, regardless of the interacting user. The scenarios tested different situations, by implying both a single user and multiple users, with simple behaviour thread and complex pre-emptive behaviours. We added several videos[1] to offer an example of the performance of systems presented in the thesis.

9.1. Contributions

The thesis introduces several new techniques related to visual information extraction for autonomous applications. The original contributions of the work presented in this thesis are as follows:

- A framework suitable for general social robots which can be applied for socially assistive applications. We combine data acquired from multiple sensors to create a homogeneous platform that can provide sophisticated actions grouped into complex behaviours. The framework offers capabilities for navigation, based on robot lasers, for visual understanding, based on RGB and depth cameras, for vocal interaction, based on microphones and speakers. The required sensors are basic for any social robotic device. Some parts of the framework were developed in collaboration with Mihai Nan, Alex Awada and Alexandru Sorici.

- An architecture for planning the sequencing of the tasks. The architecture allows on-line dynamic changes when new commands are triggered by various sources: user voice commands, user visual interface and automatic external platform. The tasks are pre-empted and interchanged based on their associated priorities. Each command received by the framework starts a specific defined behaviour which consists of a composition of tasks. The robotic device executes one task at a time, allowing the possibility to pause and resume behaviours based on their importance.

- An architecture for a real-time people trajectory prediction system. The system incorporates three pieces of information to generate the final predictions: individual movement, environmental information and social influence. The individual movement is collected based on the observed

[1] `https://ctipub-my.sharepoint.com/:f:/g/personal/stefania_a_ghita_upb_ro/EoCy0oZ--RZOre5ktqPXioAB2DjgsNExNk2kn0mAvQWzgw?e=fM4cEa`.

data, representing the positions where the person was located before the prediction, the environmental information is generated using a semantic segmentation network which processes corresponding RGB images, and the social influence is integrated by a pooling layer which exchanges information between neighbouring people.

- An analysis on the advantages of integrating a people trajectory prediction system into a robotic framework. The analysis includes experiments performed using a social robotic device. The method is integrated into the robotic framework as a new component which exposes additional capabilities for the robot. The defined experimental scenarios prove the improvement of the performance of a socially assistive robot in various critical situations.

- A method for changing the visualization angle of an image from an eye-level view to a bird-eye view. The method applies a mathematical perspective transformation of a general image using parameters generated based on a semantic segmentation mask. The semantic segmentation mask is processed to extract the vanishing point of the image by determining the ground label. Based on the vanishing point, the method computes the parameters required by the homography matrix which is used to apply the perspective transformation.

- An architecture for a real-time people tracking and re-identification system. The architecture combines a simple tracking technique with a people trajectory prediction system to re-identify the same people through a sequence of images. The method is designed for streams of images acquired by an autonomous device, as it can encode camera movement, changes in angle views and speed variations. It assigns identification numbers to each person detected in an image in a short amount of time, making it suitable for autonomous applications. The evaluation phase proved the validity of the method in three different contexts: assistive social robotics, autonomous driving and autonomous drones.

- A self-acquired data collection used to validate the people tracking and re-identification system. The data collection includes videos acquired in indoor and outdoor environments from an eye-level point of view. The data contains streams of RGB images acquired in two different scenarios: from the point of view of a robotic platform and from a moving car. The videos include

occlusions, changes of the camera angle and variation of the speed movement. The corresponding bounding boxes for each image were computed using an external people detection system.

9.2. Perspectives

The goal of this thesis is to design a reliable system for the task of people tracking in autonomous contexts and to integrate it in a framework developed for socially assistive robots. The newly introduced systems can be improved considering near-term objectives and long-term development.

The robotic framework involves multiple aspects when considering future work in terms of individual functionality modules. The set of functionality modules can be extended with the ability to perceive emotions from RGB image data. Such a feature can enhance the general behaviour of a social robot, making it more conscientious of what the user needs. The voice command module can be extended to include additional languages and a richer set of commands. This improvement would allow a more ample testing in deployment scenarios involving end-users from other countries. Furthermore, the 3D coordinates estimation module can be furthered improved by replacing the information coming from a depth sensor with a more reliable data source. As the values of all components of the 3D coordinate are computed based on the recorded depth, eliminating the errors in the depth data is expected to result in more accurate estimations. In addition, the behaviour management module can be extended to interweave predefined action sequences with planning results, based on the ROSPlan framework, thereby enabling a flexible and more extendable behaviour composition functionality.

The trajectory prediction module can be further improved by integrating additional information into the system. A possible development would imply including the influence coming from all the moving objects in the images that could alter the movement of a person. The movement of cars or animals can significantly influence the trajectory of a person by causing sudden changes or breaks. Furthermore, the social influence can be better distributed between participants. In the current form of the architecture the people in the images share the information about their trajectories regardless of their destination or orientation. A possible improvement consists in analysing the destinations of the people and distributing information only between people that are or will be in

proximity. Moreover, we can boost the environmental information extracted by the system by including probability maps that represent possible destination goals for the people in the scene. Considering this system, one of our objectives is to generate even more information for an autonomous system, such as collision detection and groups trajectory analysis, based on the results of the trajectory prediction module.

The people tracking and re-identification module is an essential component which need to be further improved for even more reliable re-identifications. One main improvement the system can experience is the replacement of the simple tracking technique which generates the first three observed positions of a person. By replacing this method with a more accurate one the general performance of the system would be greatly improved. Another possible development consists in integrating visual features extracted by a neural network to better differentiate the cases where the predicted future positions of different people are in proximity. In the current architecture, if the trajectories of two people collide and only one person is visible in the image, the system assigns the visible person the identification number of the person that was previously considered to be closer to the camera, more specifically with a larger bounding box. By integrating visual cues about the general appearance of the people, such problems would be solved with a higher accuracy. Furthermore, we can improve the system by combining the information extracted by the people trajectory prediction method with the detected postures of the people in the images, to validate the trajectories with possible activities. This development has two directions: one in which we validate the trajectories based on recognized actions and one in which we integrate the postures directly into the trajectory prediction system for more informed results.

References

[1] D. Feil-Seifer and M. J. Mataric, "Defining socially assistive robotics," in *Rehabilitation Robotics, 2005. ICORR 2005. 9th International Conference on*, pp. 465–468, IEEE, 2005.

[2] H. Nap, S. Suijkerbuijk, D. Lukkien, S. Casaccia, R. Bevilacqua, G. Revel, L. Rossi, and L. Scalise, "A social robot to support integrated person centered care," *International Journal of Integrated Care*, vol. 18, no. s2, 2018.

[3] C. Antonopoulos, G. Keramidas, N. S. Voros, M. Hübner, D. Goehringer, M. Dagioglou, T. Giannakopoulos, S. Konstantopoulos, and V. Karkaletsis, "Robots in assisted living environments as an unobtrusive, efficient, reliable and modular solution for independent ageing: The radio perspective," in *International Symposium on Applied Reconfigurable Computing*, pp. 519–530, Springer, 2015.

[4] D. Fischinger, P. Einramhof, K. Papoutsakis, W. Wohlkinger, P. Mayer, P. Panek, S. Hofmann, T. Koertner, A. Weiss, A. Argyros, *et al.*, "Hobbit, a care robot supporting independent living at home: First prototype and lessons learned," *Robotics and Autonomous Systems*, vol. 75, pp. 60–78, 2016.

[5] S. Coşar, M. Fernandez-Carmona, R. Agrigoroaie, J. Pages, F. Ferland, F. Zhao, S. Yue, N. Bellotto, and A. Tapus, "Enrichme: Perception and interaction of an assistive robot for the elderly at home," *International Journal of Social Robotics*, pp. 1–27, 2020.

[6] D. Portugal, P. Alvito, E. Christodoulou, G. Samaras, and J. Dias, "A study on the deployment of a service robot in an elderly care center," *International Journal of Social Robotics*, vol. 11, no. 2, pp. 317–341, 2019.

[7] N. Hawes, C. Burbridge, F. Jovan, L. Kunze, B. Lacerda, L. Mudrova, J. Young, J. Wyatt, D. Hebesberger, T. Kortner, *et al.*, "The strands project: Long-term autonomy in everyday environments," *IEEE Robotics & Automation Magazine*, vol. 24, no. 3, pp. 146–156, 2017.

[8] G. Wilson, C. Pereyda, N. Raghunath, G. de la Cruz, S. Goel, S. Nesaei, B. Minor, M. Schmitter-Edgecombe, M. E. Taylor, and D. J. Cook, "Robot-enabled support of daily activities in smart home environments," *Cognitive Systems Research*, vol. 54, pp. 258–272, 2019.

[9] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–971, June 2016.

[10] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2255–2264, June 2018.

[11] Y. Xu, Z. Piao, and S. Gao, "Encoding crowd interaction with deep neural network for pedestrian trajectory prediction," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5275–5284, June 2018.

[12] J. Amirian, J. Hayet, and J. Pettré, "Social ways: Learning multi-modal distributions of pedestrian trajectories with gans," *CoRR*, vol. abs/1904.09507, 2019.

[13] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, and S. Savarese, "Sophie: An attentive GAN for predicting paths compliant to social and physical constraints," *CoRR*, vol. abs/1806.01482, June 2018.

[14] D. Ridel, N. Deo, D. Wolf, and M. Trivedi, "Scene compliant trajectory forecast with agent-centric spatio-temporal grids," *IEEE Robotics and Automation Letters*, vol. PP, pp. 1–1, 02 2020.

[15] J. Liang, L. Jiang, J. C. Niebles, A. G. Hauptmann, and L. Fei-Fei, "Peeking into the future: Predicting future person activities and locations in videos," *CoRR*, vol. abs/1902.03748, 2019.

[16] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. S. Torr, and M. K. Chandraker, "DESIRE: distant future prediction in dynamic scenes with interacting agents," *CoRR*, vol. abs/1704.04394, 2017.

[17] P. Dendorfer, A. Osep, and L. Leal-Taixé, "Goal-gan: Multimodal trajectory prediction based on goal position estimation," *CoRR*, vol. abs/2010.01114, 2020.

[18] J. Gu, C. Sun, and H. Zhao, "Densetnt: End-to-end trajectory prediction from dense goal sets," *CoRR*, vol. abs/2108.09640, 2021.

[19] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, pp. 1–45, 01 2006.

[20] T. Huang and S. Russell, "Object identification in a bayesian context," in *IJCAI*, vol. 97, pp. 1276–1282, Citeseer, 1997.

[21] J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, and Q. Tian, "Spatial-temporal graph convolutional network for video-based person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3289–3299, 2020.

[22] P. Voigtlaender, J. Luiten, P. H. Torr, and B. Leibe, "Siam r-cnn: Visual tracking by re-detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6578–6588, 2020.

[23] Y. Zhong, X. Wang, and S. Zhang, "Robust partial matching for person search in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6827–6835, 2020.

[24] S. Gao, J. Wang, H. Lu, and Z. Liu, "Pose-guided visible part matching for occluded person reid," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11741–11749, 2020.

[25] 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.

[26] N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 748–756, IEEE, 2018.

[27] P. Bergmann, T. Meinhardt, and L. Leal-Taixé, "Tracking without bells and whistles," *CoRR*, vol. abs/1903.05625, 2019.

[28] A. Milan, L. Leal-Taixé, I. Reid, and S. Roth, "Mot16: A benchmark for multi-object tracking," *CoRR*, vol. abs/1603.00831, March 2016.

[29] R. Martín-Martín, H. Rezatofighi, A. Shenoi, M. Patel, J. Gwak, N. Dass, A. Federman, P. Goebel, and S. Savarese, "Jrdb: A dataset and benchmark

for visual perception for navigation in human environments," *CoRR*, vol. abs/1910.11792, Oct. 2019.

[30] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 261–268, Sep. 2009.

[31] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," *Computer graphics forum*, vol. 26, pp. 655–664, Sep. 2007.

[32] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 304–311, IEEE, 2009.

[33] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *European conference on computer vision*, pp. 549–565, Springer, 2016.

[34] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018.

[35] "Pepper Robot." https://www.softbankrobotics.com/emea/en/pepper. [Online; accessed 11 July 2022].

[36] "Robot Operating System." https://www.ros.org/. [Online; accessed 11 July 2022].

[37] A. Ş. Ghiță, A. F. Gavril, M. Nan, B. Hoteit, I. A. Awada, A. Sorici, I. G. Mocanu, and A. M. Florea, "The amiro social robotics framework: deployment and evaluation on the pepper robot," *Sensors*, vol. 20, no. 24, p. 7271, 2020.

[38] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–10, 2008.

[39] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European conference on computer vision*, pp. 17–35, Springer, 2016.

[40] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe, "Hota: A higher order metric for evaluating multi-object tracking," *International Journal of Computer Vision*, pp. 1–31, 2020.