TÉCNICO LISBOA

People Recognition and Identification in Service Robots

Vicente Palma Figueira Castro Pinto

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisor: Prof. Rodrigo Martins de Matos Ventura

Examination Committee

Chairperson: Prof. João Fernando Cardoso Silva Sequeira
Supervisor: Prof. Rodrigo Martins de Matos Ventura
Member of the Committee: Dr. Plinio Moreno Lopez

November 2021

DECLARATION

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

I would like to thank Joana for all the love and care throughout these years that has helped me through the stressfull times and made my life beautiful. I would like to thank my parents for all the support and good advices, for investing in my education and for always encouraging my critical thinking. I would like to thank my two sisters for always making me see the world in a different perspective, Maria for always being someone I look up to and Sofia for always making me want to be someone she looks up to. I would like to thank my grandparents for showing me the value of the simple things in life. I also must thank Dee and Dum, my two little companions who were always with me during the writing of this thesis, giving me the occasional excuse to take a break.

I would like to thank João Ramalho, Ruben Sirgado, João Sousa and the rest of my friends and colleagues for all the unforgettable moments of joy, companionship and music. I would like to thank Martim Pereira and Tiago Jacinto for the technical help. I would also like to thank my colleagues from the SocRob@Home team, that welcomed me and helped me understand the challenges of mobile robotics.

I would also like to acknowledge my dissertation supervisors Prof. Rodrigo Ventura and Rui Bettencourt for their insight, support and sharing of knowledge that was fundamental for this thesis.

Last but not least, to all the people that I have crossed paths with so far, which in one way or another have shaped who I am and how I view the world. Thank you.

To each and every one of you – Thank you.

Abstract

Service robots provide services to humans such as helping in domestic chores or serve as companion to elderly people. To accomplish a good social behaviour, the robot should be able to recognize and differentiate people in the scene, since this skill enables personalized human-robot interaction. People re-identification in service robots is key for their acceptance in people's homes, as well as for performing a wide variety of tasks. People re-identification and tracking are two closely related tasks. Existing Re-ID based tracking methods designed for mobile robots have some limitations since they either assume constrained conditions on the environment and the movement of people or they are not robust enough in challenging conditions such as the presence of obstacles or similar targets. This thesis proposes a Re-ID based multi-people tracker suitable for mobile robots. It combines existing methods such as: a people detector, a people localizer, a Re-ID feature extractor and a Kalman filter framework with simple data association and track management approaches. A novel RGB-D Re-ID multi-people 3D tracking dataset recorded with a moving camera in an environment with obstacles and target's occlusions and appearance changes is presented. Experimental evaluation shows that the method achieves very good tracking and re-identification performance on the proposed dataset, at a high frame-rate, and that it outperforms another state-of-the-art method on an open-space dataset. The proposed system is lightweight, robust and suitable for real-world applications, allowing for an improvement of human-robot interaction.

Keywords

Human-robot interaction; People re-identification; People tracking; Multiple Kalman-filter; RGB-D dataset.

Resumo

Os robôs de serviço providenciam serviços tais como ajuda nas tarefas domésticas ou companhia para idosos. O robô deve ser capaz de diferenciar as pessoas que o rodeiam e ter uma interação humano-robô personalizada. A re-identificação de pessoas é crucial para a sua aceitação em ambientes domésticos. A re-identificação e o tracking de pessoas são duas tarefas que se relacionam intimamente. Os métodos de re-identificação e *tracking* existentes desenvolvidos para robôs móveis têm limitações já que, ou assumem condições restritivas do espaço e do movimento das pessoas, ou não são robustos relativamente a situações complexas, tais como a presença de obstáculos ou pessoas com aparência semelhante. Esta tese propõe um *tracker* 3D de múltiplas pessoas baseado em re-identificação, adequado para robôs móveis. O sistema combina métodos existentes tais como um detetor e um localizador de pessoas, um extrator de caraterísticas de re-identificação e uma estrutura de *Kalman Filters* com estratégias simples de associação de dados e gestão de trajetórias. É apresentado um conjunto de dados RGB-D de re-identificação e *tracking* de múltiplas pessoas, gravado com uma câmara móvel num ambiente com obstáculos, oclusões e mudança de aparência nas pessoas presentes. A avaliação experimental mostra que o método tem um bom desempenho de re-identicação e *tracking*, a um *frame-rate* alto, no dataset proposto e que tem melhor performance que outro método do estado da arte, num conjunto de dados em espaço aberto. O método proposto é computacionalmente leve, robusto e aplicável em situações reais, proporcionando uma melhoria da interação humano-robô.

Palavras Chave

Interação humano-robô; Re-identificação de pessoas; *Tracking* de pessoas; Múltiplos *Kalman-filter*; Conjunto de dados RGB-D.

Contents

List of Figures

List of Tables

List of Algorithms

Acronyms

CNN	Convolutional Neural Network
CV	Computer Vision
DPM	Deformable Part Models
EKF	Extended Kalman Filter
ELF	Ensemble of Localized Features
FDA	Fisher Discriminant Analysis
GPU	Graphics Processing Unit
HA-CNN	Harmonious Attention Convolutional Neural Network
HOG	Histogram of Oriented Gradients
HSV	Hue Saturation Value
KTP	Kinetic Tracking Precision
LBP	Local Binary Pattern
LDA	Linear Discriminative Analysis
LOMO	Local Maximal Occurrence
LRF	Laser Range Finder
LSTM	Long Short-Term Memory
MBOT	MOnarCH Robot
MOTA	Multiple Object Tracking Accuracy
MOT	Muliple Object Tracking

MOTP Multiple Object Tracking Precision

PCA Principal Component Analysis

Re-ID Re-Identification

R-FCN Region-based Fully Convolutional Network

RFID Radio Frequency Identification

RGB Red-Green-Blue

RGB-D Red-Green-Blue-Depth

ROS Robot Operating System

SCNCD Salient Color Names Based Color Descriptor

SDALF Symmetry-Driven Accumulation of Local Features

SIFT Scale Invariant Feature Transform

SVM State Vector Machine

UKF Unscented Kalman Filter

XQDA Cross-view Quadratic Discriminant Analysis

1

Introduction

Contents

1.1 Motivation

Service robots have received increased attention in recent years, covering a great variety of applications and system designs. Service robots are robots that are fully or partially autonomous, that provide services to humans or equipment [3]. They differ from industrial robots in their applications and are usually mobile or manipulative robots. The range of services they can provide are vast: refuelling, cleaning, maintenance, surveillance, office automation, firefighting, serve as a restaurant waiter or an hotel receptionist, entertainment and much more. These robots can be extremely helpful since they can replace humans in hazardous situations, allow humans to save time for other activities, help physically handicapped people and serve as a companion to elderly people or children.

To provide these services, the robot needs to perform complex tasks such as navigation, sensor-fusion, manipulation and perception of the environment. Besides that, some of these robots are in constant interaction with humans, which requires additional skills and functionalities to provide a natural and efficient human-robot interaction, including speech recognition, people identification, ability to reproduce emotions and communication. Their ability to interact naturally with humans and be social, while requiring minimal intervention from the user, is critical for their usefulness [4].

A sub-group of service robots are the domestic robots. These are robots designed to help humans at home in their daily domestic chores. Their tasks can range from simple ones such as vacuuming or cleaning to more challenging ones such as providing care for elderly at home. For these specific type of robots, human-robot interaction is very important and is determinant for their acceptance in people's homes. They should be able to communicate and understand humans. Besides that, their appearance, movements and interface influences how people perceive and react to them [5].

The MOnarCH robot (MBOT) is a service robot originally designed to interact with children in hospitals. The MOnarCH Robot (MBOT) was adapted for robotic competitions in domestic scenarios by SocRob@Home [6] and is capable of navigating autonomously, understand spoken commands, detecting and manipulating objects, tracking and following people. Considering its application in domestic environments, the human-robot interaction is also a very important aspect to consider in the system's design.

In order to accomplish a good social behaviour, the robot should be able to recognize and identify humans. Besides that, the robot should be able to re-identify individuals, that is, determine if a certain person is present in a set of candidates and recall that person's identity through time. The ability to differentiate different people and re-identify a certain person in different points in time is also very useful for a better interaction between the robot and humans. This skill enables personalized interaction between the robot and the people in his surroundings. For instance, the robot can address people by their personal names and recall personal preferences and details, as well as communication patterns that allow for further improvement of the human-robot interaction. All of these build up the robot's personality

and adaptability, which are key factors for increasing trust in the robot [7].

Re-Identification (Re-ID) of people also improves people tracking and following, in cases where there are occlusions or where the target is lost. This improves the perception that the robot has of the people in the scene, allowing for the execution of more personalized tasks such as delivering objects requested by a specific person, count the number of different people present, keep track of the movements and behaviours of a specific person and much more.

Although people re-identification is a critical skill for improving human-robot interaction and people perception and thus increasing people's acceptance and trust in service robots, it has not been deeply investigated in this context and there is still need for a robust and practical approach.

1.2 Research Goals

The main goal of this thesis is the development of a Re-identification based system to be deployed in a mobile robot that robustly recognizes, identifies and tracks the different people present in the robot's surroundings, assigning to each one of them a unique ID, while guaranteeing a real-time performance. Since practical applicability is one of the main goals, the evaluation of the method in real-case scenarios is key. Considering this, the following research goals can be identified:

- 3D position tracking of multiple people in an environment with obstacles and occlusions caused by other people
- Recognition and identification of the different people seen by the robot during the system's execution
- Re-identification of a target that was previously seen, by assigning it the same ID as before
- Real-time performance and practical applicability of the method in a mobile robot

1.3 Contributions

The scientific contributions of this thesis are three-fold:

- Integration of existing modules and methods (people detector, people localizer, Re-ID feature extractor and Kalman filter) in the development of a novel Re-ID 3D multi-people tracker
- Construction of a RGB-D Multi-people tracking and Re-Identification dataset recorded using a moving camera, including people 3D position ground-truth in an indoor and occluded scenario, representative of a domestic environment
- An experimental evaluation of the method proposed in a real-word dataset

1.4 Dissertation Outline

The remaining of this document is organized as follows:

- Chapter 2 provides an overview of the background on people Re-ID and its application on service robots. First, the main concepts and components of a person re-identification system are presented, with several examples of state-of-the-art methods, as well as benchmark datasets. Secondly, the main components of person re-identification systems applied to service robots and existing methods are presented.
- Chapter 3 presents the proposed Re-id based multi-people tracker. It gives an overview of the system architecture and the coordinate frames that are considered in this work. The different modules that compose the system are then detailed. It also presents the methodology of the multi-people tracker, explaining the main stages: track and state estimation, data association and track management.
- Chapter 4 first presents the state-of-the-art on Multi-object tracking datasets and then presents the newly created Re-ID Multi-people tracking dataset.
- Chapter 5 details the experiments that were conducted and then shows the evaluation results of the proposed system on the proposed dataset and on other multi-target tracking benchmarks.
- Chapter 6 presents the conclusions that can be drawn from this work and discusses further improvements and future research topics.

2

Background

Contents

2.1 Computer Vision Person Re-Identification

In the context of Computer Vision (CV) and pattern recognition, people re-identification is the task of retrieving the occurrences of a certain person (probe) from a set of person candidates (gallery) [8]. This task is mostly useful for surveillance systems and is very challenging because a person's appearance varies a lot with illumination, pose and viewpoint changes, obstructions and resolution. This factors can lead to cases where the difference between the same person's appearance in two images is larger than the difference between the appearance of two different people.

The gallery and probe are represented by bounding boxes that enclose the person. The appearance information of the probe and the gallery candidates is extracted from the bounding boxes and is represented by a feature descriptor. Feature descriptors are then compared using a similarity function, which measures how similar two instances are.

Re-ID can be classified into closed-world and open-world. In this context, the world is the environment in which the system is operating in, including all the people seen by the camera. Closed-world Re-ID is based on the assumption that the probe is present in the gallery and consists in a matching task. The goal is to find the pair of images for which the appearance of the probe and the gallery candidate is more similar. Open-world Re-ID is a more general case, where there is no guarantee that the probe is present in the gallery. This case is more representative of real-life and is usually the context of practical applications. In the open-world scenario, Re-ID implies a verification task.

2.1.1 Re-ID methods

Hand-crafted Features

Some Person Re-ID methods make use of different histograms and segmentation techniques to construct the appearance descriptor for each person. That descriptor is then compared using a distance metric to identify the same person across different frames or images. These methods are usually combined with metric learning algorithms, which will be covered in the next section.

One way of constructing a feature descriptor is using Ensemble of Localized Features (ELF) [9]. This approach allows for simple color and texture features to be combined into a single similarity function.

Another approach consists in extracting Symmetry-Driven Accumulation of Local Features (SDALF) [10]. These features represent three distinct aspects of the human appearance and are extracted by computing Hue Saturation Value (HSV) histograms, Maximally Stable Colour Regions and Recurrent Highly Structured Patches. This method achieves robustness against very low resolution, occlusions and pose, viewpoint and illumination changes.

Yang et al. presented a feature representation based on Salient Color Names Based Color Descriptor (SCNCD) [11] which allowed for feature computation to be done very fast, if SCNCD of each color was

computed in advance. There is also another feature representation called Local Maximal Occurrence (LOMO) [12] that describes the horizontal occurrence of local features, which is stable against viewpoint changes.

Another approach is to use Local Binary Pattern (LBP) [13], which is a simple texture operator that assigns each pixel of an image a binary number, which is the result of thresholding the neighbourhood of that pixel. A feature descriptor composed of LBP is computationally very cheap and is very efficient.

Matsukawa et al. presented a descriptor based on an hierarchical distribution of pixel features [14]. A local region of an image is described using an hierarchical Gaussian distribution which includes mean and covariance. Each region is then described by a set of gaussian distributions that represent the appearance of a local patch, including color and texture information. The parameters of the Gaussian distributions are also described by a Gaussian distribution.

Hand-crafted features are a fast and simple way of computing person feature descriptors, although their discriminative power can be limited, which makes the performance of the methods very dependent on the robustness of matching techniques. Matching images based on this type of features using standard metrics such as the Euclidean distance leads to poor performance due to large variations in pose and illumination.

Metric Learning

Metric Learning is a field of Machine Learning with the objective of learning distances from the data, improving similarity-based methods [15]. In Person Re-ID methods, metric learning can be used to learn an appropriate distance metric to compare feature descriptors such as the ones presented in the previous section. Metric learning can improve the matching performance, by grouping data points that belong to the same person together, while pushing away data points belonging to different people.

One of the earliest works that applied distance learning to the task of person Re-ID was the Probabilistic Relative Distance Comparison model [16], that aimed at maximizing the matching accuracy regardless of the feature representation method. The main novelty presented was that the goal is to maximize the probability of a pair of a true match having a smaller distance than that of a wrong match, instead of trying to minimize intra-class variation while trying to maximize inter-class variation.

To tackle the issue of having images from different cameras and how the transition of one camera to another can impact the distance metric learning, a relaxed pair-wise learned metric was proposed by Hirzer et al. [17] that learns a metric from pairs of samples from different cameras, allowing less-sophisticated features to be used while maintaining matching performance.

Another popular approach is kernel-based metric learning, which allows for a dimensionality reduction of the data being compared. This approach has been applied to person Re-ID and several variations have been implemented and evaluated such as regularized Pairwise Constrained Component Analysis, kernel Fisher Discriminant Analysis (FDA), Marginal Fisher Analysis and a ranking ensemble voting

scheme, which have shown improvements in performance [18].

Along with the presentation of LOMO features, a subspace and metric learning approach called Cross-view Quadratic Discriminant Analysis (XQDA) was proposed [12]. A discriminating metric is learned by learning a discriminant low dimensional subspace by cross-view quadratic discriminant analysis.

Metric learning assumes great importance in Re-ID methods, especially when using hand-crafted features, since it improves significantly the re-identification performance, allowing for the use of simple feature representations that are computationally lightweight.

Deep Networks and Attention Networks

With the development of deep learning in the recent years, several deep Re-ID methods have been gaining relevance and achieving the best performance on the most challenging datasets [19]. These methods consist in deep-network architectures that focus on feature representation and metric learning together. Each architecture differs in the way features are computed, the distance metric learned and the way both are combined.

The first deep network approaches for the Re-ID task introduced a new concept of jointly learning the color feature, texture feature and metric, all in the same framework [20] and a better handling of misalignment, photometric and geometric transforms, occlusions and background clutter.

Deep networks contributed to overcome several difficulties and obstacles in Re-ID. One of those problems is that hand-crafted features are usually not discriminative or robust enough. To tackle that, Wu et al. proposed an hybrid deep architecture with Fisher vectors and multiple supervised layers [21]. The network was trained with an Linear Discriminative Analysis (LDA) criterion that approximates inter and intra-class variations such that the deeply non-linear features become linearly separable. Patches from a person image are extracted and described by Principal Component Analysis (PCA)-projected Scale Invariant Feature Transform (SIFT) descriptors.

While feature extraction is critical, a lot of methods don't focus enough in the similarity learning task, applying simply a cosine or an Euclidean distance to the feature vectors which is not very discriminative and leads to overfitting, causing the need for larger datasets, as the networks get deeper. A solution was proposed with the deep hybrid similarity learning method [22], it consists of a metric learning module, which includes a hybrid similarity function to measure person similarity, and a feature learning module which is a light convolutional network with three convolutional layers to extract features. The hybrid similarity function is realized by learning a group of weight coefficients to project the element-wise difference and multiplication of a Convolutional Neural Network (CNN) learning feature pair into a similarity score. It is much more discriminative and requires less parameters than the Mahalanobis distance.

Many of the deep network for the Re-ID task are built by fine-tuning already existing architectures, successfull on other tasks. This paradigm originates global representations that are efficient for Re-ID,

but can suffer from the misalignment inherent to human pose variations and person detection errors. While some methods overcome that problem by adding attention models, extra annotations or explicit alignment of body parts, good results can also be obtained using a simple deep architecture where the critical design choices are thoroughly examinated and an appropriate training strategy is selected [23]. For example, using triplet-loss and its variations for Re-ID shows very good results [24], either for models trained from scratch or for pre-trained ones. In his work, Hermans et al. proposed two network architectures, TriNet and LuNet, that were used to perform end-to-end deep metric learning.

To improve perfomance in the case of person detection errors, Han et al. joined person detection and person Re-ID in an end-to-end framework [25]. The detection is optimized under the supervision of the Re-ID loss, in order to produce more reliable bounding boxes.

Attention-networks have also been widely implemented recently, with many deep-network architectures including attention modules where attention cues are deduced to construct more discriminative feature representations. One of these architectures is the Harmonious Attention Convolutional Neural Network (HA-CNN) that aims to simultaneously learn hard region-level and soft pixel-level attention within arbitrary person bounding boxes along with re-id feature representations [26]. Another similar approach is an Hybrid-attention guided network that fuses high-level features with low-level features, which aims to enhance the representation capacity of the CNN models to discriminately learn the features [27]. Attention-networks have achieved state-of-the-art performance on the most challenging datasets, specifically the Multi-level-attention Embedding and Multi-layer-feature Fusion Model which is currently one of the best performing models for Re-ID [28]. It uses ResNet-50 pre-trained on ImageNet as a baseline. Multi-level-attention embedding (spatial-level and channel-level attention blocks) and mulit-layer-feature fusion model were developed to obtain richer and more representative features.

Unsupervised Re-ID methods have been showing promising results [29, 30] but they still require labeled data, their performance is limited by the scale of the dataset and they ignore relations between the source and the target dataset, since they are based in unsupervised domain adaptation. Recently, a framework for unsupervised Re-ID was proposed consisting of a multi-scale network (MN), a multi-label learning module (ML) and a self-paced clustering module (SC), using ResNet as a backbone [31]. The MN module extracts global and local multi-scale features. The ML module generates a multi-label vector for each image. The SC removes noisy samples by density-based clustering algorithm and assigns pseudo-labels for multi-class training.

Deep Re-ID models can achieve very high performance but their real-world application is still a challenging task, considering that they require large amounts of training data and that they are usually computationally expensive.

2.1.2 Types of data used for Re-ID

Most of Re-ID methods are based on RGB data since RGB images are the most frequently available data in Re-ID tasks. Nonetheless, there are other types of data that can also be used for Re-ID and that have advantages relative to using only RGB.

In comparison with RGB images, depth images vary less in cases of low illumination or poor color information. They can also be used to obtain information about the body structure of a person, which can be very helpful to re-identify a person that has changed clothes, for example.

One way of making use of those advantages is to re-identify people based on biometric features such as the body volume [32]. In this method, body segmentation is applied for feature extraction, by computing several geometric body distances like the height of the person, distance between shoulder points, face's length, head, upper torso and lower torso volume. The matching and classification phase is done using an State Vector Machine (SVM). Another method uses the full body point cloud for the re-identification task [33]. The point clouds are warped to a standard pose and a similarity score between them is computed for the matching.

Depth information can also serve as input for neural networks, allowing models to learn discriminative features of the body shape or motion dynamics of a person. Haque et al. presented an attention-based model effective in identifying people in the absence of RGB information [34]. A combination of recurrent and convolutional neural networks allow to identify small and discriminative regions indicative of a person identity.

While using depth information has shown interesting results, specially in cases of low lighting or in clothes-changing scenarios, its practical application to the Re-ID task still faces some challenges [8]: first, depth cameras are not suitable for outdoor environments because depth information decreases rapidly with an increase in the distance between the camera and the target; secondly, body structure information obtained with depth cameras can be indistinguishable or not discriminative enough, specially as the viewpoint varies. Using depth information only for Re-ID may not be optimal in most cases, but it can be very useful in combination with other data, such as in multi-modal matching, which will be covered in the following section.

There are other types of data that can be useful for the Re-ID task, if available. One of them is skeletal data, which is a representation of the human body as an articulated system composed of rigid sections and joints [35]. It can be obtained by using a 3D skeletal tracker [36], an RGB-Depth camera or other methods that extract body joints explicitly. This representation is robust to view-point and scale variations and it is relatively easy to generate. Hence, there are some Re-ID methods that use skeletal data in their approach, mainly for person segmentation, which allows the computation of local feature descriptors in each joint through hand-crafted techniques or deep models, followed by feature matching. These methods are more robust to illumination, viewpoint and pose changes. Examples of these methods are

the ones presented in [37] and [38].

There are other methods that use thermal data, which allows for the identification of people under low lighting conditions [39–41]. The biggest challenge for thermal Re-ID is the cross-modality person matching using both color images and thermal images. It is also common in video investigation applications to have access to eye witnesses natural speech or text statements about the targeted person. Text-to-image Re-ID is a field of investigation with the goal of matching text descriptions of a person to their corresponding images [42]. Considering the hardware present in the MBOT, these methods were not investigated further since thermal images are not available and text-to-image is out of context.

All the different types of information that can be used for person Re-ID can be combined, making the most of their different advantages. This is done through multi-modal matching. This approach requires fusing different features that represent different information. This fusion can be done at feature-level [43, 44], where feature vectors are concatenated and a distance metric is learned for that combined vector; at score-level [45, 46], where the feature vectors are scored independently and their matching scores are then fused.

2.1.3 Re-identification datasets

There are several datasets that are currently used for benchmarking in people Re-ID. These datasets are extremely useful to evaluate the performance of the different methods. Besides that, they are used to train some of the deep learning models described before.

They can be divided into image-based and video-based datasets, based on the type of the data. Image datasets are more common and are usually bigger, while video datasets can be used for methods that make use of other data besides RGB such as temporal attributes. Most of these datasets include challenging cases like viewpoint variations, illumination variations, detection errors, occlusions, background clutter and low-resolution images [47]. Some of the most popular image datasets are VIPer [9], GRID [48], CUHK01-03 [49], Market-1501 [50], DukeMTMC [51], Airport [47] and MSMT17 [52] and some of the most popular video datasets are PRID-2011 [53], MARS [54], Duke-Video [55], Duke-Tracklet [56], LPW [57]and LS-VID [58]. A comparison between image and video datasets is presented in Table 2.1 and Table 2.2, respectively. The datasets vary in number of different people (unique ID's), the number of bounding boxes of targets, the number of cameras and the way the labels were annotated, either manually or automatically using a person detector. We can see that image and video datasets have become larger as time passes, which is related to the growing need of bigger datasets for the deployment of deep learning methods. We can also see that recently the datasets have been including automatically generated bounding boxes, considering that the increasing size of the datasets makes it very time-consuming to annotate all of the images. Automatically-generated labels are also useful to get a better indication of how a method performs in a practical application since it usually receives as input

a detection from a person detector and these bounding boxes are not as perfectly aligned as manually anotated ones.

Table 2.1: Image Re-ID datasets

Dataset	Year	#ID	#BBs	#Cameras	Label
VIPer	2007	632	1264	2	hand
GRID	2009	250	1275	8	hand
CUHK01	2012	971	3884	2	hand
CUHK02	2013	1816	7264	10	hand
CUHK03	2014	1467	13164	2	auto/hand
Market-1501	2015	1501	32668	6	auto/hand
DukeMTMC	2017	1404	36411	8	auto/hand
Airport	2017	9651	39902	6	auto
MSMT17	2018	4101	126441	15	auto

Table 2.2: Video Re-ID datasets

Dataset	Year	#ID	#Tracks	#Cameras	Label
PRID-2011	2011	200	400	2	hand
MARS	2016	1261	20715	6	auto
Duke-Video	2018	1812	4832	8	auto
Duke-Tracklet	2018	1788	12647	8	auto
LPW	2018	2731	7694	4	auto
LS-VID	2019	3772	14943	15	auto

2.2 Person Re-Identification and Tracking on Service Robots

On the context of mobile robotics, people Re-ID can be extremely helpful. One of the most common tasks in mobile robotics is the tracking of multiple targets [59]. Besides tracking their positions, knowing their identities and being able to differentiate between different individuals is very valuable, as it allows for a personalized interaction between the robot and the people in its surroundings. Hence, people Re-ID methods are used to assign unique ID's to the targets being tracked.

Applying Re-ID in robotics differs from the computer vision approach described earlier, because its goal is to identify the same people in different points in time, while in CV the task is to identify the same people across cameras, usually with short differences in time. Therefore, the use of people Re-ID methods in mobile robotics is usually integrated in a pipeline that contains three modules: a person detector, a person Re-ID module and a tracker [60]. As described before, person Re-ID is applied in regions of an image or video that represent a person, eg. a bounding box that enclosures the target. Hence, a person detector is required to generate the bounding boxes that will be the input for the Re-ID method. Additionally, a people tracker is commonly used to keep track of the position of the targets, while exchanging information with the Re-ID module. The fusion between the tracker and the person re-identifier is key for achieving good performance and is one of the main differences between different approaches.

When applying Person Re-ID methods to mobile robotics, it is also crucial to consider the limitations

of hardware and computational power of the system. A trade-off between performance and computational efficiency is often required.

2.2.1 Person Detection

Person detection consists in locating people in a frame or in an image, without the need for identifying their identity. The detected people are represented by a bounding box, which is a rectangle that provides the person position and size on the image.

There are several methods to detect people. One way is by using an Histogram of Oriented Gradients (HOG) [61], an adaptation of the SIFT approach [62]. This approach consists in computing normalized local histograms of gradient orientations in a dense grid, which represent well local object appearance and shape. These HOG features are then evaluated using an SVM. Variations of this approach have been implemented, such as the Multi-class HOG [63].

Another popular approach are Deformable Part Models (DPM) [64]. These methods create a model of an object based on a global root filter and several part models, which are computed using HOG features. It is able to represent the high discriminability of the full body, while being robust to occlusions [65, 66].

Bourdev et al. proposed a new definition of a body part called poselet, which is based in 2D human annotations and 3D human pose annotations [67]. This method allows the identification of torsos and body keypoints such as the left shoulder, nose and others.

Currently, one of the most popular approaches is training a CNN to detect people, considering their high performance on image-based tasks. These models are trained using large-scale person datasets such as PASCAL-VOC [68] or MS-COCO [69] and have shown very high reliability. Several models and their variations have been implemented such as RCNN [70], R-FCN [71], SSD [72] and YOLO [1]. There are also deep learning-based 2D human poses detectors such as OpenPose [73] and DensePose [74], that can be used for the task of people detection.

Laser sensors can also be used to detect people, by detecting legs [75], or the upper-body [76]. This method is robust to illumination changes and allow the detection of people without RGB data.

2.2.2 Tracking

Tracking is the task of keeping track of the position of a target through time. This task is challenging due to illumination changes, occlusion, clutter, camera motion, low contrast, specularities and more [77]. This task increases in complexity if we consider Muliple Object Tracking (MOT), which aims at tracking the positions of multiple objects at the same time. These objects can be of all types, but I will focus mainly on people tracking, considering the context of this work. The tracking task implies a data association

step, which determines how to associate new detections to existing tracks.

One of the common approaches is the use of a Bayesian estimator such as the Kalman Filter [78] or the Particle Filter [79]. These methods estimate the position of the targets, represented by a state. The state is predicted using a motion model and updated using the incoming measurements. There are two relevant variations of the Kalman Filter developed to work with non-linear systems: the Extended Kalman Filter (EKF) [80], which aims at solving the problem of non-Gaussian distributions of the measurements and prediction models by linearizing them using Taylor series expansions; the Unscented Kalman Filter (UKF) [81], which uses sigma points to give a better aproximation of the behaviour of the system. The particle filter is a sequential Monte Carlo algorithm, that uses particles to represent the posterior distribution of the states given state observations and can deal with nonlinearities [82].

Volkhardt et al. presented a real-time people tracking method based on multi-modal measurements, such as an HOG detector, a face detector, a leg detector and a motion detector, that are fed to a Kalman Filter [83]. Another method also uses an HOG detector to extract person features and estimates their position using a Kalman Filter, while improving detection by using the tracking estimation to narrow the scale of detection [84]. Kalman Filters can also be used to track people skeleton data, providing information about the correlation in time and between body parts [85]. Similar systems have been proposed using the particle filter, for example using histograms of color and edge orientation as person features [86]. In [87, 88], a particle filter models the probability distribution of the position of the target, which is updated using information given by a Radio Frequency Identification (RFID) tag. A Correlation Particle Filter was proposed, consisting in the combination of a correlation filter and a particle filter that, through search region padding and particle refinement, effectively reduces the number of particles needed for accurate tracking [89].

Bazzani et al. compared the Multi-Hypothesis Kalman filter and the particle filter, with results showing that the latter is more robust to occlusions [90]. Other works compared the UKF, the EKF and the particle filter and showed that the UKF can work as well as a particle filter in terms of accuracy and robustness [91]. Therefore, the UKF can be a better option in case the computational resources are limited, which is common in mobile robots.

One example of a particle-filter based tracker is presented in [92], which uses person features obtained by LBP-AdaBoost and HOG-SVM and a greedy data association algorithm.

A multi-target and multi-people detector for mobile robots was presented by the Beta Robots at the RoboCup@home challenge [93], which is based on a particle filter that fuses several sensors such as a laser leg detector, a body detector based on a SVM with HOG, a face detector and a skeleton detector. The data association step is done using a probabilistic tree for each detector. Another approach designed for mobile robots is the Selected Online Ada-Boosting [94], which combines a Online Ada-Boost tracking algorithm with depth images, to increase robustness to occlusions and light or pose

sudden changes.

Neural networks can also be used for tracking, for instance the bilinear Long Short-Term Memory (LSTM) [95], which improves long-term apperance models by using a recurrent network.

In [96] many other trackers are compared and evaluated against common benchmarks. One of the most relevant conclusions drawn is that the common attribute of top performing models is a strong affinity model and that deeply learned models are currently showing the best performance.

Some works have shown the benefits of combining Re-ID with tracking, since both tasks can complement each other. Chen et al. proposed a multi-people tracking system with deeply learned candidate selection and person Re-ID [97]. The person candidates are generated both from detections and tracking, the tracking is done with a Kalman filter and tracklet confidence and data association is done by extracting features using a Region-based Fully Convolutional Network (R-FCN). While Re-ID allows for a better association between candidates and tracks, tracking gives robustness when handling missing detections in crowded scenarios.

A step was made towards end-to-end tracking, in the context of multi-camera Re-ID and tracking using optimal bayes filters [98]. The data association step is avoided, by using LuNet to generate ID-specific measurements and the need of bounding boxes is eliminated by keeping full probability maps, without any assumption about their underlying distribution.

Another method used Re-ID to improve the performance on long-term associations, taking into account appearance changes [99]. A person feature descriptor is computed using LOMO features followed by PCA and motion prediction of the targets is done using a Kalman Filter. This work showed that, when the appearance does not change significantly, it is possible to re-identify a target in a distant position after a long time without seeing it.

Chen et al. investigated the fusion of appearance and spatio-temporal models for Re-ID and tracking [100], combining results from both using linear weighting influenced by a decay function and a rule-based system. For the Re-ID module, appearance is described using color histograms (HSV and LBP), followed by PCA, and classification is done using sequential k-means. The tracking module consists in a Kalman filter, where the measurement model checks the closest previous position of the target.

These methods show the advantages of combining tracking and Re-ID, specially in the data association step, increasing robustness in cases where targets walk out of the scene, i.e. the camera view, and re-enter it, crowded environments and noise. On the other hand, tracking can handle better cases where the target's appearance changes.

2.2.3 Related Work

There are existing methods that implement tracking and Re-ID in the context of mobile robotics. They vary in terms of the type of features they compute and the tracking algorithms used.

Most of the methods are based on hand-crafted features. One method uses color, height and gait features to identify a specific person using a people-following robot [101]. Tracking is based on a Laser Range Finder (LRF), the position of a person is obtained by the LRF and then the upper-body is detected using a cascade HOG classifier. Online boosting tracking is implemented, using three weak classifiers that use one of the three features: color, represented by a Hue-Saturation histogram; height, by determining the sinciput of the head region and then calculating the height based on camera geometry; gait, using the LRF data. This method showed good performance in indoor and outdoor environments, being able to re-identify robustly the target that is being followed. However, the system does not have a predictor of the target's position and it is designed specifically for the task of following a person.

Another method tracks not only the 3D position of the targets but also their upper-body orientation [102]. Person detection is done using a leg detector and HOG upper-body detector and the tracking is done using a multi-hyphotheses Kalman filter. This method evaluates edges and color and also learns the texture of the upper-body of each person, which is used for re-identification. Since it relies on the motion of the targets, it is generally suitable to track the positions of walking persons, while cases where targets are static are not thoroughly evaluated.

As stated before, the relation between tracking and the Re-ID approach is key for the success of the method, and that is the main focus of the work presented by Wengefeld et al. [103]. Several person detectors are evaluated and Part-HOG is chosen as the best performing one. A 7D Kalman-filter based tracker, along with a template-based visual tracker keep track of people's position and upper-body orientation. People re-identification is done through an appearance-based approach, that computes person descriptors using weighted color histograms and Maximum Stable Color Regions, from the SDALF approach. The distance metric learning method used is kernel Local FDA and features are fused at score-level using the PROPER approach [104]. The final decision is then made using probabilistic voting. The Re-ID module improves on wrong ID switches by proximity and the tracking module improves on Re-ID problems due to illumination changes. This work shows once more the importance of fusing Re-ID and tracking, although the actual performance of the method could be better, since the spatio-temporal model is not robust enough to noise and the appearance Re-ID method, which is composed of hand-crafted features, does not perform well when two targets have similar appearances.

As mentioned before, with the development of deep learning, Re-ID methods based on these types of models have been implemented recently, which improve performance comparing to methods that are based on hand-crafted features. One of the main limitations of using these methods in mobile robots is their computational requirements. However, there are some methods that have used deep learning Re-ID models for mobile robots and shown promising results. One of them uses online transfer learning [105]. This method suggests using three CNN's: one for person detection, one for person feature extraction and one for person re-identification. The person detection CNN is the Yolov2 model

trained in the COCO dataset [106]. 512-feature vectors of the detected targets are extracted using another CNN with a triplet-loss function. Finally, the Re-ID CNN receives the feature vectors as input and ouputs a person ID. It incorporates transfer learning since the lower layers are equal to the person feature extraction layer and only the upper layers parameters are updated. A human-in-the-loop online learning approach is also proposed. This method has some limitations: it requires a large dataset for training the neural networks and its performance relies heavily on the quality and similarity of this dataset with the environment where the method will be applied; it was developed assuming a perfect target gallery, where the target is always present in the scene.

Carslen also proposed two new CNN's called LuNet Light and LuNet Lightest with the purpose of implementing Re-ID in mobile robots [107]. The networks use LuNet as a baseline and are trained in the MARS datase [54], using batch hard triplet loss. The features are matched using mean-feature matching between frames using the Euclidean distance as the distance metric. The resulting models achieve close to state-of-the-art performance, while being much lighter, making them suitable for robotic applications, although a deployment on a real robot and an integration with a complete pipeline including a person detector and a tracker was not experimented.

Recently, a novel T-D-R framework for quadruped robots was proposed, including a visual tracker based on a correlation filter, a person detector based on deep learning and a Re-ID module also based on a deep learning model [108]. This system is designed for a real-time tracking and following of a leader in long-term. For this purpose, the result of the tracker and the detector are compared to improve tracking performance, while the Re-ID module handles distractions and occlusions caused by other people. The correlation filter discriminates the leader from the rest of the people present by recording the appearance in long-term. This method is robust in handling occlusions, appearance changes and illumination variations. Although this method shows a very good tracking performance, it is designed for tracking and following a single-target.

The methods presented so far use mainly RGB data and some use laser data for the people detection task. However, depth data is frequently available in mobile robots. Hence, there are some methods that use this type of information. Liu et al. proposed a method for people detection and tracking using Red-Green-Blue-Depth (RGB-D) cameras for mobile robots, that also re-identifies targets through association [109]. A point cloud is divided into subsclusters using meanshift clustering with an Epanechnikov kernel. After that, human candidates are detected in each subcluster, by using plan view maps to describe a spatial region of interest. A human candidate is described by a depth-weighted histogram and is tracked using a particle filter, using Global Nearest Neighbour for data association. Although this method provided good insight into the use of RGB-D data for this task, it does not perform well when the targets are highly occluded by obstacles or other people.

One of the main considerations one must have when developing systems for mobile robots is the

computational efficiency and the system's running frame rate. In [2], a very fast RGB-D people tracking method for service robots is proposed, that can run in real-time at a very high frame-rate even without using Graphics Processing Unit (GPU). It features a novel depth-based sub-clustering method that allows to detect people within groups and standing next to walls. To reduce identity switches, an online appearance classifier is used featuring a three-term joint likelihood. This method achieved state-of-the-art performance in several RGB-D datasets, while being very fast, although it has some limitations such as high ratio of misses when people are gathered in groups and difficulty to track trajectories that do not follow a constant velocity motion model.

2.2.4 Critical discussion

The methods described so far provide meaningful insights and show progresses in developing a multi-target tracker based on a Re-ID module to be deployed in a mobile robot. They show that the integration of person re-identification with tracking benefits performance and that there are lightweight methods for feature extraction that are discriminative and allow for robust person re-identification in a mobile robot. However, they have some limitations since they either rely on constrained conditions on the environment, such as open-spaces without occlusions, and the movement of people, such as standard poses or walking motion, or their tracking and re-identification is not robust enough in challenging conditions such as the presence of obstacles or similar targets. It is also important to note that there are not many existing methods designed for multi-people tracking using Re-ID features on mobile robots. Hence, there is need for a development of a Re-ID based multi-people tracker designed to be deployed in a mobile robot working in an environment with obstacles and occlusions.

3

Methodology

Contents

The system proposed in this thesis is a multi-people tracker based on a Re-ID module for deployment in a service mobile robot. This chapter presents the system architecture and the details of its several components, followed by a description of the track's state estimation, data association and track management procedures. The chapter is organized as follows: Section 3.1 gives a brief introduction to the relevant coordinate frames, Section 3.2 gives an overview of the system architecture and its main components, Section 3.3 presents the tracks and state estimation, Section 3.4 details the data association step and finally Section 3.5 presents the track management approach.

3.1 Coordinate frames

Before presenting the overall system architecture, it is important to define the relevant coordinate frames of our problem.

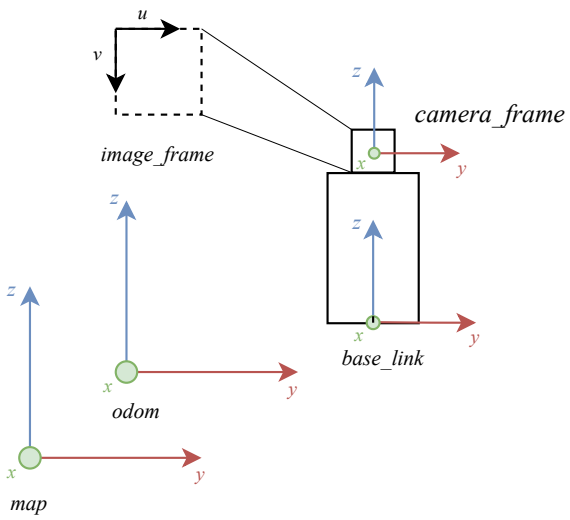


Figure 3.1: Coordinate frames

Figure 3.1 contains a schematic showing the relevant coordinate frames. If we consider the robot, four relevant frames can be identified: the 3D frame centered in the base of the robot, $base\ link$, the 3D world frame of the odometry of the robot, $odom$, the 3D frame centered in the camera of the robot which moves along with the camera, $camera\ frame$, and the 2D frame that represents pixels on the camera image, $image\ frame$. We also have the map coordinate frame which is fixed and represents the world and the environmnet where the robot is moving. The transformations between $odom$ and $base\ link$ and between $odom$ and map change based on the odometry errors. Corrections are calculated based on the localization of the robot and are introduced in the transformation between $odom$ and map, in a way that the transformation between $odom$ and $base\ link$ represents the odometry of the robot and at the same time the transformation between $base\ link$ and map represents the localization of the robot.

People tracking can be done in 2D, if each target is tracked in the image plane, or in 3D, if the target is tracked in a world coordinate frame. Considering our goal is to track people in the world so that the robot knows their location and can interact with them, the tracking in this work is done in the map frame, i.e. in the world frame, hence a person's position is given by 3 coordinates, (X, Y, Z). The other coordinate frames presented will be used by several modules of the system.

3.2 System architecture overview

The proposed system aims at tracking multiple people in the scene while assigning an unique ID to each one of them. The system is going to be deployed in a mobile robot, which requires a computationally lightweight solution.

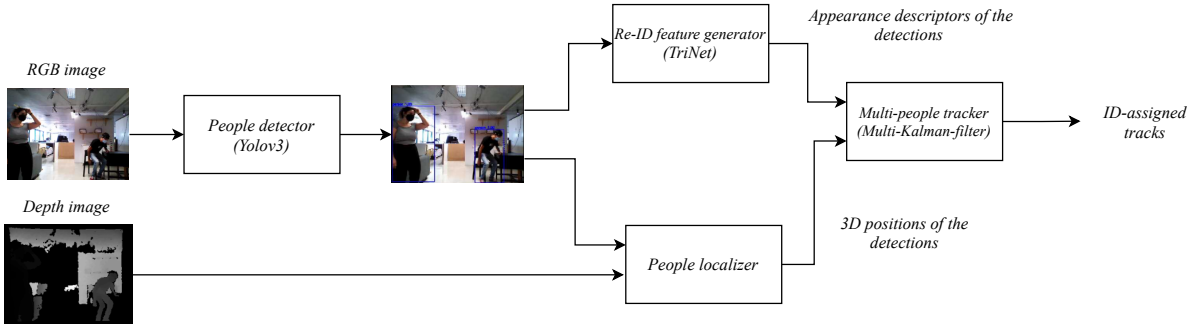


Figure 3.2: System architecture overview

An overview of the system architecture is presented in Figure 3.2 and is summarized here:

Input: The input of the system consists in a RGB image and a depth image, captured by the robot with a RGB-D camera positioned on the head. The RGB image is fed to the people detector and the depth image serves as input for the people localizer module.

People detector: The people detector module receives as input the RGB image and outputs the detected people in the image, in the form of bounding boxes in image coordinates. People detection is done using Yolov3 [110], which is an object detection convolutional neural network.

People localizer: This module takes as input the depth image and the bounding boxes from the people detector and outputs the 3D position of the detected people in the map coordinate frame.

Re-ID feature generator: The Re-ID feature generator is responsible for generating appearance feature descriptors of the detected people. It takes as input the bounding boxes from the detections and outputs a 128-feature descriptor of each target. This module uses the deep neural network TriNet [24].

Multi-people tracker: The multi-people tracker is composed of multiple Kalman-filters, one for each track. It receives as input the positions in the map frame and the appearance descriptors of each

target and outputs ID-assigned tracks, that correspond to the people being tracked in the scene.

3.2.1 People detector

The people detector module is responsible for detecting the people present in the scene. It takes as input the RGB image taken by the robot's camera and outputs bounding boxes in the $image\_frame$, that represent the detected people. Bounding boxes are rectangles in the image plane that enclosure a detection and are represented by (u, v, h, w), where (u, v) is the bounding box center position and (h, w) are the height and width of the bounding box, respectively.

For this task, a trained model of Yolov3 is used, which is an improvement of the object detection convolutional network Yolo [1]. This network was chosen because it is very fast and robust, making it a very reliable and suitable solution for person detection in a robotic context. An example of detections generated by Yolo can be seen in Figure 3.3.

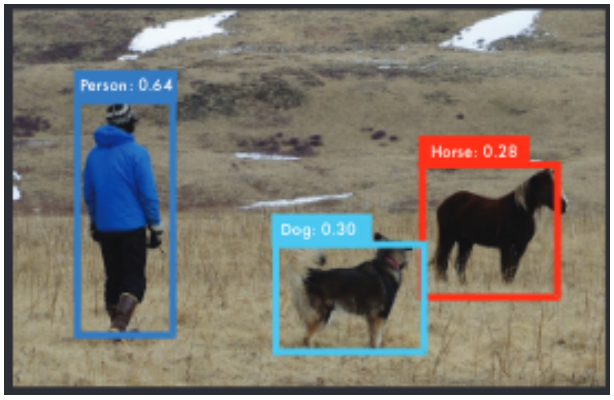


Figure 3.3: Example of Yolo detections, taken from [1]

The network architecture consists on 24 convolutional layers followed by 2 fully connected layers. The convolutional layers extract features from the image and the fully connected layers predict the output probabilities and coordinates. The network is trained in the large-scale object detection dataset COCO [106]. Yolo and Yolov3 can be used to detect a great variety of objects but in this work only the detections belonging to the class 'person' are considered, since we are only focused on people tracking. An example of a bounding box generated by the people detector module of the proposed system, can be seen in Figure 3.4

3.2.2 People localizer

The people localizer module converts detections in the image plane to 3D positions in the world frame. For that, it takes as input the bounding boxes from the people detector and the depth image taken by the robot's camera. First, it takes the center of the bounding box and, using the $image\_geometry$

Figure 3.4: Example of a person detection

Robot Operating System (ROS) package [1], it calculates the unit vector in the $camera\ frame$ that passes through the pixel corresponding to the center of the bounding box in the $image\ plane$. By default, the unit vector (x, y, z) has z equal to 1. The unit vector is multiplied by a depth value to obtain the position of the target in the map frame. The depth value is determined by finding the region in the depth image that corresponds to the bounding box and getting the 25th percentile of the depth values from that region. This is a good estimate of the depth of the person relative to the $camera\ frame$ because the region in the depth image enclosing the person will have some high depth values originated by the background, as can be seen in Figure 3.5, that should not be taking into consideration. Therefore, taking the 25th percentile of the depth values of that region will give a good estimation of the lower values which are more representative.

In this work people tracking is done in the world coordinate frame, therefore the positions obtained by the people localizer, which are in the $camera\ frame$, are then converted to the map frame. To do that, they are first tranformed to the $odom$ frame and then transformed to the map frame. The transformation between $odom$ and map depends on the localization of the robot, which is running in parallel. After the conversion, we get the 3D position in the world frame of every target present in the scene. Regarding the z position, since the point obtained by the people localizer refers to the center of the bounding box of the detection, it will represent approximately the height of the center of the body of the person. This information can be useful to determine if a person is sitting or laying down, but cannot be used to compare people's heights, for instance.

[1] http://wiki.ros.org/image_geometry

Figure 3.5: Depth image showing two people. In this image, a brighter colour represents points that are further away from the camera. Totally black represents a NaN point where the depth could not be obtained. We can see here that a person is surrounded by background points that have bigger depth.

3.2.3 Re-ID feature generator

To be able to differentiate people in the environment and re-identify them when they exit the scene and reappear, a Re-ID module is required. This module computes feature descriptors that represent a target's appearance. These feature descriptors are then compared to decide if a person has been seen or not and who that person is. As mentioned before, the goal is to assign an unique ID to a person and keep that same ID troughout the execution of the system, even if the person exits and re-enter the scene.

As presented in the previous chapter, there are several methods that can be used to generate feature descriptors and each of them has their own advantages and disadvantages. In this work, the neural network TriNet [24] is used, due to its robustness and light computational effort, which is key for the deployment of the method in a mobile robot. Comparing to hand-crafted techniques, neural networks achieve much better re-identification performance and considering our tracking environment has several challenges such as illumination, viewpoint and pose changes and occlusions, robustness is a very important requirement. Hermans et. al [24] was also one of the few works that made pre-trained models available to deploy out-of-the-box as feature extractors. Besides that, variations of this model had already been used in other works with good results [98, 107].

TriNet uses ResNet-50 pre-trained on the ImageNet dataset as a baseline. The two last layers of ResNet-50 are discarded and two fully connected layers are added. The first has 1024 units, followed by batch normalization and ReLu. The second has 128 units and it's the output layer of the network. TriNet is trained with batch hard triplet loss and the model used in this work was trained in the MARS dataset.

This already trained model was chosen over a model trained on the Market-1501 dataset because MARS is a video dataset and the tracking will be made on video.

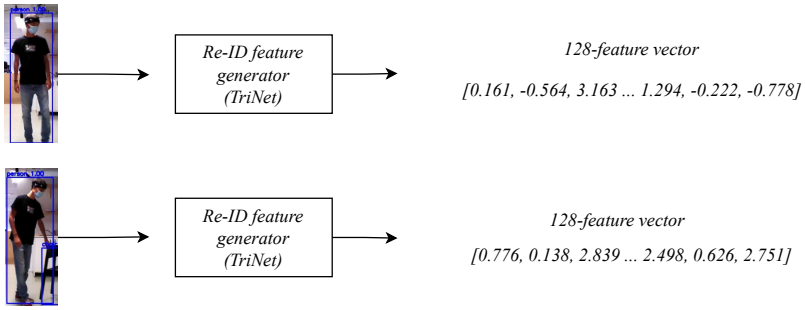


Figure 3.6: Example of Re-ID feature extraction using the Re-ID feature generator.

The Re-ID feature extractor takes as input the people detections from the people detector, feeds them to TriNet and outputs a 128-dimensional feature vector, which is the appearance descriptor, for each detection. An example is shown in Figure 3.6. Each feature vector contains 128 float values that describe the characteristics of the input image and that can be compared to match similar people. In each layer of the network, different features are computed. For instance, the lower convolutional layers compute low-level features such as color, texture and edges. The layers deeper in the network gradually compute higher-level features such as the separation between upper-body and lower-body, the shape of the body and other features that are useful for re-identification.

The set of appearance descriptors belonging to each target will then be used in the Multi-people tracker, to assign ID's to the tracks and manage them.

3.2.4 Multi-people tracker

The Multi-people tracker implemented in this thesis is composed of multiple single-hypothesis Kalman filters and frame-by-frame data association using appearance descriptors and was inspired by Deep SORT [111]. A Kalman filter approach was chosen because it is lightweight while achieving good tracking performance and, when combined with an appearance metric, it allows for fast and robust tracking of multiple targets.

The tracker is composed by a set of Kalman filters, one for each track. The several Kalman filters do not relate to each other, each one of them represents one and only one track. Each person is tracked using a simple Kalman filter, that predicts and updates the person's position in the map coordinate frame. The appearance descriptor generated by the Re-ID feature extractor is used to associate detections to tracks and to manage the creation and elimination of tracks. At each frame, the tracker decides which tracks to keep, delete or create, along with the Kalman filters associated with them. The track management and data association methodology are described in the next chapter. The multi-tracker

execution loop is ilustrated in Figure 3.7.

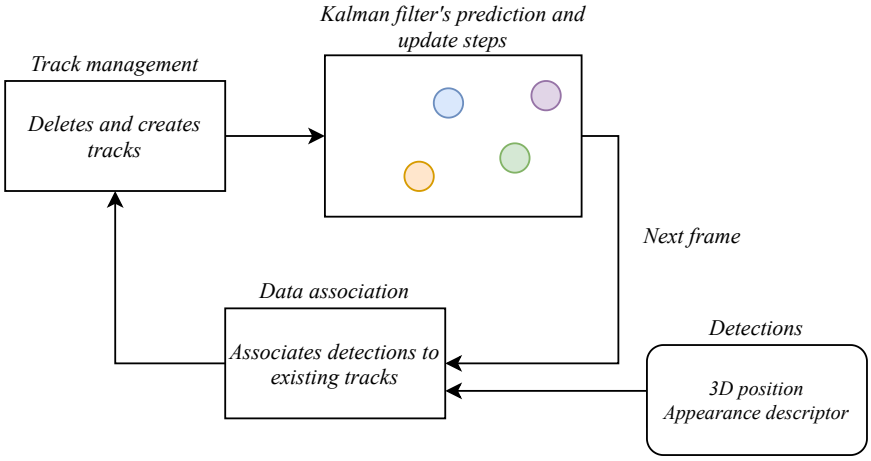


Figure 3.7: Multi-tracker execution loop. At each frame, the tracker receives as input people detections, including their 3D position and their appearance descriptor. Next, the data association step associates detections to existing tracks, followed by a track management step that deletes and creates tracks when needed. Finally, each Kalman filter associated with existing tracks executes a prediction and an update step. In the figure, each colored circle ilustrates a single Kalman filter, corresponding to an existing track. This loop repeats every frame during the execution of the system.

A general overview of the Kalman filter algorithm that is applied for each track is presented in Figure 3.8. When a track is initialized, a new Kalman filter is initialized with an initial state and covariance. At each timestep, which in this case corresponds to a frame, the state is then predicted using a motion model that models the person's movement from one frame to another. The state is then updated using a measurement of the position of that person, if available. The measurement is a vector containing the 3D position of the target in the map frame, given by the people localizer module. At each timestep, the Kalman filter outputs the estimated track state. The motion and estimation models, the state composition and the relevant covariance matrices are presented and explained in the next chapter.

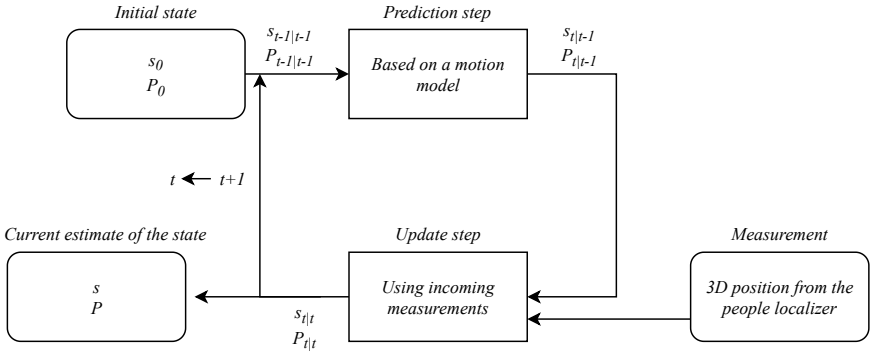


Figure 3.8: Kalman filter algorithm, where s, P and t are the state, the covariance matrix and the timestep, respectively

3.3 Tracks and state estimation

In order to track a target's position through time, a Kalman filter is initialized for each new track. The Kalman filter will predict and update the target's position in the world at each frame. Each track's state is modelled as

$$\hat{x} = (x, y, z, vx, vy, vz), \tag{3.1}$$

where x, y and z are the positions in the X, Y and Z axis of the world frame, respectively, and vx, vy, vz are their corresponding velocities. A new track is initialized with the position of the target, initial velocities are considered zero and an uncertainty is also assigned to the state, represented by the following covariance matrix:

$$P_0 = \begin{bmatrix} {\sigma_x}^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & {\sigma_y}^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & {\sigma_z}^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & {\sigma_{vx}}^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & {\sigma_{vy}}^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & {\sigma_{vz}}^2 \end{bmatrix} \tag{3.2}$$

where σ is the standard deviation of each of the state variables, with the following values, that were previously determined experimentally:

$$\sigma = \begin{bmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \\ \sigma_{vx} \\ \sigma_{vy} \\ \sigma_{vz} \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 1 \\ 1 \\ 1 \end{bmatrix} \tag{3.3}$$

At each frame, the states are predicted using a constant velocity model for the x and y positions and a zero velocity model for the z position, described by matrix A. The prediction step is described by:

$$\hat{x}_{k|k-1} = A_k \hat{x}_{k-1|k-1} \tag{3.4}$$

$$\hat{x}_{k|k-1} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \hat{x}_{k-1|k-1} \tag{3.5}$$

Using the constant velocity model for the x and y positions is a common approach taken in 3D people tracking, since from frame to frame we can assume that people keep their velocity constant in those directions. Regarding the z position, which in this work represents approximately the height of

the center of the body of the target, it is reasonable to assume that this position won't change from one frame to another. Significant changes in z occur if a person changes its pose by standing, sitting down or bending, which are cases that cannot be predicted to occur from one frame to another. This cases will impact the value of the z measurement and will affect the state in the update step.

Uncertainty about the state increases in the prediction step, so the covariance is recalculated. The process noise covariance matrix Q is the same as the initial covariance matrix P_0 and it is used to update the covariance matrix, P in the prediction step:

$$P_{k|k-1} = A_k P_{k-1|k-1} {A_k}^T + Q_k \tag{3.6}$$

At each frame, a measurement of the target's position can be received. This measurement Z is given by the people localizer module and gives information on the 3D position of the target, (x, y, z), in the world frame. The update step is performed using a linear observation model where the target's position Z, is taken as a direct observation of the target's state, using the following measurement matrix:

$$H_k = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{3.7}$$

Using the measurement of the targets position Z, the current state vector s, and the measurement matrix H, an innovation factor y is obtained:

$$y_k = Z_k - (H_k \hat{x}_{k|k-1}) \tag{3.8}$$

The uncertainty associated with the measurement is also calculated. The measurement uncertainty is the following:

$$R_k = \begin{bmatrix} {\sigma_x}^2 & 0 & 0 \\ 0 & {\sigma_y}^2 & 0 \\ 0 & 0 & {\sigma_z}^2 \end{bmatrix} \tag{3.9}$$

The measurement uncertainty matrix indicates how reliable the values of the measurements are. Using the above measurement uncertainty matrix, the innovation covariance associated with the measurement step S is calculated:

$$S_k = H_k P_{k|k-1} {H_k}^T + R_k \tag{3.10}$$

Using the above calculations, the Kalman gain K is computed. The Kalman gain is the weight given to the measurements and the state estimation, stating which one should be trusted more. It is given by:

$$K_k = P_{k|k-1} {H_k}^T {S_k}^{-1} \tag{3.11}$$

The elements of K will be larger if the measured values do not match the predicted state and will decrease otherwise. The state and the state covariance are then updated:

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + (K_k y_k) \tag{3.12}$$

$$P_{k|k} = (I - K_k H_k) P_{k|k-1}, \tag{3.13}$$

where I is the identity matrix.

The track's states are predicted and update with incoming measurements at each frame, however, deciding which measurement is associated to which track is key to maintaining a track's correct ID and to perform a good state estimation. This task is handled by a data association method, that will be described in the next section.

3.4 Data association

The data association task is handled as an assignment problem where we have a set of detections that have to be assigned to a set of existing tracks and a detection can be associated to one and only one track and vice-versa. Each detection has its own position, (x, y, z), and an appearance descriptor. Each existing track has also a 3D position given by its state and an appearance gallery associated with it.

In this problem, two metrics are used: a spatial distance metric and an appearance metric. The reasoning behind using these two metrics is that they both provide different information about how likely a detection is to be related to a track. A detection can be spatially very close to a track but if their appearances are very different, then they probably are not the same person. On the other hand, a detection can appear very similar to a track's known appearance, but be spatially very distant from it, which is an indication that it might not belong to the same target.

The spatial distance metric between a detection d, and a track t, is computed by calculating the squared Mahalanobis distance D, following this expression:

$$s(d,t) = D(d,t)^2 = (x_d - m_t)^T \cdot {P_t}^{-1} \cdot (x_d - m_t), \tag{3.14}$$

where x_d is the vector containing the x and y position of the detection, m_t is the vector containing the x and y position of the track and P_t is the covariance matrix associated to the track. Using the Mahalanobis distance gives a better indication of how close a measurement is to the position of the track, since it also takes into account the uncertainty associated with it. The z position is not taken into account in this metric because, as stated before, it represents approximately the height of the center of the body of a person and that is not a variable that allows to match detections to tracks. Although the z position of a detection and a track that belong to the same person should be similar, a person can change its pose

dramatically from one frame to another, which will affect the z position and would erroneously affect the spatial distance metric. The x and y position will remain close from one frame to another, and that is why they are the variables used to account for spatial proximity.

The appearance metric is calculated by computing the smallest Euclidean distance between the 128-dimensional appearance descriptor of the detection, and the appearance descriptors present in the track appearance gallery. This gallery is composed of all the appearance descriptors associated to the track since its initialization. The appearance metric computation between a detection and a track is given by:

$$c(d,t) = min(d(l_d, l_i)|l_i \epsilon L_t) \tag{3.15}$$

where

$$d(v,u) = \sqrt{(v_1 - u_1)^2 + (v_2 - u_2)^2 + ... + (v_{127} - u_{127})^2 + (v_{128} - u_{128})^2}, \tag{3.16}$$

l_d is the detection appearance descriptor, l_i is the i-th appearance descriptor of the track and L_t is the track gallery containing all the appearance descriptors associated with the track.

By storing all the previous appearances of a target in the gallery and finding the minimum appearance distance to one of its elements of the gallery, this metric evaluates how similar a detection is to a target, based on all the history of appearances of that target. This approach was chosen over an average of all the appearances because a target's appearance can change greatly with time, due to illumination or viewpoint changes, and averaging the appearance descriptors would result in a less informative and discriminative representation of a person's appearance. Using this approach, if a person's apperance changes significantly because of a sudden illumination change, for instance, the appearance descriptor associated with that frame will be saved and if in a later point in time that person suffers the same appearance change, this metric will find a very small distance between that current appearance and the previously saved appearance.

Algorithm 3.1 describes the data association algorithm. The algorithm receives as input the set of detections and existing tracks in the current frame. Empty sets are initialized for associations, unmatched detections and unmatched tracks. The assignment problem is represented by a cost matrix and the cost between a detection and a track is the appearance distance, as can be seen in line 5. A distance matrix is also computed, with the spatial distances between detections and tracks, in line 6. Then, all possible associations are assessed considering both metrics, from line 7 to 12. The goal of this step is to assign an "infinite" cost, which in our case is a very large number ($10e5$), to associations which are not admissible considering their combination of appearance and distance metrics. A range of values of the appearance metric between two thresholds, T_{lower} and T_{upper}, is considered, where the spatial distance determines if the association is admissible or not. If the appearance distance is in that range and the

Algorithm 3.1: Data association

1 **Input:** Tracks T, Detections D
2 $A \longleftarrow \emptyset$ // Initialize set of associations
3 $U_d \longleftarrow \emptyset$ // Initialize set of unmatched detections
4 $U_t \longleftarrow \emptyset$ // Initialize set of unmatched tracks
5 $C = \left[c_{d,t}\right]$ // Compute cost matrix
6 $S = \left[s_{d,t}\right]$ // Compute distance matrix
7 **for** *each detection* d **do**
8 **for** *each track* t **do**
9 **if** $T_{lower} < C[d,t] < T_{upper} \wedge S[d,t] > S_{max}$ **then**
10 $C[d,t] =$ *INFINITE COST*
11 **else if** $C[d,t] \geq T_{upper}$ **then**
12 $C[d,t] =$ *INFINITE COST*
13 $A, U_d, U_t \longleftarrow hungarian\_algorithm(C, T, D)$
14 **for** *each association* (d,t) *in* A **do**
15 **if** $C[d,t] =$ *INFINITE COST* **then**
16 $A \longleftarrow A \setminus (d,t)$
17 $U_d \longleftarrow U_d \cup d$
18 $U_t \longleftarrow U_t \cup t$

spatial distance is above S_{max}, the association is given an infinite cost. This step is important because it discards associations where a detection is not similar in appearance to a track. At the same time, it keeps associations where, although the appearance is not that similar, the spatial distance between them is very close, which strongly indicates that they belong to the same target. An example where this procedure proves useful is when a target's appearance changes suddenly from one frame to the next and thus the appearance metric is heavily affected, but if the target did not move unexpectedly, the spatial distance remains very small and the association is considered admissible. The upper limit of the appearance range is used to discard completely an association where the detection and the track are not similar at all and, even if they are spatially very close, they cannot belong to the same target. This can occur if a person is occluding the other or if they are standing very close. The spatial and appearance thresholds were determined experimentally, through several tests, and the values used are:

$$T_{lower} = 300,\ T_{upper} = 700,\ S_{max} = 0.05m \tag{3.17}$$

After checking both metrics, the assignment problem is solved using the Hungarian algorithm [112], in line 13. This algorithm solves the association between detections and tracks with the minimum possible cost. Given that one detection can be assigned to one track only and vice-versa, if the amount of

detections and tracks is not the same, unmatched detections and tracks will be identified. From the algorithm's output, the association set is filled. Finally, the association set is iterated and the cost of each association is checked. If it is the infinite cost, it means that this association was considered not admissible in the previous steps and it is discarded, with the respective detection and track being added to their respective sets.

3.5 Track management

Following the data association stage, we get a set of associated detections and tracks and a group of detections and tracks that were not associated. Unmatched detections are used to create new tracks and unmatched tracks are used to delete existing tracks, which is done in a track management step.

Each track has an associated state indicator, which can be $Confirmed$, $Tentative$ or $Deleted$. When a track is initialized it is assigned the $Tentative$ state. A $Tentative$ track changes to $Confirmed$ if there is an association with a detection for three consecutive frames. A $Confirmed$ track is considered $Tentative$ if there is no association at the current frame. A $Tentative$ track changes to $Deleted$ if there is no association for five consecutive frames. When a track is $Deleted$, the corresponding track appearance gallery, L_t, containing all the appearance feature descriptors previously associated with that track, is saved to memory. The three state indicators and the conditions that determine when to change states are represented in Figure 3.9 by a state-machine.

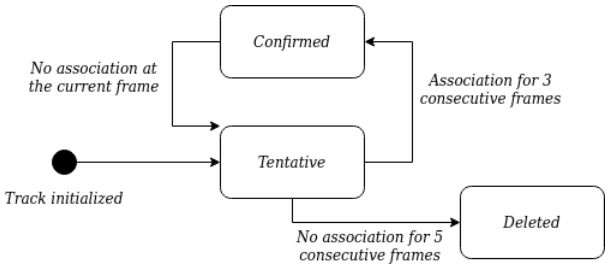


Figure 3.9: Track state indicators

At each frame, both $Confirmed$ and $Tentative$ tracks serve as input to the data association algorithm. Based on the data association result, the state indicator of each track can change and finally only $Confirmed$ tracks are the output of the overall system. They identify the different people in the scene that the tracker is tracking with confidence. $Tentative$ tracks serve two purposes: one is to prevent keeping tracks that were created based on an erroneous detection, eg. a false positive, and that's why when a track is initialized it remains as $Tentative$ for three frames until it can be considered $Confirmed$; the other purpose is to keep tracks of targets that are occluded only for a short time, and should continue to be tracked after they reappear in the scene, and that's why a $Confirmed$ track switches to $Tentative$ for five frames before being deleted.

The initialization of new tracks is done based on the unmatched detections at each frame. For each unmatched detection, a new track is initialized. Assigning an ID to a new track is a very important step. If the new track corresponds to a person that was seen before, the ID that was previoulsy assigned to that person should be assigned to the new track. At the same time, we have to be careful to not assign a previously used ID to a track that does not belong to the same person. To accomplish that, the appearances of all previously seen people are saved in memory in a dictionary I, which is updated by appending the track gallery L_t, to a list containing all the appearance descriptors previously associated with that track's ID. This track gallery was introduced in the previous section in Equation (4.15). In this dictionary, all the previous IDs and their corresponding appearance descriptors, combined in track galleries, are listed. Every frame, this dictionary is updated with the track galleries of the tracks that were deleted in that frame. When those tracks have an associated ID that is already present in the dictionary, the track gallery is appended to that ID's list of galleries. Otherwise, a new entry is added to the dictionary. An example of the structure and update step of the ID dictionary is shown in figure 3.10.

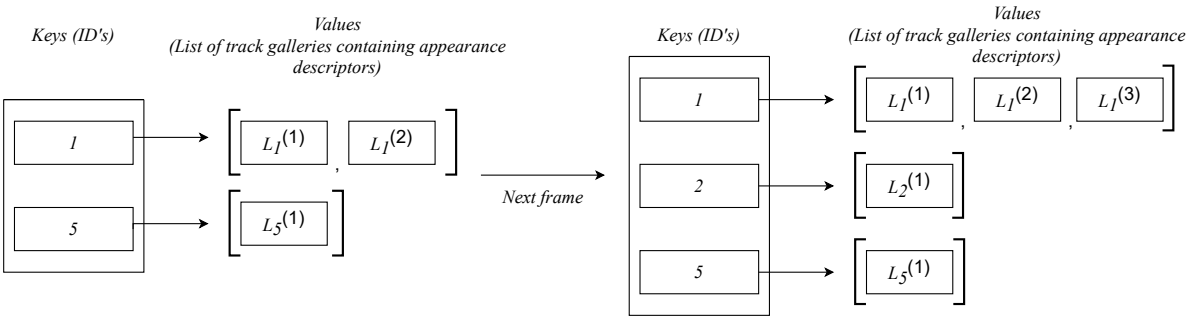


Figure 3.10: ID dictionary, I example. Each dictionary entry contains a list of track galleries, L_t. Each track gallery contains several 128-feature vectors. In this example, from one frame to the next, two tracks are deleted: one with the ID 1 and the other with the ID 2. Hence, the dictionary is updated: a new L_1 is appended to the ID 1 list and a new ID, with the number 2, is added to the dictionary with its corresponding track gallery being saved.

To determine which ID is assigned to a new track, the algorithm described in Algorithm 3.2 is performed. The appearance distance between the new track and all previously seen people is computed using the appearance distance metric described before. The exception is that ID's that are associated with tracks currently being tracked are not possible ID's to be recovered and assigned to a new track, hence, are excluded from this comparison. The appearance of the new track is given by the detection that initialized it. The appearances of all the previously seen people are available in the ID dictionary. The appearance distance metric is designed to be computed between a detection and a track, therefore, in this case, we take the new track as the detection and the track galleries stored in the dictionary as the tracks, as can be seen in line 5. By finding the minimum distance between the new track and all the track galleries of a specific ID, we get the minimum distance of the new track to that ID, which is stored in a list in line 6. That procedure is repeated for every ID in the dictionary, except if the ID is being tracked at the

current frame, and the minimum distance to an ID is calculated (lines 3 to 7). That distance represents the best possibility of the new track belonging to a previously seen ID. We check that distance against an appearance threshold, $T_{recover}$, and if it is smaller, that ID is assigned to the new track (lines 7 and 8). If the minimum distance to an ID is above $T_{recover}$, the new track is assigned a new ID, in line 11. The threshold $T_{recover}$ was determined experimentally and has the following value:

$$T_{recover} = 400 \tag{3.18}$$

This value is higher than the value of T_{lower}, because in the experiments it was observed that usually when a person re-enters the scene its current appearance is less similar to the previous than if we compare them frame to frame, hence this threshold is higher in order to recover more frequently previously seen people. It was observed that if the threshold is lower, there are some cases where a person that was seen before gets a new ID because the appearance distance was slightly above the threshold. At the same time, considering that usually newly seen people are not that similar to previously seen targets, increasing the threshold did not lead to cases where a new track was assigned a previous ID erroneously.

Algorithm 3.2: ID assignment to a new track

```
1  Input: New track t, ID dictionary I, list of active ID's B, next unused ID, next_id
2  Z ⟵ ∅                                          // Initialize list of appearance distances
3  for each k in I do
4      if k is not in B then
5          V = [c(t, I[k]^(i))]                   // Compute appearance distance vector
6          Z ⟵ min(V)                             // Append minimum of V to list Z
7  min_distance = min(Z)                          // Find minimum of Z
8  min_id = argmin(Z)                             // Find respective ID
9  if min_distance < T_recover then
10     Track t's ID = min_id
11 else
12     Track t's ID = next_id
13     next_id = next_id + 1
```

4

Re-ID Multi-Tracking Dataset

Contents

In this chapter, first a review of the state-of-the-art of 3D Multi-object tracking datasets is presented, where the current short-comings are pointed out. Next, a novel re-identification multi-people tracking dataset is presented. The composition of the dataset and examples of the dataset sequences are provided.

4.1 3D Multi-Object tracking datasets: State-of-the-art review

Regarding multi-object tracking, the most important benchmark is the MOT Challenge [113]. It provides some of the largest datasets for pedestrian tracking, including ground-truth and detections, recorded with static and moving cameras. The detections are provided since the quality of the detections heavily impacts the performance of the tracker but usually the tracking module is independent from the detection, thus, comparing the performance of different trackers is easier if they all use the same detections [114]. Some of the datasets available in the MOT Challenge are MOT15 [115], MOT16 [116], MOT19 [117] and KITTI [118]. The great majority of these datasets provide ground-truth of the positions of the targets in image coordinates, i.e. in 2D. MOT15 provides ground-truth of the 3D position of targets but only on 4 of the 22 sequences, where the camera is static, and the 3D positions were obtained using camera geometry calculations rather than with physical sensors, which introduces error. The datasets provided by the MOT Challenge include mostly crowded and open-space outdoor areas, featuring with pedestrians walking, which is not representative of the typical environment of the application considered in this work. A domestic environment is indoors and it is not an open space, targets will be frequently occluded by furniture and they can be sitting down or assuming other poses. Besides that, MOT Challenge datasets do not provide depth images, only RGB, hence they cannot be used to evaluate and test the method proposed in this thesis, since it requires depth information of the targets.

Table 4.1: RGB-D multi-human tracking datasets

Dataset	Year	#Sequences	#Frames	#Camera	Track ID's	Environment
ETH	2008	8	5017	Static	No	Busy pedestrian zones
UHD	2011	3	1130 per sequence	Static	Yes	University Hall
StanfordRGB-D	2012	35	4500 per sequence	Static (17) and moving (18)	Yes	Office, hallways and corridors
KTP	2012	5	8475	Static (1) and moving (4)	Yes	Office
KingstonRGB-D	2014	6	1000 per sequence	Static	Yes	Laboratory
SD	2015	10	-	Static	No	Indoor shop

There are several RGB-Depth multi-human tracking datasets that have been presented in the past years. Some of the most popular ones, refered here with the same names as in [119] for clarity, are the following: the ETH dataset [120], the University Hall Dataset (UHD) [121], the StanfordRGB-D dataset [122], the Kinect Tracking Precision Dataset (KTP) [2], the KingstonRGB-D dataset [123] and the SD dataset [124]. A comparison of these datasets is summarized in Table 4.1. The datasets vary in number of sequences and frames, in the status of the camera (static or moving), the existence of annotated track

ID's and the environment in which they were taken. Out of the datasets considered, only two of them were obtained using moving cameras, which is very important to evaluate a method to be deployed in a mobile robot, since the camera movement impacts heavily the accuracy of the position estimation and introduces other errors. The two datasets where the camera is moving were taken in an open-space environment where targets are not occluded by other objects such as furniture and they do not include sequences where targets are assuming different poses such as sitting down.

Analyzing the state-of-the-art on 3D Multi-object tracking datasets, we can see that there is an overall lack of datasets aimed at tracking in 3D and that the ones that exist are very limited in the conditions in which they were taken. They depict very crowded scenes and most of them were taken using a static camera. The existing datasets show outdoor areas or open-space indoor areas. There is also a lack of robust and reliable ground-truth of 3D positions of targets, along with depth information besides RGB, which is data commonly used by 3D tracking methods. Considering the application of the method proposed in this thesis, there is a need for a multi-target dataset with track ID's and target 3D position ground-truth taken in an apartment-based environment with occlusions caused by obstacles in the scene and taken by a moving camera. A dataset with these characteristics was not found in the literature, therefore a novel Re-ID multi-target tracking dataset is proposed, in the next section.

4.2 Re-ID Multi-Tracking Dataset

In this section a new RGB-Depth dataset is proposed, called Re-ID Multi-people tracking dataset, acquired from a mobile robot moving in a domestic environment testbed equipped with a motion capture system. This dataset was built to test and evaluate 3D position accuracy and people re-identification performance of multi-target tracking methods based on RGB-D data.

4.2.1 Data and ground truth collection

The data was collected by teleoperating the MBOT in the ISRoboNet@Home Testbed[1] (Figure 4.1) with up to 3 targets moving in the environment. The robot is equipped with a tilt-controlled Orbbec Astra RGB-D camera positioned on the head that captures RGB and depth images with 640 x 480 pixel resolution at 30Hz. The testbed is an apartment-like environment designed to benchmark service robots and is equipped with a motion capture system composed of 12 OptiTrack® "Prime 13" cameras (1.3 MP, 240 FPS), which provides real-time tracking data of rigid bodies with sub mm precision in 6 dimensions with low latency (4.2ms).

Although the camera frequency is 30Hz, the recording of the dataset was done at a lower frequency of approximately 10Hz, resulting in a total of 3144 RGB images, 3437 depth images and 2154 people

[1] https://welcome.isr.tecnico.ulisboa.pt/isrobonet

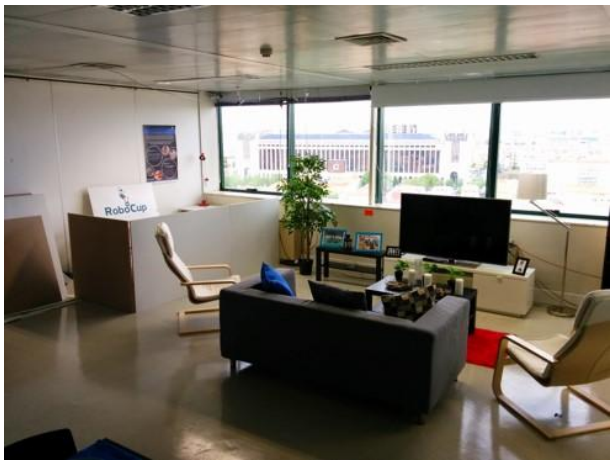


Figure 4.1: ISRoboNet@Home testbed

instances.

The RGB-D images, camera information, a map of the environment on the form of an occupany grid along with map metadata, odometry of the robot, transforms along reference frames and ground-truth are made available as ROS bag files[2].

People detections originated by the people detector module of the proposed system described earlier have also been included in the dataset. This information cannot be considered as ground truth because it is generated by an automatic people detector, but it was included because it can be of great utility for future works that focus solely on the tracking algorithm and require people detections beforehand. Besides that, as mentioned before regarding the MOT Challenge, providing public detections is very useful for comparing the tracking performance of different methods, independently of their detection performance.

As ground-truth, 3D positions of people in environment were obtained using the motion capture system, together with ground-truth of the robot's position. Markers were placed on the targets and on the robot. Track ID's were also obtained directly by the motion capture system. 3D ground-truths of targets that were out of the field of view of the robot or completely occluded were manually deleted. Besides that, there were frames where the motion capture system failed, due to the positioning of the cameras and markers that were not visible, and the 3D ground-truth of some of the targets was not registered. In these cases, ground-truth was not associated with these frames and they should not be considered when evaluating performance metrics. After this process, ground-truth is associated with approximately 70% of the frames. This process is important because keeping frames with lack of ground-truth of some targets would lead to errors when evaluating methods on this dataset, such as the occurrence of incorrect false positives (cases where the method outputs a track that is not present in the

[2] https://ulisboa-my.sharepoint.com/:f:/g/personal/ist187134_tecnico_ulisboa_pt/Emz8wKesZThJs0_TNoR1mTkBWJ6JjriW-0le5CsfWRd3ig?e=KfJtlT

(a) $Still$ **(b)** $Moving\ head$ **(c)** $Moving\ base$ **(d)** $Chairs$

(e) $People\ following$ **(f)** $Changing\ clothes\_1$ **(g)** $Changing\ clothes\_2$

Figure 4.2: Dataset sequences examples. From (a) to (f) we can see an example of a single frame of each sequence of the dataset. On (g) three frames of the $Chaning\ clothes\_2$ sequence are shown, where the target has different clothes in each one of them.

ground-truth).

4.2.2 Dataset sequences description

The dataset consists of 7 videos with durations ranging from 40s to 1:10s. Each video contains different characteristics (camera and people movement) and represents different cases, so that the dataset is representative of several situations that can occur in an environment with multiple people and obstacles. The 7 sequences (videos) present in the dataset are the following:

- $Still$: sequence recorded with a static camera. Three targets move around freely in front of the camera without being occluded by obstacles.
- $Moving\ camera$: sequence recorded with the camera rotating while the robot's base does not move. Three targets move around freely in front of the camera without being occluded by obstacles.
- $Moving\ base$: sequence recorded with the robot moving around the environment. Three targets are present and are frequently occluded by obstacles. One of the targets also sits down and gets up again during the sequence.

- $Chairs$: sequence recorded with the robot moving around the environment. Three targets are sitting down in chairs around two tables and during the sequence they get up, walk around and switch places several times.
- $People\ following$: sequence recorded with the robot being teleoperated to follow a specific person around the environment. During the sequence, three targets are present and there are several occlusions caused by obstacles and people crossing paths.
- $Changing\ clothes\ 1$: sequence recorded with the robot moving around the environment. Two targets are present. Both of the targets change their clothing during the sequence while in front of the camera.
- $Changing\ clothes\ 2$: sequence recorded with the robot moving around the environment. Two targets are present. One of the targets exits the scene and re-enters with different clothes twice.

These sequences cover most of the common cases that can occur in a domestic environment. There are several occlusions caused by furniture such as chairs, tables and a sofa or caused by other people when targets cross paths with each other. A specific case where the robot is following a person was also recorded, since this is a common task executed by mobile service robots. The last two sequences represent cases where targets change their clothes during the sequence, which is a challenging scenario for people re-identification. This dataset also has the particularity that all of the people present are wearing cirurgical masks, due to the Covid-19 pandemic. An example of RGB-D frames from the sequences described above is presented in Figure 4.2. Statistics of the sequences and the overall dataset are also presented in Table 4.2

Table 4.2: Re-ID Multi-tracking dataset statistics

Sequence	Duration (s)	#RGB images	#Depth images	#People instances	%Frames with ground-truth	#ID's
Still	39.5	398	454	326	90.3	3
Moving head	42.1	374	457	237	74.1	3
Moving base	60	590	574	490	71.7	3
Chairs	52.2	424	544	366	66.5	3
People following	39.0	399	394	174	42.1	3
Changing clothes 1	68.0	621	676	347	67.6	2
Changing clothes 2	56.7	338	338	214	75.4	2
Total	357.5 (5:75s)	3144	3437	2154	70.1	-

Each of the sequences pose different challenges and can be very helpful when developing or evaluating a multi-people tracker. The two $Changing\ clothes$ sequences are aimed at evaluating the re-identification performance of multi-people trackers or other methods, since the challenging scenario where a target changes clothes is presented. Since the maximum number of targets present in a single sequence is 3, the dataset is not suitable to evaluate a tracker's performance in crowded environments, but it is representative of a domestic or office environment where usually there are not many people present in the scene at the same time.

The proposed Re-ID Multi-Tracking Dataset contains 7 sequences that are very different from each other and that represent common cases in multi-people tracking in a indoor and occluded environment, making it a suitable dataset for developing and evaluating 3D multi-target tracking methods that use RGB-D data.

5

Experimental Results

Contents

In this chapter first the system implementation on the MBOT is described. The evaluation metrics that were used are described, the experiments that were conducted are detailed and the evaluation results on two datasets are presented. The results are analyzed and compared with the results of another state-of-the-art method in a common dataset.

5.1 Implementation

The system was deployed in the MBOT, that features two on-board computers with i7 processors, one dedicated for navigation and the other for human-robot interaction and a NVIDIA GeForce 1060 6gb GPU. As stated before, the robot also features a tilt-controlled Orbbec Astra RGB-D camera positioned on the head.

The system was implemented using ROS and Python and consists in the following ROS nodes:

$darknet\_ros\_py$: It was already implemented in the MBOT and is used as the people detector in the system. This module source code was written in Python.

$mbot\_object\_localization$: It was already implemented in the MBOT and is used as the people localizer module. This module source code was written in C++.

$re\_id\_tracker$: Developed in the context of this thesis, including the Re-ID feature generator and the multi-people tracker. This module source code was written in Python.

The $darknet\_ros\_py$ and $mbot\_object\_localization$ nodes publish their outputs in ROS topics which are subscribed by the $re\_id\_tracker$. One implementation challenge was the synchronization of the messages that were being received by the $re\_id\_tracker$ node, considering that the publishing frequency of the people detector and the people localizer were not equal. The people localizer module receives as input the detections generated by the people detector so the people localizer messages always come with a small delay. To overcome this issue, people detection messages are stored in a queue and when a people localizer message arrives, both messages are matched by timestamp.

5.2 Evaluation metrics

Before detailing the experiments and evaluation results, it is important to clarify which evaluation metrics were used and what they consist in. All the experiments that will be detailed below were evaluated using these metrics and understanding them is key to analyze the performance of the method.

All the experiments conducted in this thesis were evaluated using the CLEAR MOT metrics [125]. These metrics are the most frequently used metrics for evaluating multiple object tracking performance and are one of the adopted metrics on the multi-object tracking benchmark MOT Challenge [113].

CLEAR MOT are composed of two separated metrics that tackle different aspects of tracking performance: the MOTP and the MOTA.

Before detailing the calculation of these metrics, it is important to refer how the matching between tracker hypotheses and ground-truth objects is done in the CLEAR MOT framework. The first step to evaluate a tracker's performance is to establish a correspondence between the tracker hypotheses and the real objects present in the scene. In CLEAR MOT, the correspondence is done based on the spatial distance between them, as long as the distance between a pair of hypothesis-object does not exceed a threshold. This threshold represents the maximum distance that can be considered an error in position estimation. Beyond that threshold, we consider that the tracker hypothesis has to belong to someone else. In the case of this thesis, the value chosen for this threshold is 1 meter. This means that if a tracker hypothesis is more than 1 meter away from an object, they cannot be associated. Besides this consideration, consistent association through time is also a concern, hence pairs of hypotheses-objects that were associated in the previous frame are preferred over other pairs in the current frame, even if their distance is larger (although it has to be always below the threshold). Taking into account this considerations, the matching is done using the Hungarian algorithm. These are the most important aspects of the matching procedure, but a more detailed description can be consulted in [125].

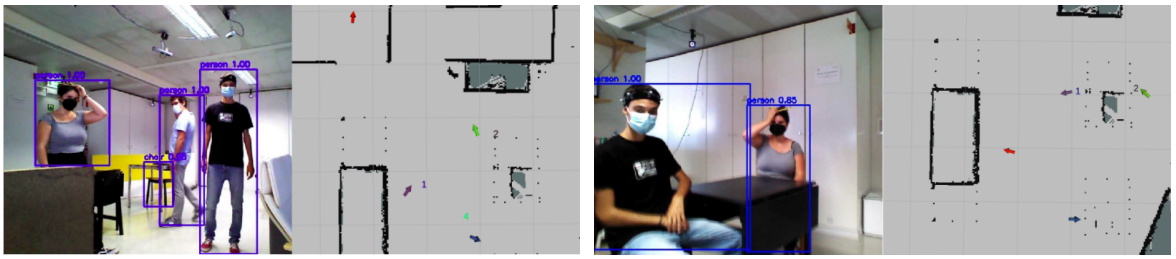
Following the matching of hypotheses and objects, there may be hypotheses and objects that were not matched. The hypotheses that were not matched are counted as false positives, fp. The objects that were not matched are counted as misses, m. Besides that, CLEAR MOT also counts the number of mismatches, mme, or ID switches, that represent the number of times where an ID was wrongly associated with a person. This can occur in two ways: when a person is being tracked with an ID and then switches to another ID and when a new person is seen and it is given an ID that was previously associated with another person. This counts are indicative of these specific types of errors and will be used for the computation of the MOTA metric.

The MOTP shows the ability of the tracker to estimate the position of the targets, regardless of its skill in assigning identities and keeping trajectories. It is a metric that represents how big is the error of the tracker in estimating positions of tracked people. It is given by:

$$MOTP = \frac{\Sigma_{i,t} {d_t}^i}{\Sigma_t c_t}, \tag{5.1}$$

where ${d_t}^i$ is the distance between the hypothesis and the object of the ith matched hypothesis-object pair of frame t and c_t is the number of matches made in frame t. MOTP is the total error in estimated position for matched hypothesis-object pairs averaged by the total number of matches made.

The MOTA takes into account all object identity and track errors, such as false positives, misses and mismatches. It measures how well a tracker assigns identities to targets, keeps trajectories and recognizes people in the scene. It is given by:



(a) $Moving\ base$ **(b)** $Chairs$

Figure 5.1: Examples of the tracker output in the sequences $Moving\ base$, figure (a), and $Chairs$, figure (b). On the left, the RGB image and the output of the people detector are shown. On the right the map of the environment is shown, with the ground-truth positions of the targets and the robot and the tracker output displayed. Only a 2D position (X and Y) is displayed for easier visualization. The red arrow and the other three arrows represent the ground-truth position of the robot and the ground-truth position of the three people in the environment, respectively. The output of the tracker is represented by a number. Each number indicates a track and is located on the estimated position of that person.

$$MOTA = 1 - \frac{\Sigma_i(m_t + fp_t + mme_t)}{\Sigma_t g_t} \tag{5.2}$$

where m_t, fp_t, mme and g_t are the misses, false positives, mismatches and objects seen in frame t, respectively. MOTA is the sum of all object configuration errors averaged over the total number of objects seen.

These two metrics assess different abilities that a tracker must have and, combined with the other errors, provide a good evaluation of a tracker's performance.

5.3 Experiments on the Re-ID Multi-Tracking Dataset

5.3.1 Evaluation results

The method was evaluated on the novel Re-ID Multi-Tracking Dataset using the CLEAR MOT metrics. Examples of the system's output for two sequences of this dataset are presented in Figure 5.1. Examples of the target's trajectories in every sequence and the tracks generated by the system are presented in Appendix A.

Results divided by sequence and in total can be seen in Table 5.1. The MOTP in meters and the MOTA score in percentage are shown in Figure 5.2. The results presented in the table are: the number of objects seen, i.e. number of ground-truth of people registered, the number of matches between a tracker hypothesis and a ground-truth object as computed by the CLEAR MOT evaluation procedure described earlier, the number of misses and the ratio of misses relative to the number of objects, the number of false positives and the false positives ratio relative to the number of objects, the number of

Table 5.1: Re-ID Multi-Tracking Dataset experiment results divided by sequence and in total. The ratios of misses, false positives and ID switches, in percentage, are given relative to the number of objects seen.

	Objects	Matches	Misses	Misses (%)	False Positives	False Positives (%)	ID Switches	ID Switches (%)	Recall errors
Still	330	267	15	4.55	4	1.21	0	0.00	0
Moving head	261	213	10	3.83	3	1.15	0	0.00	0
Moving base	522	296	51	9.77	16	3.07	13	2.49	4
Chairs	429	160	26	6.06	11	2.56	6	1.40	3
People follow	186	72	11	5.91	5	2.69	4	2.12	0
Changing clothes 1	280	172	15	5.36	6	2.14	1	0.36	0
Changing clothes 2	226	173	34	15.04	17	7.52	9	3.98	3
Total	2234	1353	162	7.25	62	2.78	33	1.48	10

ID switches, i.e. mismatches. Besides these results, an additional count is reported, that is not included in the CLEAR MOT metrics, which is the number of recall errors. The recall errors are the number of ID switches that occurred by initializing a track with an ID that was previously associated with a different person. This count is useful to access how the tracker recalls previously seen people and how many times a person entering the scene is assigned a and ID that belonged to someone else. The number of recall errors is always a portion of the ID switches and the smallest the better.

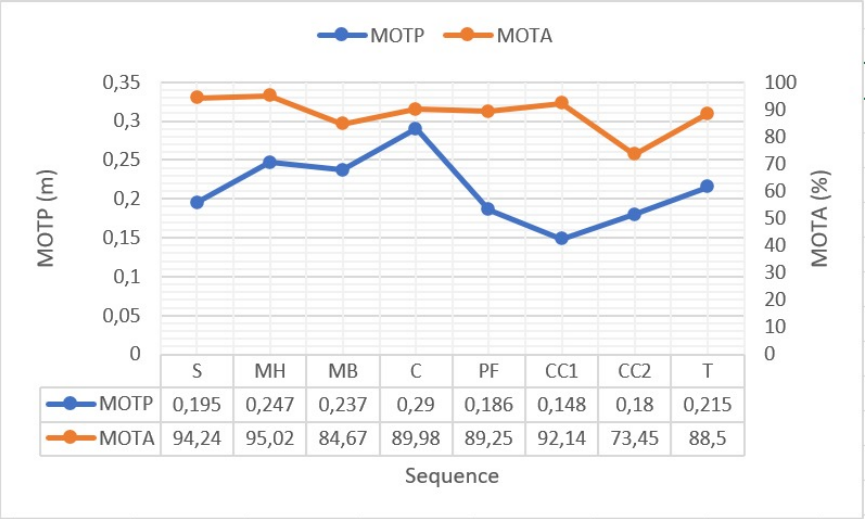


Figure 5.2: Re-ID Multi-Tracking Dataset MOTA and MOTP scores achieved by the system, divided by sequence and in total. The letters in the horizontal axis represent the differente sequences: Still (S), Moving head (MH), Moving base (MB), Chairs (C), People follow (PF), Changing clothes 1 (CC1), Changing clothes 2 (CC2) and Total (T).

The total MOTP value is 0.215 meters which shows a good target position estimation, since it means that the average error in people position estimation was only 22cm, approximately. The system was able to track the target's positions accurately in every sequence, although we can see greater error in three sequences, $Chairs$, $Moving\ base$ and $Moving\ head$, which are the sequences where the targets and the robot are moving the most. As the robot moves, the error in the robot's localization increases, which also impact the error in the transformation calculation between coordinate frames. The error in

the transformation between frames will increase the error in the position given by the people localizer module, increasing the error in the tracker's position estimation.

In the $Chairs$ sequence the MOTP is the highest, with almost 30cm of error, which can be explained by the fact that in this sequence the people are constantly stopping and sitting down. Since a constant velocity motion model is used in the Kalman Filter prediction step, when people suddenly stop and keep their position, the predicted positions will follow a constant velocity trajectory and will have a greater error until a measurement is made and the position is updated. It is interesting to note that in the $People\ follow$ sequence, where the robot and the targets move a lot around the room, the error is below 20cm. This is due to the fact that, since the robot is following a target, the matches made in this sequence are mostly belonging to a single person that is being seen by the robot almost all the time. The position estimation error is low because the target's position is being constantly updated. Another reason may be that this sequence has the lowest percentage of frames with ground-truth (see Section 5.2.2) which results in less ground-truth 3D positions to compare against the system's output.

The total MOTA score is 88.50% which shows a very good performance in assigning unique ID's to targets, keeping trajectories and identifying people in the scene. This result shows that almost 90% of people instances that were seen were correctly tracked with a unique ID. The highest MOTA was achieved in the $Moving\ head$ and $Still$ sequences, where the robot's base is static. In these sequences the targets are being seen by the robot almost all of the time and their movements are mostly linear, which reduces the complexity in estimating their identities and keeping track of their trajectories. The lowest MOTA is obtained in the challenging $Changing\ clothes\ 2$ sequence, where the maximum number of misses, false positives and ID switches occur.

The ratio of misses and false positives on the complete dataset is 7.25% and 2.78%, respectively. These ratios show that the system produces a very low number of false positives and that only less than 10% of people instances were missed. The misses occur because when a new track is initialized, the system only outputs a track after 3 frames with a correct association, as described in section 4.3. This approach is important to decrease the number of false positives caused by erroneous people detections, but inevitably increases the number of misses in those 3 frames before a track is outputed by the system after the detection of an unseen person or the initialization of a new track due to an ID switch.

As can be seen in the table, the number of misses is directly correlated with the number of ID switches, since most of ID switches represent cases where a new track is initialized. There are other cases that can lead to misses, such as when a person is heavily occluded and no detection is given by the people detector or when a person's position estimate is too far away from the ground-truth and no match between a tracker hypothesis and an object is made. The latter situation is rare, but when it occurs, it also produces the false positives that are reported. The ratio of misses and false positives are much higher in the $Changing\ clothes\ 2$ sequence than in the other sequences, which also leads to

the lowest MOTA score. This is due to two reasons: this sequence is the most challenging in terms of targets appearance changes, since one of the targets changes clothes twice and exits the scene several times, which leads to many ID switches and consequently more misses; this sequence was the only one that was recorded in a different day than the others and there are ground-truth errors that could not be completely eliminated by manually removing frames, such as errors in the estimation of the position and in the ID's assigned to the targets. Evaluation in this sequence is reported anyways because it shows that, even with some ground-truth errors and in a very challenging scenario with heavy appearance changes, the system still achieves a MOTA above 70%, which is acceptable.

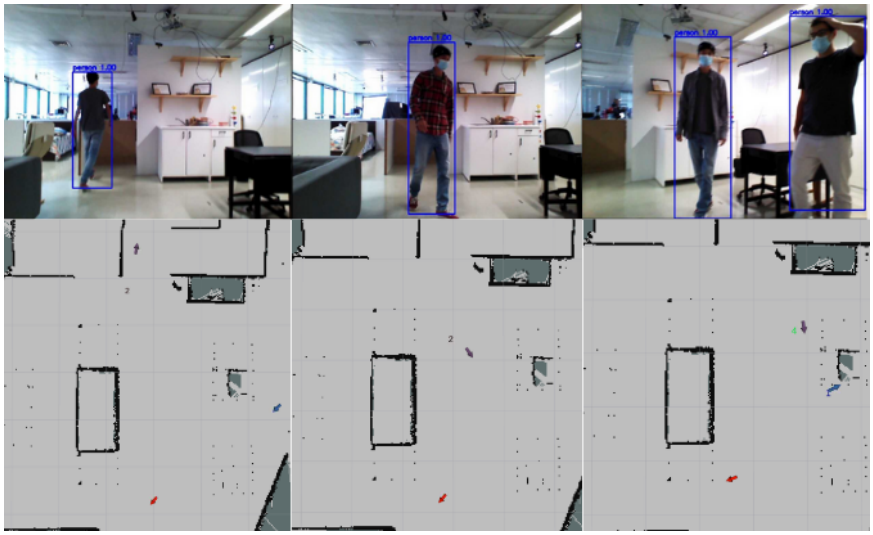


Figure 5.3: Experiment on the $Changing\ clothes\ 2$ sequence. The images and the elements in the map represent the same as in Figure 5.1. Three frames from the sequence and the respective tracker output are shown. In the first frame, a target is identified with ID 2. The other target is not being seen by the camera, so it has no track associated with it. In the second frame, in the middle of the figure, we see a correct re-identification of the target seen in the first frame, now wearing different clothes, and the ID 2 is re-assigned to that target. In the last frame, the first target is wearing different clothes again and we see that an ID switch occurs, because the tracker assigns the ID 4 to that target. In the last frame we can also see the atribution of ID 1 to a target that was not shown before.

The number and ratio of ID switches are some of the most relevant results because, considering the system is based on a Re-ID module, one of the main goals is to have the least amount of ID switches as possible. In the complete dataset, the system produces only 1.48% ID switches, which shows a very good performance in keeping target's ID's and associating a specific ID to only one person. Out of the total 33 ID switches, 10 were recall errors, caused by assigning a previously seen ID to a wrong person when initializing a track. The recall errors were caused by different targets with similar appearances, that the system could not tell apart.

We can see that in the $Still$ and $Moving\ head$ sequences there were no ID switches, since as stated before the targets were in the camera view most of the time and the system could easily distinguish targets using both spatial distance and appearance metrics. In the $Moving\ base$ sequence, the system

produces the second worst ID switch ratio, which can be explained by the fact that in this sequence the robot moves a lot around the room and the targets are often occluded by obstacles and other people, besides exiting and re-entering the field of view often. All of theses situations make the targets appearance change a lot, so when they re-enter the scene they are sometimes assigned a different ID than before.

It is also interesting to note that the ID switches ratio in the $People\ following$ is higher than the average. In this sequence there are cases where targets cross between the camera and the person being followed, and in this cases the camera only sees a very small portion of the people that is crossing, since it is very close. This results in a track with a very specific appearance, corresponding to only a part of the person's body. When the system detects this person again, and the full-body is seen, the person's appearance is completely different and a different ID is assigned, leading to an ID switch. In this sequence there are effectively no ID switches associated to the track of the target that is being followed by the robot, only to the tracks of people passing by.

The results in the $Changing\ clothes\ 1$ sequence show that changing clothes while in the camera view does not pose an identification challenge for the system, leading to only one ID switch. Since the system constantly updates the target's appearance, a change in appearance while the person is being tracked is registered in the track's appearance gallery and that person continues to be assigned the same ID. When the target is later seen and its appearance corresponds to one of the appearances previously associated with its ID, that same ID is assigned to it again.

Finally, in the $Changing\ clothes\ 2$ sequence the results show the highest ratio of ID switches. In this case, contrary to the previous, the person changes clothes while out of the camera view. Although in some cases when it re-enters the room the system still manages to assign the same ID, there are other cases where a different ID is assigned, as ilustrated in Figure 5.3, which leads to the number of ID switched reported. In the figure, we can see that when the target that was first identified with ID 2 exits the scene and re-enters with different clothing, the system correctly re-identifies it. On the last frame, an example is shown where a re-identification error occurs, since the target is now identified with ID 4. In this example, the target's clothes are very similar to the clothes of the other person in the scene, which caused a recall error. Although it is not shown in the figure, the ID 4 was previously assigned to the other target.

Another important consideration is the real-time performance of the method. In this dataset, the system achieved a 33Hz frame rate, which is suitable for real-time robotic applications.

5.3.2 Parameters fine-tuning

The system has several parameters that were fine-tuned to achieve the best performance possible. The parameters that were analyzed and tuned were T_{lower}, T_{upper}, S_{max} and $T_{recover}$. The values of these

parameters were previously chosen when developing the system, by a qualitative analysis of the output of the system. To find the optimal values for these thresholds and check the initial assumptions, the CLEAR MOT metrics were computed by running the method on the full Re-ID Multi-Tracking Dataset, i.e. in every sequence, with varying values of the parameters. Additionally, the recover errors were also counted. In Figure 5.4 and Figure 5.5, the values of MOTA, MOTP, ID switches and recall errors in total are reported, when varying S_{max} and $T_{recover}$. In these figures, MOTP is given in percentage, for better visualization. This percentage represents the position accuracy relative to 1 meter, which was the threshold used to determine matches in the CLEAR MOT procedure, as proposed in [125]. In Table 5.3, the values of MOTA, MOTP, ID switches and recall errors are also reported for different combinations of T_{lower} and T_{upper}.

First, the value of the $T_{recover}$ threshold was fine-tuned. This threshold determines when a new track is assigned a previously seen ID. A higher threshold means that the appearance distance between a new track and a previously seen person can be higher and will increase the amount of times that a new track is assigned previously seen ID's. A small value of this threshold leads to the opposite. Next, the value of S_{max} was also looked into. This parameter determines the maximum spatial distance that a detection can be from a track in order to be considered that it belongs to the same person, when the appearance threshold is between the range [T_{lower}, T_{upper}], as detailed in section 4.2.

Looking at the evaluation results on the Re-ID Multi-Tracking Dataset on Figure 5.4 and Figure 5.5 we can see that varying $T_{recover}$ and S_{max} does not impact the MOTA and MOTP scores greatly. Although the number of ID switches varies, the MOTA score is practically constant in every experiment which can be explained by the fact that the number of ID switches is always low when comparing to the number of objects seen, thus, the impact on MOTA is not high. On the other hand, the values of these parameters affect the data association and track management steps, which are not related to the track's position estimation, so it was expectable that the MOTP would not change either. To compare the optimal values of these two parameters, the number of ID switches and recall errors is compared. Although the MOTA score is similar, it is relevant to compare the number of times a track switches ID's and, out of the total ID switches, how many times a previously seen ID is wrongly associated to a target. The smaller these numbers are the better. Following this logic, it is clear that, out of the 5 values tested, the optimal values for $T_{recover}$ and S_{max} are 400 and 0.05m, respectively.

For values of $T_{recover}$ below 400, the number of recall errors drops, since the recover threshold is low, while the number of ID switches rises. This shows that, for values below 400, the tracker assigns previoulsy seen ID's to new tracks much less, which leads to assigning different ID's to the same person when the target exits and re-enters the scene, for instance. For values above 400, the number of recall errors and ID switches increase because more targets are considered to be the same person.

In Table 5.2, the mean and the minimum values of the appearance distances between a detection

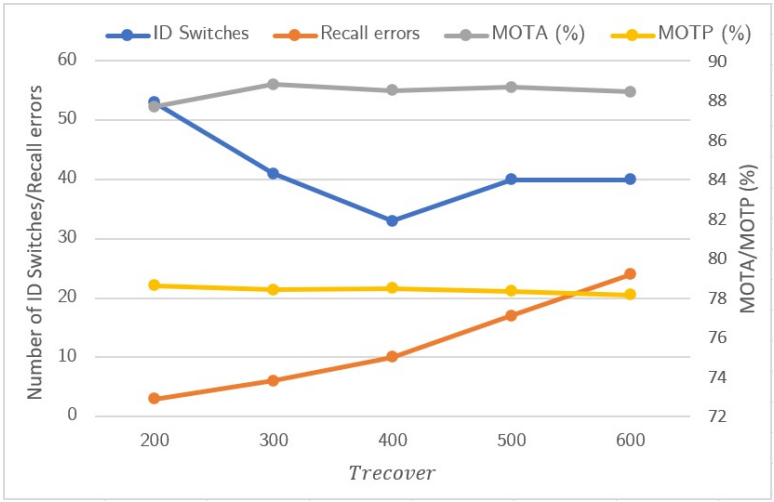


Figure 5.4: Evaluation metrics on the Re-ID Multi-Tracking Dataset when varying the value of $T_{recover}$. MOTA and MOTP are given in percentages and ID switches and recall errors are the total number of these occurrences on the complete dataset. The values of the other parameters are: S_{max} = 0.05m, (T_{lower},T_{upper}) = (300,700)

	Mean	Distance
Appearance distance	301.7	40

Table 5.2: Mean and minimum values of the appearance distance between a detection and a previously seen target, for the cases where an ID was recovered, on the Re-ID Multi-Tracking Dataset. The value of $T_{recover}$ when running this experiment was 600.

and a previously seen target, for the cases where an ID was recovered are reported. The results were obtained by running the system on the complete dataset, using a value of 600 for the $T_{recover}$ threshold. The goal was to check the average appearance distance while using the highest value considered for the threshold. The average appearance distance between a track's appearance and a previously seen ID is around 300 and the minimum of that distance in the complete dataset was 40. This shows that the appearance distance between two instances of the same person can be as low as 40, but at the same time, due to the frequent appearance changes, it can also be much higher. This result backs the decision of using a value of 400 for the $T_{recover}$ threshold, since this is a value that is just above the average, which will allow for the recall of ID's even in case of appearance changes. Increasing the threshold more does not make sense, since it would increase the recall errors, as stated before.

When S_{max} is zero, it corresponds to the case where the spatial distance metric has no impact and the data association step is done using solely the appearance metric. In this case, the MOTA score is slightly lower than in the other cases and the number of ID switches is the highest. Since the spatial distance is not being considered, in this case detections are matched to tracks based only on appearance and that leads to many errors when targets have similar appearances. This result shows that combining the spatial and appearance metric is very important for the performance of the tracker. By increasing the S_{max} value to 0.25, the number of ID switches increases by just 1, while the MOTA score

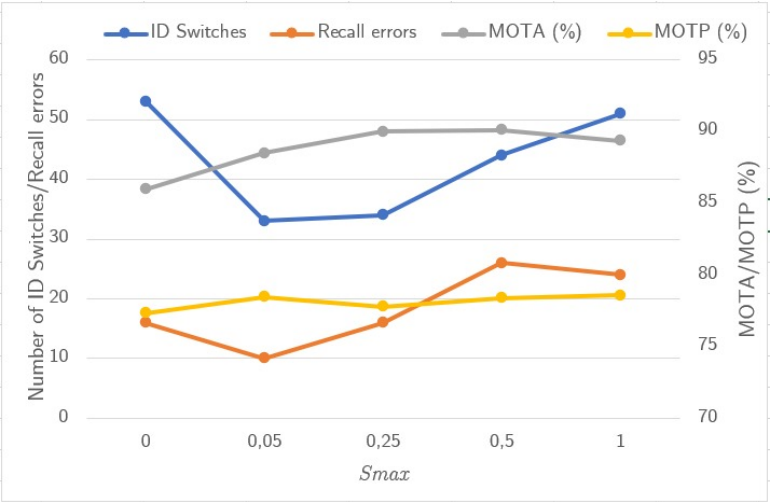


Figure 5.5: Evaluation metrics on the Re-ID Multi-Tracking Dataset when varying the value of S_{max}. MOTA and MOTP are given in percentages and ID switches and recall errors are the total number of these occurrences on the complete dataset. The values of the other parameters are: $T_{recover}$ = 400, (T_{lower},T_{upper}) = (300,700)

Tlower	Tupper	ID switches	Recall errors	MOTP (m)	MOTA (%)
100	700	70	25	**0.213**	86.41
100	1000	72	24	0.215	86.33
300	700	**33**	**10**	0.215	88.50
300	1000	37	16	0.220	89.85
400	600	34	19	0.218	**90.17**

Table 5.3: Evaluation metrics on the Re-ID Multi-Tracking Dataset wen varying the values of Tlower and Tupper. MOTP is given in meters and represents position estimation error, hence, the lower the better. MOTA is given in percentages and ID switches and recall errors are the total number of these occurrences on the complete dataset. The best results for the various metrics are highlighted in bold. The values of the other parameters are: $T_{recover}$ = 400, S_{max} = 0.05m

is slightly better, but there are 6 more recall errors. The value of 0.05 for S_{max} was chosen because it is more important to have less recall errors, which in practice impact more the perception of the robot of the identities of the people present in the scene, specially since the increase in MOTA for an S_{max} of 0.25 is not substantial and the number of ID switches is almost the same.

To determine the optimal values for T_{lower} and T_{upper} it is necessary to consider the combination of both, since they define a range of values of appearance distance between a detection and a track where the spatial distance metric will also be taken into account. Five pairs (T_{lower},T_{upper}) were tested and the results can be seen in Table 5.3. The MOTP is practically constant in every case, because this parameters also do not impact the track's position estimation. A big range of values, (100, 1000), corresponds to the case where the spatial distance has much more weight in the data association step than the appearance distance, and we can see that it does not perform as well as the other combinations: the number of ID switches is the highest and the MOTA is the lowest. It is interesting to note that even if

we decrease T_{upper} to 700, while keeping T_{lower} at 100, the number of ID switches is almost the same and the number of recall errors even increases by one. This shows that what is really affecting performance is the low value of T_{lower}, that is causing a lot of ID switches by discarding correct associations between detections and tracks based on the spatial distance. The best performance is achieved for the pair (300,700). Although the MOTA is higher for the pair (400,600), the difference is very small and the number of ID switches and recall errors show that the pair (300,700) is the optimal combination. For the combination of values (400,600), the range of appearance distance values for which the spatial distance is taken into consideration in the data association step is smaller, which leads to more associations based only on appearance distance, which causes more errors when similar people are present in the scene.

Considering the results obtained by testing on the Re-ID Multi-Tracking Dataset, the optimal values for the system's parameters and that lead for the best tracking results are:

$$T_{recover} = 400,\ S_{max} = 0.05,\ T_{lower} = 300,\ T_{upper} = 700$$

5.4 Evaluation on a test sequence

The proposed system performance was evaluated in a test sequence. This sequence was recorded in the same conditions as the Re-ID Multi-Tracking Dataset. The ground-truth were obtained using the same method described in chapter 5. The sequence statistics are presented in Table 5.4.

Table 5.4: Test sequence statistics

Sequence	Duration (s)	#RGB images	#Depth images	#People instances	%Frames with ground-truth	#ID's
Test sequence	112 (1:52s)	1301	813	824	50.5	3

This test sequence includes challenging scenarios such as targets sitting down, crossing paths with each other and frequent occlusions by obstacles and other people. The sequence was recorded with the robot moving around the environment while the targets walked randomly and assumed different poses. **This sequence was not used to tune the system's parameters**, therefore the performance of the tracker in this sequence provides valuable insight into how the system performs in unseen scenarios. The tracking results on the test sequence are reported in Table 5.5.

The results show a very good performance on the test sequence. The system achieves a MOTA

Table 5.5: Tracking results on a test sequence, compared with the results on the Re-ID Multi-Tracking Dataset

	Objects	Matches	Misses(%)	False Positives (%)	ID Switches	ID Switches (%)	Recall errors	MOTP (m)	MOTA (%)
Test sequence	690	174	7.82	2.86	14	1.99	4	0.190	87.25
Re-ID Multi-Tracking Dataset	2234	1353	7.25	2.78	33	1.48	10	0.215	88.50

Table 5.6: Tracking results for the Kinetic Tracking Precision dataset of the proposed system and the system presented in [2], divided by situation. Best results by situation are shown in bold.

	Situation	ID switches	Misses (%)	False Positives (%)	MOTP (m)	MOTA (%)
Proposed system	Back and forth	**0**	**0**	**0**	0,306	**1**
SOAM	Back and forth	1	8,5	2,4	**0.196**	88.97
Proposed system	Random walk	23	**8,9**	**4,7**	0,355	**85.30**
SOAM	Random walk	**20**	18,9	9,8	**0.171**	70.93
Proposed system	Side by side	**5**	**5,9**	1,9	0,386	**89.35**
SOAM	Side by side	**5**	11,6	**1,2**	**0.146**	87.22
Proposed system	Running	**2**	5,66	2,0	0,350	88.68
SOAM	Running	4	**4,4**	**1,1**	**0.143**	**94.57**
Proposed system	Group	30	**11,68**	**2,1**	0,364	**80.98**
SOAM	Group	**26**	42,53	9,1	**0.181**	47.91

score of 87.25% which is very close to the MOTA score on the Re-ID Multi-Tracking Dataset and a MOTP of 0.190m which is even lower than the MOTP achieved on the proposed dataset. The miss, false positive and ID switches ratios are almost the same, although slightly higher on the test sequence. The system produced some ID switches and recall errors but most of the time the targets were consistently identified and re-identified.

5.5 Experiments on the Kinetic Precision dataset

The system was also evaluated in the Kinetic Tracking Precision (KTP) dataset [2], that was mentioned in the previous chapter, where RGB-D data was recorded using a Microsoft Kinect mounted on top of a mobile robot moving inside an open-space room equipped with a motion capture system. The dataset consists of 4 videos of around 1 minute each. The robot moves differently in each video, to test tracking performance for different robot motion. The videos are named with the movement that the robot performs: *Still*, *Translation*, *Rotation* and *Arc*. In each video, the same five cases occur: *back and forth*, where a single target walks back and forth once, *random*, where three targets walk randomly for about 20 seconds, *side-by-side*, where two targets walk side-by-side in a linear trajectory, *running*, where one person runs across the scene and *group*, where five persons get together in a group and then exit the room.

Along with the presentation of the KTP dataset, Munaro and Menegatti [2] proposed a RGB-D tracking system for service robots, which will be referred to as State of The Art Method (SOAM) for the remainder of this thesis. The results of this thesis's system for the KTP dataset are compared with the results of that method, as reported in their work, in Table 5.6.

Overall the system proposed in this thesis shows a better tracking performance on the KTP dataset than SOAM. The MOTA scores are higher in every situation except one. The ratio of misses and the ratio of false positives are also lower for most of the situations. The number of ID switches is almost the same

Table 5.7: Tracking results of the system for the KTP dataset, divided by video.

	Misses (%)	False Positives (%)	ID Switches	ID Switches (%)	MOTP (m)	MOTA (%)
Still	6.31	2.35	18	2.60	0,340	87.77
Translation	7.91	2.06	10	1.60	0,354	87.20
Rotation	10.42	4.75	20	3.66	0,371	79.26
Arc	10.04	3.06	12	2.60	0,356	82.09
Total	8.46	2.98	60	2.59	0.354	85.98

in every situation for both methods, with a slightly better performance of SOAM if we consider the total amount of ID switches and the fact that in the two most challenging situations (*random walk* and *group*) it produces less ID switches than the system proposed. The MOTP, in meters, is approximately two times lower in every situation for SOAM, which shows a better position estimation of the targets. Nonetheless, the MOTP of the proposed system is always below 40cm, which is still an acceptable result for people tracking, since it still provides a good estimation of the target's position for most tasks. It is important to note that, due to difficulties in synchronizing the ground-truth with the image frames provided in the KTP dataset, the MOTP error of the proposed system is inflated, which explains the bigger error in these experiments when comparing to the experiments presented in the previous section. The ratio of misses has the most impact on the MOTA scores when comparing the two methods. The ratio of misses of the proposed system is much smaller than the one of SOAM in every situation except one. One of the reasons for this is that the people detector used in this work, the Yolov3, has a much better performance than the people detector used in SOAM, which is a HOG detector. The number of ID switches shows that the proposed system struggles more when there is a group of people present in the scene, such as in the $random\ walk$ and $group$ situations. The KTP dataset also features 5 different people, while the Re-ID Multi-Tracking Dataset only features 3 different targets, which also increases the difficulty in keeping track's identities.

Tracking results on the KTP dataset, divided by video, are also reported in Table 5.7. In [2], 3D tracking results divided by video are not reported, so it is not possible to compare the two methods in this case.

The results show that the system has a good tracking performance in all of the videos, producing a low ratio of misses, false positives and ID switches. The MOTA scores are higher in the videos where there is less camera and robot motion, because camera movement introduces more errors in the position estimation and when targets exit the field of view, their re-identification is more challenging.

The experimental results in the KTP dataset show that the proposed system is robust to different scenarios and situations and, compared with SOAM, it achieves an overall better tracking performance.

5.6 Discussion

Several experiments were conducted by running the system on the MBOT using two different datasets and a test sequence. Evaluation results were obtained using standard tracking evaluation metrics, the CLEAR MOT metrics, with the addition of a count of the recall errors produced by the system.

The system's parameters were fine-tuned through experiments on the Re-ID Multi-Tracking Dataset. First, in these experiments it became clear that the value of the parameters T_{lower}, T_{upper}, S_{max} and $T_{recover}$ does not have a big impact on the MOTP and MOTA scores. This thresholds impact mostly the number of ID switches and recall errors. The number of misses and false positives is determined mainly by the performance of the people detector module which is constant in every experiment. Since the people detector detects very accurately the people present in the scene, the number of misses is never high and is only caused by the track initialization delay of 3 frames, that is needed to prevent false positives. This trade-off results in a very low number of false positives, which in these experiments can be due to ground-truth errors that could not be completly eliminated.

Secondly, the experimental results when varying S_{max} and pairs of (T_{lower}, T_{upper}) showed that a combination of the spatial and the appearance distance metrics is key for achieving a better re-identification performance. For the values of these parameters where the spatial or the appearance metric have much more impact in the data association step, the number of ID switches increases, due to either appearance similarities between people or unexpected movement from a target which increases the spatial distance to a point where the association is not performed. Combining both metrics trough the optimal values of the parameters leads to an optimal performance where both spatial and appearance considerations are taken in the data association step.

Thirdly, the value of $T_{recover}$ has a great impact on the number of recall errors, since a target's appearance can change a lot during the system's execution, due to illumination changes for instance, but two different people can also have small appearance distances between each other. Hence, choosing the right value implies a trade-off between correctly re-identifying targets and reducing ID switches, while keeping recall errors low.

The system's performance on the Re-ID Multi-Tracking Dataset shows robust target tracking and identity assignment with precise position estimation, achieving a MOTA score of 88.50% and MOTP of 0.215 meters. The system was also tested in an unseen test sequence, in which the performance was also very good, achieving a MOTA score of 88.50% and MOTP of 0.190 meters. Even in the sequence were targets changed clothing, the ID switches ratio did not increased to values above 4%. These results show that the proposed system is robust at tracking and re-identifying people in an environment with multiple targets, obstacles and frequent occlusions. Considering the dataset was recorded using a moving camera mounted on a mobile robot and that the system achieved a frame rate of 33Hz, this shows that the system can be applied for mobile robotics with good performance.

Finally, the system was also evaluated on an open-space dataset with up to 5 different targets in the scene, the KTP dataset, and was compared against a state-of-the-art method, proposed with that same dataset. The system achieved an overall MOTA score of 85.98% and MOTP of 0,354m on this dataset and produced overall better results than SOAM, the method proposed in [2].

The results show that the proposed system is robust at multi-target tracking and re-identification in an indoor environment with challenging scenarios such as occlusions and obstacles. The system is also suitable for robotic applications, considering the real-time performance of the method.

6

Conclusion

Contents

6.1 Conclusions

Service robots are designed to interact with humans which means they should have a good social behaviour. In order to improve human-robot interaction, the robot should be able to locate and differentiate the different people in the environment, which allows for personalized interactions. For this purpose, this thesis aimed at developing a 3D position tracking system to be deployed on a mobile robot that robustly recognized and identified multiple people in the scene, besides re-identifying targets previously seen, while maintaining real-time performance, on an environment with obstacles and occlusions caused by other people. The results show that the proposed Re-ID based multi-people tracker achieves these goals and even outperforms another state-of-the-art method.

Multiple limitations in existing Re-ID and tracking methods and datasets were pointed out in this thesis. Most Re-ID and tracking methods are not suitable for robotic applications since they are computationally too demanding or they are based on underlying assumptions that not hold for real-world scenarios such as a perfect gallery of targets or constrained movement of the robot and people in the scene. This work combines existing methods suitable for robotic applications such as a people detector, a people localizer, a Re-ID feature extractor and a Kalman filter framework with simple data association and track management approaches that result in a lightweight and robust Re-ID based multi-people tracker suitable for real-world scenarios and applications.

Regarding the existing datasets, the great majority of the tracking datasets are aimed at 2D tracking only and the 3D tracking datasets are very limited in terms of camera movement and environments. If we consider datasets that use depth information, it was found only one dataset available recorded with a moving camera. To overcome this lack of datasets, a novel RGB-D Re-ID multi-people tracking dataset recorded with a moving camera mounted on top of a mobile robot was constructed. This dataset is representative of real-world scenarios, it was recorded on an apartment-like environment with obstacles and features up to three different targets that suffer several occlusions and appearance changes. The dataset includes target's 3D position and identities ground-truth. This dataset has characteristics that are new and fullfills the lack of a moving camera RGB-D 3D tracking dataset in an environment with obstacles.

An experimental analysis was conducted and evaluation of the method was performed in the proposed dataset, as well as in a test sequence and in a state-of-the-art dataset. The system achieved a MOTA score above 85% in all of them and a MOTP always below 0.4m. The proposed method is robust to appearance changes, such as clothing, pose and illumination changes and occlusions. The proposed method runs at 33Hz, which is suitable for real-time robotic applications.

To conclude, this thesis presents a robust 3D tracking method focused on people re-identification that can be used for robotic applications. This method was designed for the MBOT, but can be easily deployed in other mobile robots to accomplish tasks that rely on people re-identification and tracking,

improving human-robot interaction and allowing for personalized robot behaviour, which is key for a better acceptance of robots in domestic environments. The results and findings of this thesis can be a contribution for further developments of Re-ID based tracking systems suitable for mobile robots.

6.2 System Limitations and Future Work

Although the experimental results show that the proposed system is robust to various scenarios and achieves a very good tracking performance, there are still improvements that can be made.

There are still some cases where the system produces ID switches, which introduce error in the perception of the people present in the scene. The ideal case would be for the system to produce zero ID switches and there are several research paths that can be taken to try to accomplish this. First, the data association step could be improved. Currently, the data association relies on the fine-tuning of the system's thresholds which are rigid and will always discard correct associations due to appearance changes that lead to distances slightly above the threshold. A better performance could be achieved by using a probabilistic model to model the combined appearance and spatial distance between detections and existing tracks. This approach could better model the changes in appearance and spatial distance, resulting in a combined probabilistic metric that could be used to assign detections to tracks.

Secondly, the track management step can also be changed to reduce recall errors, which will also reduce ID switches. Instead of deleting tracks when the targets are not being seen, those tracks could be kept with an *Inactive* track indicator, described by a random walk motion model, for instance. Following this approach, when deciding whether to assign a previously seen ID to a new track or not, the spatial distance could also be taken into account. Other option would be, besides the random walk motion model, keep the position where the track was deleted and check if the new track is being initialized in that position. This option would help in recalling a person that may had not moved since the track was lost. Combining these two options in a probabilistic way, e.g. by assuming equal probability to both cases, could further improve this approach.

Besides these improvements in the algorithm, more experiments could also be conducted in order to access and improve the system's performance. The experiments could be evaluated using different metric, such as metrics more focused on the re-identification performance only. The proposed dataset only features a maximum of three different people at the same time in the scene, which is not much. Although in a domestic environment usually there are not more than three people at the same time, an addition to the dataset could be made with more sequences including larger groups of people, considering that the complexitiy of the people re-identification task increases with the number of people present. Additionally, more challenging scenarios could be considered, such as cases where people assume more poses such as lying down, bending down or while doing sports. It could also be interesting to test

the re-identification performance of the method in a case where every target is using the same clothes, to access how that scenario would affect the appearance distances between them. It could also be interesting to test the system in an outdoor environment, where the depth camera can be less accurate, or in sequences where the light intensity changes abruptly, to evaluate how that impacts the performance of the system.

6.3 Ethical Considerations

As a final note, it is important to mention some ethical considerations about this work. With the development of every new technology there should always be a discussion regarding its impact in our society and on human rights. Technology brings unmeasurable benefits to humans and helps us in every aspect of our lives nowadays, however, it can also have a negative effect in the way we interact with others, in our privacy and in issues such as discrimination, inequality and environmental impact. The recent developments in the field of Artificial Intelligence (AI) have been receiving attention from an ethical and legal perspective which lead to the publication of the proposal of the new EU Artificial Intelligence Act (AIA) by the European Commission [126]. This proposal shows the importance of considering ethics, legality, equality and environmental sustainability in applications and research that use AI.

People re-identification is a specially concerning topic since its most widely application is surveillance. It is even more concerning since, to the extent of my knowledge, none of the papers referred to in this thesis related to people re-identification mentioned any ethical concerns. People re-identification provides machines with a skill that can be easily used for unethical purposes such as authoritarian surveillance, ethical groups discrimination and invasion of privacy. Another concern is the construction of image and video datasets without the consent of the people that were recorded. One of the largest Re-ID datasets, DUKE MTMC Re-ID, is an extension of the DUKE MTMC dataset [51], which has now been terminated following a report from Exposing.ai[1] and an investigation from the Financial Times[2] which reported its use for discriminative surveillance applications and raised concerns regarding the image rights of the people present in the dataset, which did not explicitly authorized the recording. Despite the shutdown of the main dataset, extensions such as the ones used for Re-ID are still available and continue to be used by many researchers.

Following this remarks, the dataset constructed in this thesis was recorded with the consent of all the participants, which signed an informed consent that can be seen in Appendix B. The participants also agreed with the public distribution of the dataset for research purposes.

In conclusion, it is important to state that the dataset and the work presented in this thesis aim at improving human-robot interaction and providing service robots with the ability to conduct personalized

[1] https://exposing.ai/duke_mtmc/
[2] https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e

interactions that benefit human's well-being and helps them in their daily activities. This work should not be used in any circunstances for surveillance applications or discriminative purposes.

Bibliography

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

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

[3] R. D. Schraft and G. Schmierer, *Service robots*. CRC Press, 2000.

[4] K. Kawamura, R. T. Pack, M. Bishay, and M. Iskarous, "Design philosophy for service robots," *Robotics and Autonomous Systems*, vol. 18, no. 1-2, pp. 109–116, 1996.

[5] J. E. Young, R. Hawkins, E. Sharlin, and T. Igarashi, "Toward acceptable domestic robots: Applying insights from social psychology," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 95–108, 2009.

[6] P. U. Lima, C. Azevedo, E. Brzozowska, J. Cartucho, T. J. Dias, J. Gonçalves, M. Kinarullathil, G. Lawless, O. Lima, R. Luz *et al.*, "Socrob@ home," *KI-Künstliche Intelligenz*, vol. 33, no. 4, pp. 343–356, 2019.

[7] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human factors*, vol. 53, no. 5, pp. 517–527, 2011.

[8] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 1092–1108, 2019.

[9] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European conference on computer vision*. Springer, 2008, pp. 262–275.

[10] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 2360–2367.

[11] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *European conference on computer vision*. Springer, 2014, pp. 536–551.

[12] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2197–2206.

[13] M. Pietikäinen, "Local binary patterns," *Scholarpedia*, vol. 5, no. 3, p. 9775, 2010.

[14] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1363–1372.

[15] J. L. Suárez-Díaz, S. García, and F. Herrera, "A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges (with appendices on mathematical background and detailed algorithms explanation)," *arXiv preprint arXiv:1812.05944*, 2018.

[16] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *CVPR 2011*. IEEE, 2011, pp. 649–656.

[17] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *European conference on computer vision*. Springer, 2012, pp. 780–793.

[18] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *European conference on computer vision*. Springer, 2014, pp. 1–16.

[19] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[20] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 34–39.

[21] L. Wu, C. Shen, and A. Van Den Hengel, "Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification," *Pattern Recognition*, vol. 65, pp. 238–250, 2017.

[22] J. Zhu, H. Zeng, S. Liao, Z. Lei, C. Cai, and L. Zheng, "Deep hybrid similarity learning for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3183–3193, 2017.

[23] J. Almazan, B. Gajic, N. Murray, and D. Larlus, "Re-id done right: towards good practices for person re-identification," *arXiv preprint arXiv:1801.05339*, 2018.

[24] A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” *arXiv preprint arXiv:1703.07737*, 2017.

[25] C. Han, J. Ye, Y. Zhong, X. Tan, C. Zhang, C. Gao, and N. Sang, “Re-id driven localization refinement for person search,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9814–9823.

[26] W. Li, X. Zhu, and S. Gong, “Harmonious attention network for person re-identification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2285–2294.

[27] G. Zhang, J. Yang, Y. Zheng, Y. Wang, Y. Wu, and S. Chen, “Hybrid-attention guided network with multiple resolution features for person re-identification,” *Information Sciences*, 2021.

[28] J. Sun, Y. Li, H. Chen, B. Zhang, and J. Zhu, “Memf: Multi-level-attention embedding and multi-layer-feature fusion model for person re-identification,” *Pattern Recognition*, vol. 116, p. 107937, 2021.

[29] J. Wu, S. Liao, X. Wang, Y. Yang, S. Z. Li *et al.*, “Clustering and dynamic sampling based unsupervised domain adaptation for person re-identification,” in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 886–891.

[30] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, “Unsupervised domain adaptive re-identification: Theory and practice,” *Pattern Recognition*, vol. 102, p. 107173, 2020.

[31] Q. Li, X. Peng, Y. Qiao, and Q. Hao, “Unsupervised person re-identification with multi-label learning guided self-paced clustering,” *arXiv preprint arXiv:2103.04580*, 2021.

[32] S. Cosar, C. Coppola, N. Bellotto *et al.*, “Volume-based human re-identification with rgb-d cameras.” in *VISIGRAPP (4: VISAPP)*, 2017, pp. 389–397.

[33] M. Munaro, A. Basso, A. Fossati, L. Van Gool, and E. Menegatti, “3d reconstruction of freely moving persons for re-identification with a depth sensor,” in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 4512–4519.

[34] A. Haque, A. Alahi, and L. Fei-Fei, “Recurrent attention models for depth-based person identification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1229–1238.

[35] F. Han, B. Reily, W. Hoff, and H. Zhang, “Space-time representation of people based on 3d skeletal data: A review,” *Computer Vision and Image Understanding*, vol. 158, pp. 85–105, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314217300279

[36] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR 2011*. Ieee, 2011, pp. 1297–1304.

[37] M. Munaro, S. Ghidoni, D. T. Dizmen, and E. Menegatti, "A feature-based approach to people re-identification using skeleton keypoints," in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 5644–5651.

[38] S. Ghidoni and M. Munaro, "A multi-viewpoint feature-based re-identification system driven by skeleton keypoints," *Robotics and Autonomous Systems*, vol. 90, pp. 45–54, 2017.

[39] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "Rgb-infrared cross-modality person re-identification," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5380–5389.

[40] Y. Hao, N. Wang, J. Li, and X. Gao, "Hsme: hypersphere manifold embedding for visible thermal person re-identification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8385–8392.

[41] M. Ye, X. Lan, J. Li, and P. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[42] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1970–1979.

[43] L. An, X. Chen, and S. Yang, "Multi-graph feature level fusion for person re-identification," *Neurocomputing*, vol. 259, pp. 39–45, 2017.

[44] Z. Wang, Q. Han, X. Niu, and C. Busch, "Feature-level fusion of iris and face for personal identification," in *International Symposium on Neural Networks*. Springer, 2009, pp. 356–364.

[45] M. Eisenbach, A. Kolarow, A. Vorndran, J. Niebling, and H.-M. Gross, "Evaluation of multi feature fusion at score-level for appearance-based person re-identification," in *2015 international joint conference on neural networks (IJCNN)*. IEEE, 2015, pp. 1–8.

[46] R. Kawai, Y. Makihara, C. Hua, H. Iwama, and Y. Yagi, "Person re-identification using view-dependent score-level fusion of gait and color features," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012, pp. 2694–2697.

[47] S. karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 523–536, 2019.

[48] C. C. Loy, C. Liu, and S. Gong, "Person re-identification by manifold ranking," in *2013 IEEE International Conference on Image Processing*. IEEE, 2013, pp. 3567–3571.

[49] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 152–159.

[50] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.

[51] 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 workshop on Benchmarking Multi-Target Tracking*, 2016.

[52] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 79–88.

[53] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian conference on Image analysis*. Springer, 2011, pp. 91–102.

[54] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *European Conference on Computer Vision*. Springer, 2016, pp. 868–884.

[55] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5177–5186.

[56] M. Li, X. Zhu, and S. Gong, "Unsupervised person re-identification by deep learning tracklet association," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 737–753.

[57] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai, "Region-based quality estimation network for large-scale person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[58] J. Li, J. Wang, Q. Tian, W. Gao, and S. Zhang, “Global-local temporal representations for video person re-identification,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3958–3967.

[59] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, “Visual tracking: An experimental survey,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1442–1468, 2013.

[60] A. Bedagkar-Gala and S. K. Shah, “A survey of approaches and trends in person re-identification,” *Image and vision computing*, vol. 32, no. 4, pp. 270–286, 2014.

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

[62] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[63] C. Weinrich, C. Vollmer, and H.-M. Gross, “Estimation of human upper body orientation for mobile robotics using an svm decision tree on monocular images,” in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 2147–2152.

[64] P. Felzenszwalb, D. McAllester, and D. Ramanan, “A discriminatively trained, multiscale, deformable part model,” in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[65] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.

[66] J. Yan, Z. Lei, L. Wen, and S. Z. Li, “The fastest deformable part model for object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2497–2504.

[67] L. Bourdev and J. Malik, “Poselets: Body part detectors trained using 3d human pose annotations,” in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 1365–1372.

[68] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results,” http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[69] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[70] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[71] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, 2016, pp. 379–387.

[72] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[73] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[74] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7297–7306.

[75] K. O. Arras, O. M. Mozos, and W. Burgard, "Using boosted features for the detection of people in 2d range data," in *Proceedings 2007 IEEE international conference on robotics and automation*. IEEE, 2007, pp. 3402–3407.

[76] K. Schenk, M. Eisenbach, A. Kolarow, and H.-M. Gross, "Comparison of laser-based person tracking at feet and upper-body height," in *Annual Conference on Artificial Intelligence*. Springer, 2011, pp. 277–288.

[77] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1442–1468, 2014.

[78] G. Welch, G. Bishop *et al.*, "An introduction to the kalman filter," 1995.

[79] N. J. Gordon, D. J. Salmond, and A. F. Smith, "Novel approach to nonlinear/non-gaussian bayesian state estimation," in *IEE Proceedings F-radar and signal processing*, vol. 140, no. 2. IET, 1993, pp. 107–113.

[80] B. D. Anderson and J. B. Moore, *Optimal filtering*. Courier Corporation, 2012.

[81] S. J. Julier and J. K. Uhlmann, “New extension of the kalman filter to nonlinear systems,” in *Signal processing, sensor fusion, and target recognition VI*, vol. 3068. International Society for Optics and Photonics, 1997, pp. 182–193.

[82] R. Van Der Merwe, A. Doucet, N. De Freitas, and E. Wan, “The unscented particle filter,” *Advances in neural information processing systems*, vol. 13, pp. 584–590, 2000.

[83] M. Volkhardt, C. Weinrich, and H.-M. Gross, “People tracking on a mobile companion robot,” in *2013 IEEE International Conference on Systems, Man, and Cybernetics*, 2013, pp. 4354–4359.

[84] C. Li, L. Guo, and Y. Hu, “A new method combining hog and kalman filter for video-based human detection and tracking,” in *2010 3rd International Congress on Image and Signal Processing*, vol. 1, 2010, pp. 290–293.

[85] S. G. Konrad and F. R. Masson, “Pedestrian skeleton tracking using openpose and probabilistic filtering,” in *2020 IEEE Congreso Bienal de Argentina (ARGENCON)*, 2020, pp. 1–7.

[86] C. Yang, R. Duraiswami, and L. Davis, “Fast multiple object tracking via a hierarchical particle filter,” in *Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1*, vol. 1, 2005, pp. 212–219 Vol. 1.

[87] J. Messias, J. J. Acevedo, J. Capitan, L. Merino, R. Ventura, and P. U. Lima, “A particle-filter approach for active perception in networked robot systems,” in *International Conference on Social Robotics*. Springer, 2015, pp. 451–460.

[88] J. J. Acevedo, J. Messias, J. Capitán, R. Ventura, L. Merino, and P. U. Lima, “A dynamic weighted area assignment based on a particle filter for active cooperative perception,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 736–743, 2020.

[89] T. Zhang, S. Liu, C. Xu, B. Liu, and M.-H. Yang, “Correlation particle filter for visual tracking,” *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2676–2687, 2018.

[90] L. Bazzani, D. Bloisi, and V. Murino, “A comparison of multi hypothesis kalman filter and particle filter for multi-target tracking,” in *Performance Evaluation of Tracking and Surveillance workshop at CVPR*, 2009, pp. 47–54.

[91] N. Bellotto and H. Hu, “People tracking with a mobile robot: A comparison of kalman and particle filters,” in *Proc. of the 13th IASTED Int. Conf. on Robotics and Applications*, 2007, pp. 388–393.

[92] J.-W. Choi, D. Moon, and J.-H. Yoo, “Robust multi-person tracking for real-time intelligent video surveillance,” *ETRI Journal*, vol. 37, no. 3, pp. 551–561, 2015.

[93] A. Corominas-Murtra, J. Pagès, and S. Pfeiffer, “Multi-target amp; multi-detector people tracker for mobile robots,” in *2015 European Conference on Mobile Robots (ECMR)*, 2015, pp. 1–6.

[94] B. X. Chen, R. Sahdev, and J. K. Tsotsos, “Person following robot using selected online ada-boosting with stereo camera,” in *2017 14th conference on computer and robot vision (CRV)*. IEEE, 2017, pp. 48–55.

[95] C. Kim, F. Li, and J. M. Rehg, “Multi-object tracking with neural gating using bilinear lstm,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 200–215.

[96] L. Leal-Taixé, A. Milan, K. Schindler, D. Cremers, I. Reid, and S. Roth, “Tracking the trackers: an analysis of the state of the art in multiple object tracking,” *arXiv preprint arXiv:1704.02781*, 2017.

[97] L. Chen, H. Ai, Z. Zhuang, and C. Shang, “Real-time multiple people tracking with deeply learned candidate selection and person re-identification,” in *2018 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2018, pp. 1–6.

[98] L. Beyer, S. Breuers, V. Kurin, and B. Leibe, “Towards a principled integration of multi-camera re-identification and tracking through optimal bayes filters,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 29–38.

[99] L. Lan, X. Wang, G. Hua, T. S. Huang, and D. Tao, “Semi-online multi-people tracking by re-identification.” *International Journal of Computer Vision*, vol. 128, no. 7, 2020.

[100] A. T.-Y. Chen, M. Biglari-Abhari, and K. I.-K. Wang, “Fusing appearance and spatio-temporal models for person re-identification and tracking,” *Journal of Imaging*, vol. 6, no. 5, p. 27, 2020.

[101] K. Koide and J. Miura, “Identification of a specific person using color, height, and gait features for a person following robot,” *Robotics and Autonomous Systems*, vol. 84, pp. 76–87, 2016.

[102] C. Weinrich, M. Volkhardt, and H.-M. Gross, “Appearance-based 3d upper-body pose estimation and person re-identification on mobile robots,” in *2013 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2013, pp. 4384–4390.

[103] T. Wengefeld, M. Eisenbach, T. Q. Trinh, and H.-M. Gross, “May i be your personal coach? bringing together person tracking and visual re-identification on a mobile robot,” in *Proceedings of ISR 2016: 47st International Symposium on Robotics*. VDE, 2016, pp. 1–8.

[104] M. Eisenbach, A. Kolarow, A. Vorndran, J. Niebling, and H.-M. Gross, “Evaluation of multi feature fusion at score-level for appearance-based person re-identification,” in *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–8.

[105] Y. Murata and M. Atsumi, "Person re-identification for mobile robot using online transfer learning," in *2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS)*. IEEE, 2018, pp. 977–981.

[106] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[107] P. A. Carlsen, "Real-time person re-identification for mobile robots to improve human-robot interaction," Master's thesis, University of Oslo, 2019.

[108] L. Pang, Z. Cao, J. Yu, P. Guan, X. Rong, and H. Chai, "A visual leader-following approach with a t-d-r framework for quadruped robots," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 4, pp. 2342–2354, 2021.

[109] H. Liu, J. Luo, P. Wu, S. Xie, and H. Li, "People detection and tracking using rgb-d cameras for mobile robots," *International Journal of Advanced Robotic Systems*, vol. 13, no. 5, p. 1729881416657746, 2016.

[110] A. Farhadi and J. Redmon, "Yolov3: An incremental improvement," in *Computer Vision and Pattern Recognition*. Springer Berlin/Heidelberg, Germany, 2018, pp. 1804–02.

[111] 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)*. IEEE, 2017, pp. 3645–3649.

[112] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the society for industrial and applied mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[113] "Mot challenge," https://motchallenge.net, accessed: 2021-09-16.

[114] G. Ciaparrone, F. Luque Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep learning in video multi-object tracking: A survey," *Neurocomputing*, vol. 381, pp. 61–88, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219315966

[115] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," *arXiv:1504.01942 [cs]*, Apr. 2015, arXiv: 1504.01942. [Online]. Available: http://arxiv.org/abs/1504.01942

[116] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," *arXiv:1603.00831 [cs]*, Mar. 2016, arXiv: 1603.00831. [Online]. Available: http://arxiv.org/abs/1603.00831

[117] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé, "CVPR19 tracking and detection challenge: How crowded can it get?" *arXiv:1906.04567 [cs]*, Jun. 2019, arXiv: 1906.04567. [Online]. Available: http://arxiv.org/abs/1906.04567

[118] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013. [Online]. Available: https://doi.org/10.1177/0278364913491297

[119] M. Camplani, A. Paiement, M. Mirmehdi, D. Damen, S. Hannuna, T. Burghardt, and L. Tao, "Multiple human tracking in rgb-d data: A survey," *IET Computer Vision*, vol. 11, 06 2016.

[120] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "A mobile vision system for robust multi-person tracking," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[121] L. Spinello and K. O. Arras, "People detection in rgb-d data," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2011, pp. 3838–3843.

[122] W. Choi, C. Pantofaru, and S. Savarese, "A general framework for tracking multiple people from a moving camera," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 7, pp. 1577–1591, 2012.

[123] E. J. Almazán and G. A. Jones, "A depth-based polar coordinate system for people segmentation and tracking with multiple rgb-d sensors," *ISMAR 2014 Workshop on Tracking Methods and Applications*, 2014.

[124] J. Liu, Y. Liu, G. Zhang, P. Zhu, and Y. Q. Chen, "Detecting and tracking people in real time with rgb-d camera," *Pattern Recognition Letters*, vol. 53, pp. 16–23, 2015.

[125] 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.

[126] E. COMMISION, "Regulation of the european parliament and of the council," *EUROPIAN COMMISION, Brussels*, 2018.

7

Appendix A - Experiment's people trajectories and tracks

This appendix contains examples of people tracks generated by the proposed system compared with ground-truth trajectories, for each sequence of the Re-ID multi-people tracking dataset. In each figure, ground-truth trajectories are represented by circles and the system's output tracks are represented by squares. Different colors in ground-truth trajectories represent different people and different colors in the tracks represent different ID's assigned to the track. The examples shown here represent trajectories and tracks belonging to only a portion of the full sequences, for better visualization.

Figure 7.1: ***Still* sequence trajectories** In this figure we can see the trajectories of two targets, represented by the colors blue and black. The two tracks generated by the system show a good estimation of the target's position and there are no ID switches.

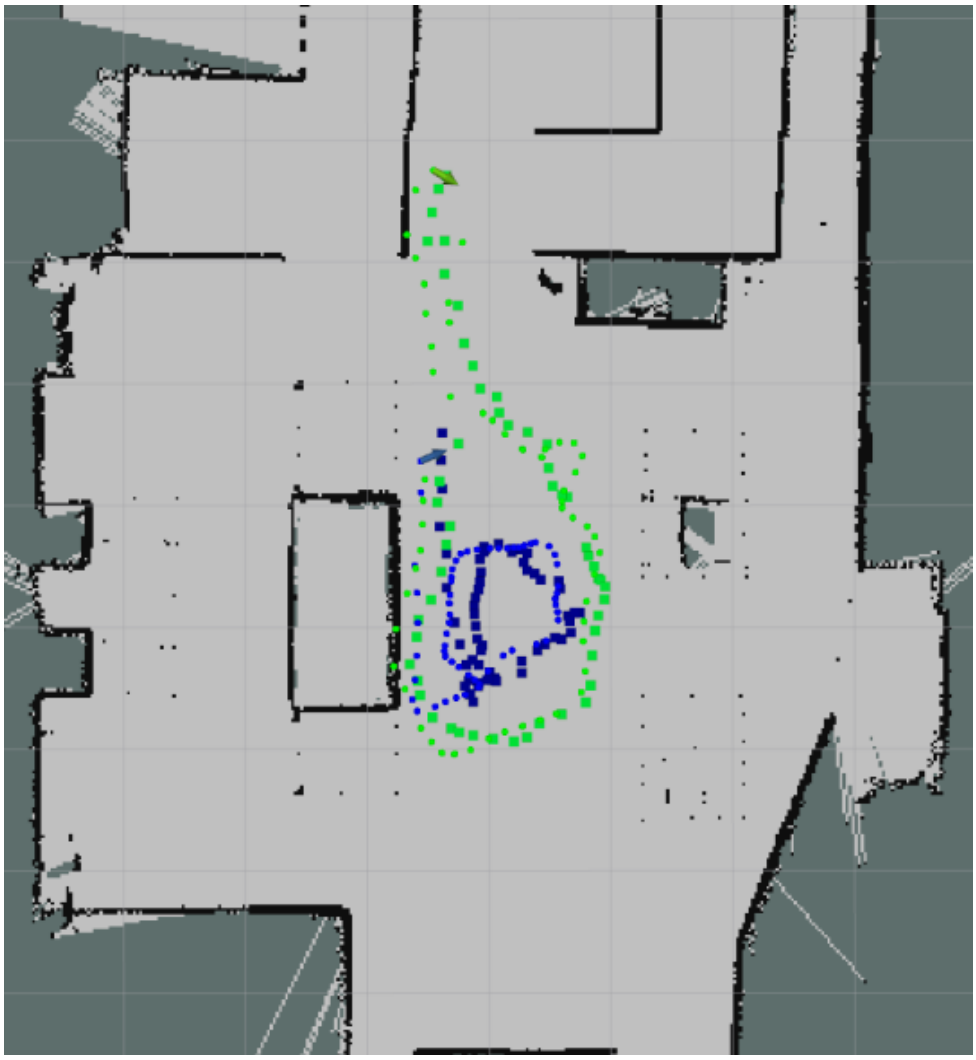


Figure 7.2: ***Moving head*** **sequence trajectories** In this figure we can see the trajectories of two targets, represented by the colors green and blue. The two tracks generated by the system show a good estimation of the target's position on the most part of the trajectories and there are no ID switches. We can see that the target represented by the color green goes around the other target, occluding it several times, and the system is still able to keep the track's IDs.

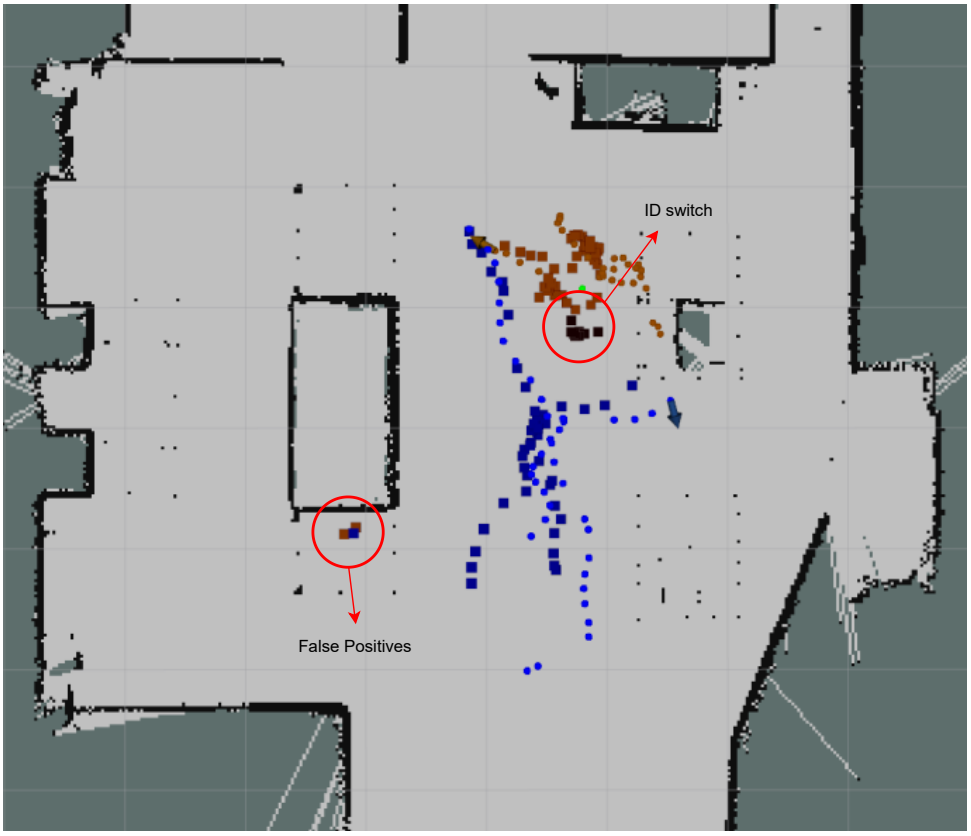
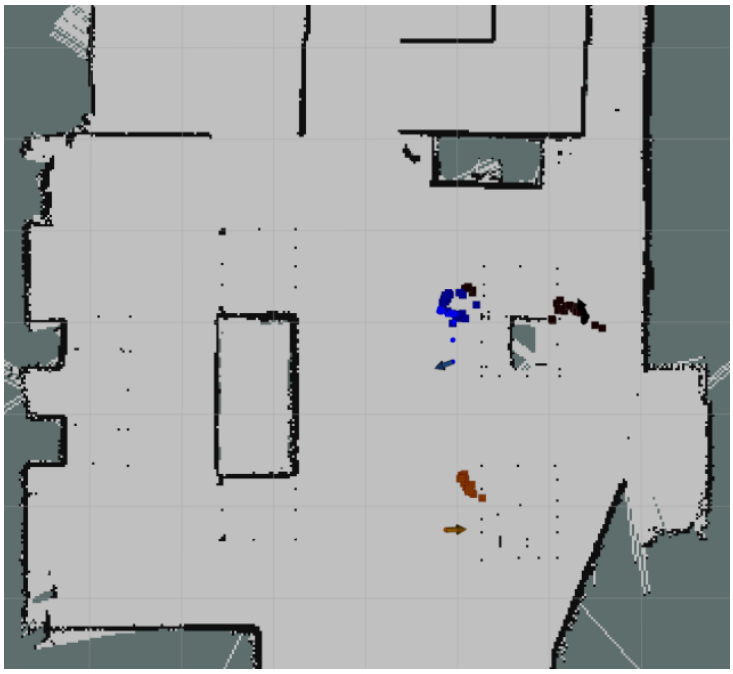
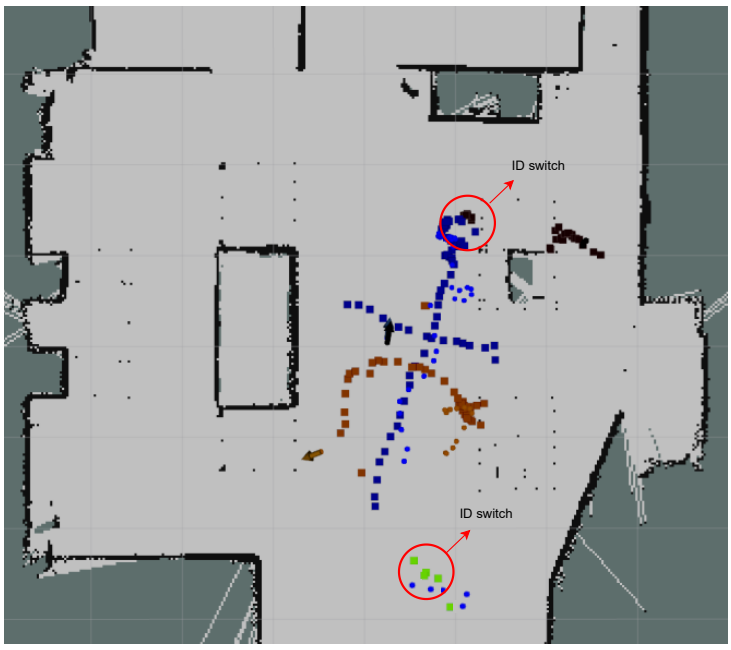


Figure 7.3: ***Moving base*** **sequence trajectories** In this figure we can see the trajectories of two targets, represented by the colors brown and blue. We can also see an ID switch, where the brown ground-truth trajectory is temporarily assigned a different ID (represented in black). We can also see some false positives generated by the system.



(a)



(b)

Figure 7.4: *Chairs* sequence trajectories In this figure we can see the trajectories of three targets, represented by the colors brown, black and blue. In (a) we can see the trajectories of the three targets while they are sitting down. In (b), two of the targets get up and walk around the environment. We can see two ID switches: one where the blue target is assigned the ID of the target that is sitting close to him (represented in black) and another one where the blue target is assigned a different ID (represented in green)

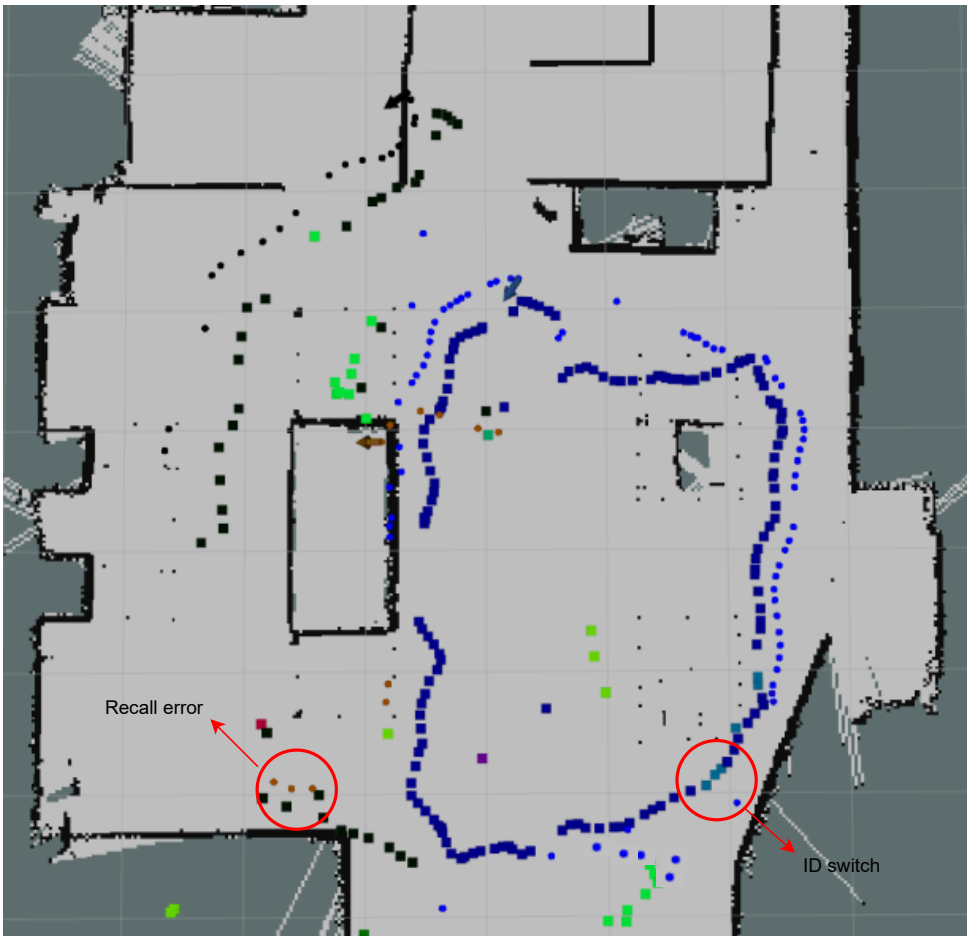


Figure 7.5: ***People following*** **sequence trajectories** In this figure we can see the trajectories of three targets, represented by the colors brown, black and blue. In this figure the main trajectory is the blue one, which belongs to the person being followed by the robot, and the other trajectories are temporary occlusions to that person. We can see an ID switch in the blue trajectory, where the ID assigned to that track switches momentarily. We can also see a recall error, where the brown trajectory is assigned an ID, black, that was previously assigned to the trajectory of other target, i.e. the black one.

Figure 7.6: ***Changing clothes 1*** **sequence trajectories** In this figure we can see the trajectories of two targets, represented by the colors black and blue. The target represented by the color black is tracked without any errors. The other target is first assigned an ID, represented by the color blue but when it re-enters the scene, a different ID, green, is assigned, which shows an ID switch.

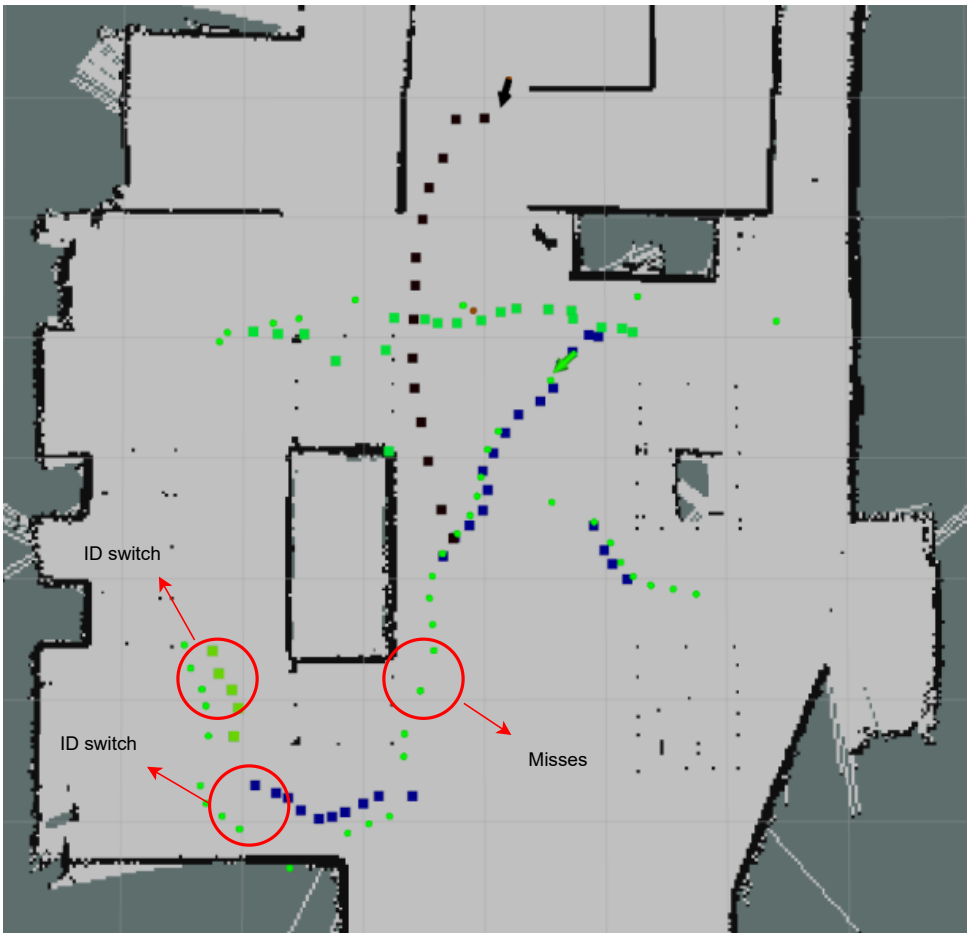


Figure 7.7: ***Changing clothes 2*** **sequence trajectories** In this figure we can see the trajectories of two targets, represented by the colors black and green. There is no ground-truth for the target represented by the black track, due to a motion capture system failure. The other target is first assigned an ID, represented by the color gree, and then continues to be tracked although it suffers two ID switches. There is also a part of the trajectory that is not tracked, which leads to misses.

8

Appendix B - Dataset recording informed consents

In this appendix the informed consent that was signed by the three participants in the videos recorded for the Re-ID multi-people tracking dataset is presented.

Institute for Systems and Robotics Lisboa

TÉCNICO LISBOA

Informed Consent

- ❑ I read and understood the procedures, duration and place of the user study.
- ❑ I read and understood what data can be recorded (robot: telemetry; participant: video and 3D position in space) for the exclusive purpose of scientific research. I authorize that the recorded data during the experimental sessions include:
  - ❑ video recording
  - ❑ 3D position recording
- ❑ I authorize the anonymous treatment of data collected under this project for the purpose of analysis, research and dissemination of results in magazines or conferences, by the researchers of the project.
- ❑ I authorize the public release of this data for access by others for research purposes, with the possibility of requesting the removal of the data from public access at any time.
- ❑ I understood that my participation in this user study does not pose any risk, discomfort or disadvantage to myself.
- ❑ I understood that my participation in this study is voluntary and that I can withdraw at any time without giving an explanation. If this happens, no penalty will occur and my data will be removed and destroyed.

I accept the crossed terms of this consent,

Name ..... Date .....

Signature .....

.....

For more information, please contact:
Vicente Pinto (Responsible for the user study and data protection)
vicenteppinto@tecnico.ulisboa.pt
Rui Bettencourt
rui.bettencourt@tecnico.ulisboa.pt
Professor Rodrigo Ventura
rodrigo.ventura@isr.tecnico.ulisboa.pt
Instituto Superior Técnico, Universidade de Lisboa & ISR

One copy of the consent is for the investigators and, if requested, another one is for the participant.

This Informed consent was written according to the guidelines of the Ethics Committee of Instituto Superior Técnico

Figure 8.1: Informed consent that was signed by the participants