People Recognition and Identification in Service Robots

Vicente Palma Figueira Castro Pinto
Instituto Superior Técnico
Lisboa, Portugal
vicenteppinto@tecnico.ulisboa.pt

***Abstract*—Service robots provide services to humans such as helping humans in their 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 too closely related tasks. However, existing Re-ID 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.**

***Index Terms*—human-robot interatction, people re-identification, people tracking, multiple kalman-filter, RGB-D dataset**

I. Introduction

Service robots have received increased attention in recent years, covering a great variety of applications and system designs. These robots can be extremely helpful since they can replace humans in hazardous situations, help physically handicapped people and serve as a companion to elderly people or children [1]. 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.The MOnarCH robot (MBOT) is a service robot originally designed to interact with children in hospitals. The MBOT was adapted for robotic competitions in domestic scenarios by SocRob@Home [2]. In order to accomplish a good social behaviour, the robot should be able to recognize, identify and re-identify humans, that is, determine if a certain person is present in a set of candidates and recall that person's identity through time. This skill enables personalized interaction between the robot and the people in his surroundings. These interactions build up the robot's personality and adaptability, which are key factors for increasing trust in the robot [3]. Re-identification 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. The scientific contributions of this work are three-fold: (1) 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, (2) 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, (3) an experimental evaluation of the method proposed in a real-word dataset.

II. Background

A. Computer Vision Person Re-Identification

In the context of computer vision 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) [4]. 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. 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. Some Person Re-ID methods make use of hand-crafted techniques such as different histograms and segmentation techniques to construct the appearance descriptor for each person [5], [6]. 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. 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 [7], [8]. 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.

B. Related Work

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 [9]. Besides tracking their positions, knowing their identities and being able to differentiate between different individuals is very valuable. Hence, people Re-ID methods are used to assign unique ID's to the targets being tracked. 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 [10].

Wengefeld et al. [11] shows 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.

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 them uses online transfer learning [12], using three CNN's: one for person detection, one for person feature extraction and one for person re-identification. Carslen also proposed two new CNN's called LuNet Light and LuNet Lightest with the purpose of implementing Re-ID in mobile robots [13]. The resulting models achieve close to state-of-the-art performance, while being much lighter than others, 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 [14]. 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 RGB-D cameras for mobile robots, that also re-identifies targets through association [15]. 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. In [16], 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 GPU.

C. Critical Discussion

The methods described above 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, specially in the data association step, increasing robustness in cases where targets walk out of the scene and re-enter it, crowded environments and noise, while on the other hand, tracking can handle better cases where the target's appearance changes [17], [18]. They show 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 assume constrained conditions on the environment and the movement of people 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.

III. Methodology

A. Coordinate frames

Before presenting the overall system architecture, it is important to define the relevant coordinate frames of our problem. 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 changes based on the odometry errors. 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).

B. System architecture and components

An overview of the system architecture is presented in figure 1. The system receives as input a RGB and a depth image and is composed of a people detector, a people localizer, a Re-ID feature generator and a Multi-people tracker. The output of the system is a set of ID-assigned tracks, that correspond to the people in the scene.

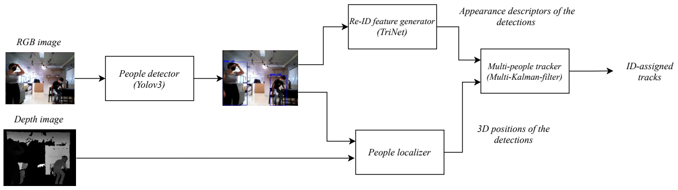

Fig. 1: System architecture overview.

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. For this task, a trained model of Yolov3 [19] is used. 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 a bounding box generated by the people detector module of the proposed system, can be seen in Figure 2

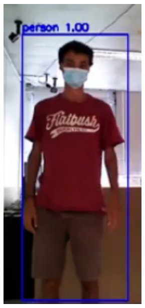

Fig. 2: Example of a person detection.

The people localizer module was already implemented in the MBOT and 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 $image\_geometry$ 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 $image\ plane$. 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, that should not be taken into consideration.

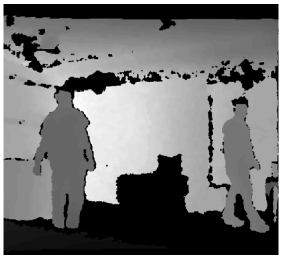

Fig. 3: Depth image showing two people. In this image, a brighter colour represents points that are further away from the camera. Totally black represents a NaN point where the depth could not be obtained. We can see here that a person is surrounded by background points that have bigger depth.

The positions obtained by the people localizer, which are in the $camera\ frame$, are first transformed 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.

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. In this work, the neural network TriNet [20] is used, due to its robustness and light computational effort, which is key for the deployment of the method in a mobile robot. TriNet is trained with batch hard triplet loss and the model used in this work was trained in the MARS dataset [21]. The Re-ID feature extractor takes as input the people detections from the people detector, feeds them to TriNet and outputs a 128-feature vector, which is the appearance descriptor, for each detection. An example is shown in Figure 4.

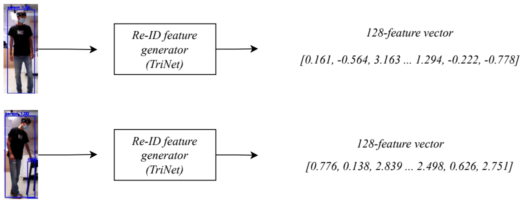

Fig. 4: Example of Re-ID feature extraction using the Re-ID feature generator.

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 [22]. 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. Each person is tracked using a simple Kalman filter, that predicts and updates the person's position in the map coordinate frame. At the same time, 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 section. The multi-tracker execution loop is ilustrated in Figure 5.

A general overview of the Kalman filter algorithm is presented in figure 6. 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

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

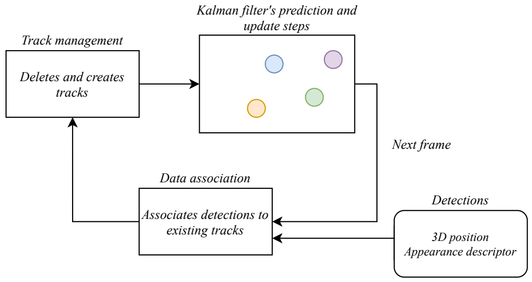
Kalman filter's prediction and update steps
Track management
Deletes and creates tracks
Next frame
Data association
Associates detections to existing tracks
Detections
3D position
Appearance descriptor

Fig. 5: Multi-tracker execution loop. 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.

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.

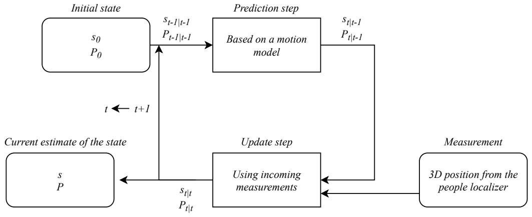

Fig. 6: Kalman filter algorithm, where s, P and t are the state, the covariance matrix and the timestep, respectively

C. Tracks and state estimation

The Kalman filter will predict and update each track's position in the world at each frame. Each track's state is modelled as:

$$\hat{x} = (x, y, z, vx, vy, vz), \tag{1}$$

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:

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

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

$$\sigma = \begin{bmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \\ \sigma_{vx} \\ \sigma_{vy} \\ \sigma_{vz} \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 1 \\ 1 \\ 1 \end{bmatrix} \tag{3}$$

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:

$$\hat{x}_{k|k-1} = A_k \hat{x}_{k-1|k-1} \tag{4}$$

$$\hat{x}_{k|k-1} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \hat{x}_{k-1|k-1} \tag{5}$$

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:

$$P_{k|k-1} = A_k P_{k-1|k-1} {A_k}^T + Q_k \tag{6}$$

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:

$$H_k = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{7}$$

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

$$y_k = Z_k - (H_k \hat{x}_{k|k-1}) \tag{8}$$

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

$$R_k = \begin{bmatrix} {\sigma_x}^2 & 0 & 0 \\ 0 & {\sigma_y}^2 & 0 \\ 0 & 0 & {\sigma_z}^2 \end{bmatrix} \tag{9}$$

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:

$$S_k = H_k P_{k|k-1} {H_k}^T + R_k \tag{10}$$

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:

$$K_k = P_{k|k-1} {H_k}^T {S_k}^{-1} \tag{11}$$

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:

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + (K_k y_k) \tag{12}$$

$$P_{k|k} = (I - K_k H_k) P_{k|k-1}, \tag{13}$$

where I is the identity matrix.

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

$$s(d,t) = D(d,t)^2 = (x_d - m_t)^T \cdot {P_t}^{-1} \cdot (x_d - m_t), \tag{14}$$

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. 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.

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:

$$c(d,t) = min(d(l_d, l_i) | l_i \epsilon L_t) \tag{15}$$

where

$$d(v,u) = \sqrt{(v_1 - u_1)^2 + ... + (v_{128} - u_{128})^2}, \tag{16}$$

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.

Algorithm .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         $U_t \longleftarrow U_t \cup t$
```

Algorithm .1 describes the data association algorithm. The assignment problem is represent 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. This step is important because it discards associations where a detection is not similar in appearance to a track, but at the same time 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. 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. The spatial and appearance thresholds were determined experimentally through several tests and the values used are:

$$T_{lower} = 300,\ T_{upper} = 700,\ S_{max} = 0.05m \tag{17}$$

After checking both metrics, the assignment problem is solved using the Hungarian algorithm [23], in line 13. From the resutl, unmatched detections and tracks are identified and the association set is filled. Finally, the association set is iterated and the cost of each association is checked, discarding unadmissible associations.

E. Track Management

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 7 by a state-machine.

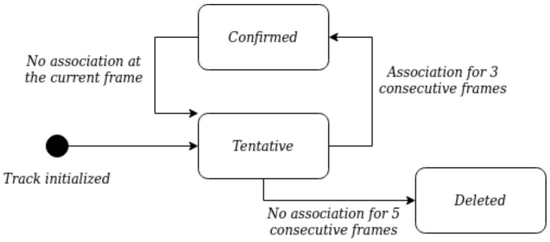
Confirmed
No association at the current frame
Association for 3 consecutive frames
Tentative
Track initialized
Deleted
No association for 5 consecutive frames

Fig. 7: 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 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.

The initialization of new tracks is done based on the unmatched detections at each frame. For each unmatched detection, a new track is initialized. The ID that is assigned to the new track is determined based on the target's appearance. 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 (15). In this dictionary, all the previous ID's 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 8.

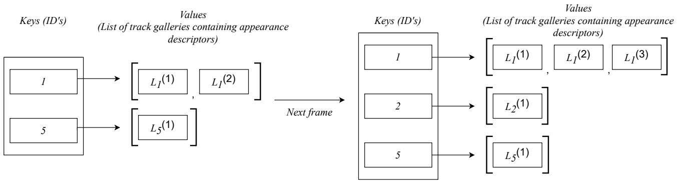
Keys (ID's)
Values (List of track galleries containing appearance descriptors)
1
5
L1(1)
L1(2)
L5(1)
Next frame
Keys (ID's)
Values (List of track galleries containing appearance descriptors)
1
2
5
L1(1)
L1(2)
L1(3)
L2(1)
L5(1)

Fig. 8: 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 .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 best match with a previously seen ID is found and , if the appearance distance between the detection and that ID's appearance is below a threshold, that ID is assigned to the new track. The threshold $T_{recover}$ was determined experimentally and has the following value:

$$T_{recover} = 400 \tag{18}$$

Algorithm .2: ID assignment to a new track

```
Input: New track $t$, ID dictionary $I$, list of active ID's $B$, next unused ID, $next\_id$
$Z \longleftarrow \emptyset$        // Initialize list of appearance distances
for each $k$ in $I$ do
    if $k$ is not in $B$ then
        $V = [c(t, I[k]^{(i)})]$        // Compute appearance distance vector
        $Z \longleftarrow min(V)$        // Append minimum of V to list Z
$min\_distance = min(Z)$        // Find minimum of Z
$min\_id = argmin(Z)$        // Find respective ID
if $min\_distance < T_{recover}$ then
    Track $t$'s ID $= min\_id$
else
    Track $t$'s ID $= next\_id$
    $next\_id = next\_id + 1$
```

IV. Re-ID Multi-Tracking Dataset

Analyzing the state-of-the-art for multi-object tracking datasets [24], 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. Considering the application of the method proposed in this work, 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.

A. Data and ground truth collection

The data was collected by teleoperating the MBOT in the ISRoboNet@Home Testbed[2] 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). The dataset consists in a total of 3144 RGB images, 3437 depth images and 2154 people instances. The RGB-D images, camera information, a map of the environment, odometry of the robot, transforms along reference frames and ground-truth are made available as ROS bag files[3]. People detections originated by the people detector module of the proposed system described earlier have also been included in the dataset. As ground-truth, 3D positions of people in environment and the robot and target's ID's were obtained using 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 as valid. After this process, ground-truth is associated with approximately 70% of the frames.

B. 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. The 7 sequences (videos) present in the dataset are: $Still$ and $Moving\ camera$ featuring three targets moving freely recorded with a static camera and with the camera rotating while the robot's base was static, respectively; 5 sequences recorded with the robot moving around the environment including $Moving\ base$, featuring occlusions, $Chairs$, featuring target's sitting and switching places, $Peoplefollowing$, where the robot is following a specific person, $Changing\ clothes$ 1 and $Changing\ clothes$ 2, where targets swith clothing in front of the camera and out of camera-view, respectively. Statistics of the sequences and the overall dataset are presented in Table I.

TABLE I: Re-ID Multi-target tracking dataset statistics

Sequence	Duration(s)	#RGB images	#Depth images	#People instances	%Ground-truth frames	#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	-

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. 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.

V. Experimental Results

A. 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 NVIDI GeForce 1060 6GB GPU. The system was implemented using ROS and Python and consists in the following ROS nodes: *darknet_ros_py*, which is used as the people detector, *mbot_object_localization*, which is used as the people localizer and *re_id_tracker*, which includes the Re-ID feature generator and the multi-people tracker.

B. Evaluation metrics

All the experiments conducted in this thesis were evaluated using the CLEAR MOT metrics [25]. CLEAR MOT are composed of two separated metrics that tackle different aspects of tracking performance: the multiple object tracking precision (MOPT) and the multiple object tracking accuracy (MOTA). These metrics are computed based on matches made between tracker hyphotesis and ground-truth objects. We considered a match to correspond to a ground-truth if their distance was below 1 meter. 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. The MOTA takes into account all object identity and track errors, such

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

[3] https://ulisboa-my.sharepoint.com/:f:/g/personal/ist187134_tecnico_ulisboa_pt/Emz8wKesZThJsO_TNoR1mTkBWJ6JjriW-0le5CsfWRd3ig?e=KfJtlT

TABLE II: 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	MOTP (m)	MOTA (%)
Still	330	267	15	4.55	4	1.21	0	0.00	0	0.195	94.24
Moving head	261	213	10	3.83	3	1.15	0	0.00	0	0.247	95.02
Moving base	522	296	51	9.77	16	3.07	13	2.49	4	0.237	84.67
Chairs	429	160	26	6.06	11	2.56	6	1.40	3	0.290	89.98
People follow	186	72	11	5.91	5	2.69	4	2.12	0	0.186	89.25
Changing clothes 1	280	172	15	5.36	6	2.14	1	0.36	0	0.148	92.14
Changing clothes 2	226	173	34	15.04	17	7.52	9	3.98	3	0.180	73.45
Total	2234	1353	162	7.25	62	2.78	33	1.48	10	0.215	88.50

as false positives, misses and mismatches. Both metrics were computed using the expressions in [25].

C. Experiments on the Re-ID Multi-Tracking Dataset

1) Evaluation results: The method was evaluated on the novel Re-ID Multi-Tracking Dataset using the CLEAR MOT metrics.Results divided by sequence and in total can be seen in Table II. Besides the CLEAR MOT metrics, an additional count is reported 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.

The total MOTP value is 0.215 meters which shows a good target position estimation. We can see greater error in the sequences where the robot is 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 thus 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 generally targets do not follow the constant velocity motion model used in the Kalman Filter predictions. 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. 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. 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, contrary to the performance in the $Changing\ clothes$ 2 sequence, where the results show the highest ratio of ID switches. In this dataset, the system achieved a 33Hz frame rate, which is suitable for real-time robotic applications.

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 $T_{recover}$. In Figure 9 and Figure 10, the values of MOTA, MOTP, ID switches and recall errors 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 [25]. In Table III, the values of MOTA, MOTP, ID switches and recall errors are also reported for different combinations of T_{lower} and T_{upper}.

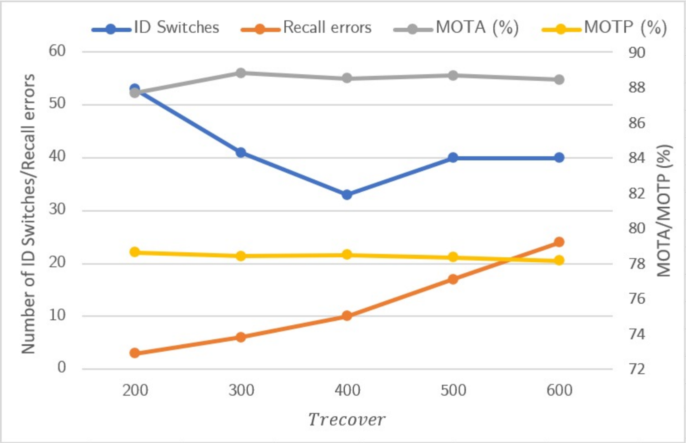

Fig. 9: Evaluation metrics on the Re-ID Multi-Tracking Dataset when varying the value of $T_{recover}$. The values of the other parameters are: S_{max} = 0.05, (T_{lower},T_{upper}) = (300,700)

Looking at the evaluation results on the Re-ID Multi-Tracking Dataset on Figure 9 and Figure 10 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. Out of the 5 values tested, the optimal values for $T_{recover}$ and S_{max} are 400 and 0.05, respectively. Five pairs (T_{lower},T_{upper}) were also tested and the results can be seen in Table III. The pair (300,700) is chosen because it produces the smallest number of ID switches and recall errors.

D. 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. 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

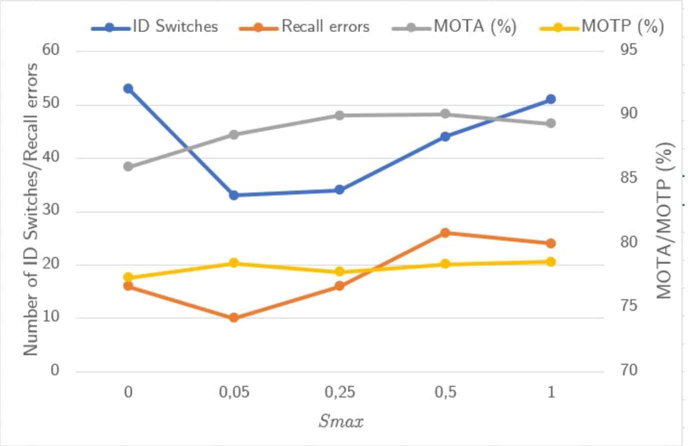

Fig. 10: Evaluation metrics on the Re-ID Multi-Tracking Dataset when varying the value of S_{max}. The values of the other parameters are: $T_{recover}$ = 400, (T_{lower},T_{upper}) = (300,700)

TABLE III: Evaluation metrics on the Re-ID Multi-Tracking Dataset when varying the values of T_{lower} and $_{Tupper}$. The best results for the various metrics are highlighted in bold. The values of the other parameters are: $T_{recover}$ = 400, S_{max} = 0.05

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**

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 IV.

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

E. Experiments on the Kinetic Precision dataset

The system was also evaluated in the Kinetic Precision dataset (KTP) [16]. Along with the presentation of the KTP dataset, Munaro and Menegatti [16] 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 work. 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 V.

Overall the system proposed in this thesis shows a better tracking performance on the KTP dataset than SOAM. The number of ID switches is almost the same 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. Nonetheless, the MOTP of the proposed system is always below 40cm, which is still an acceptable result for people tracking. 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. 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.

F. Discussion

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.

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. 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. 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. The system achieved a frame rate of 33Hz on the MBOT, which shows that the system is suitable for mobile robotics.

Finally, the system was also evaluated on the KTP dataset, achieved an overall MOTA score of 85.98% and MOTP of 0,354m on this dataset and produced overall better results than the method proposed in [16].

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.

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

	Objects	Matches	Misses	Misses(%)	False Positives	False Positives (%)	ID Switches	ID Switches (%)	Recall errors	MOTP (m)	MOTA (%)
Test sequence	690	174	54	7.82	20	2.86	14	1.99	4	0.190	87.25
Re-ID Multi-Tracking Dataset	2234	1353	162	7.25	62	2.78	33	1.48	10	0.215	88.50

	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

TABLE V: Tracking results for the Kinetic Tracking Precision dataset of the proposed system and the system presented in [16], divided by situation. Best results by situation are shown in bold.

VI. Conclusions and future work

In this work, a robust Re-ID based multi-people 3D tracker using RGB-D data to be deployed in a mobile robot is presented. A novel RGB-D Re-ID Multi-Tracking Dataset recorded with a moving camera mounted on top of a mobile robot and representative of real-world scenarios, including obstacles and occlusions, was constructed. 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 on a mobile robot, which is suitable for real-time robotic applications.

As future work, a better data association approach could be implemented using a probabilistic model to model the combined appearance and spatial distance between detections and existing tracks. The track management step can also be changed to reduce recall errors, by keeping deleted tracks with a random walk motion model combined with ”memory” of the last position where the target was seen. The experiments could be evaluated using different metrics, such as metrics more focused on the re-identification performance only and an addition to the dataset could be made with more sequences including larger groups of people and more challenging scenarios such as more complex appearance and pose changes or in an outdoor environment.

References

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

[2] P. U. Lima, C. Azevedo, E. Brzozowska, J. Cartucho, T. J. Dias, J. Goncalves, M. Kinarullathil,G. Lawless, O. Lima, R. Luzet al., “Socrob@ home,”KI-Kunstliche Intelligenz, vol. 33, no. 4, pp.343–356, 2019.

[3] 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.

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

[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, “Person re-identification bysymmetry-driven accumulation of local features,” in2010 IEEE computer society conference oncomputer vision and pattern recognition. IEEE, 2010, pp. 2360–2367.

[6] M. Pietikainen, “Local binary patterns,”Scholarpedia, vol. 5, no. 3, p. 9775, 2010.

[7] 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.

[8] 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.

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

[10] 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.

[11] T. Wengefeld, M. Eisenbach, T. Q. Trinh, and H.-M. Gross, “May i be your personal coach? bringingtogether person tracking and visual re-identification on a mobile robot,” inProceedings of ISR2016: 47st International Symposium on Robotics. VDE, 2016, pp. 1–8.

[12] Y. Murata and M. Atsumi, “Person re-identification for mobile robot using online transfer learning,”in2018 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.

[13] P. A. Carlsen, “Real-time person re-identification for mobile robots to improve human-robot inter-action,” University of Oslo, 2019.

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

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

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

[17] L. Chen, H. Ai, Z. Zhuang, and C. Shang, “Real-time multiple people tracking with deeply learnedcandidate selection and person re-identification,” in2018 IEEE international conference on multi-media and expo (ICME). IEEE, 2018, pp. 1–6.

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

[19] A. Farhadi and J. Redmon, “Yolov3: An incremental improvement,” inComputer Vision and PatternRecognition. Springer Berlin/Heidelberg, Germany, 2018, pp. 1804–02.

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

[21] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, “Mars: A video benchmarkfor large-scale person re-identification,” inEuropean Conference on Computer Vision. Springer,2016, pp. 868–884.

[22] N. Wojke, A. Bewley, and D. Paulus, “Simple online and realtime tracking with a deep associationmetric,” in2017 IEEE international conference on image processing (ICIP).IEEE, 2017, pp.3645–3649.

[23] J. Munkres, “Algorithms for the assignment and transportation problems,”Journal of the societyfor industrial and applied mathematics, vol. 5, no. 1, pp. 32–38, 1957.

[24] M. Camplani, A. Paiement, M. Mirmehdi, D. Damen, S. Hannuna, T. Burghardt, and L. Tao, “Multi-ple human tracking in rgb-d data: A survey,”IET Computer Vision, vol. 11, 06 2016.

[25] K. Bernardin and R. Stiefelhagen, “Evaluating multiple object tracking performance: the clear motmetrics,”EURASIP Journal on Image and Video Processing, vol. 2008, pp. 1–10, 2008.