2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Continuous Adaptation in Person Re-identification for Robotic Assistance

Federico Rollo[1,2,3], Andrea Zunino[1,2], Nikolaos Tsagarakis[2], Enrico Mingo Hoffman[4], and Arash Ajoudani[2]

***Abstract*— In scenarios of Human-Robot Interaction (HRI), it is often assumed that the robot should cooperate with the closest individual or that only one person is present. However, in real-life situations, such as shop floor operations, this assumption may not hold. Thus, it becomes necessary for a robot to recognize a specific target in a crowded environment. To address this problem, we propose a person re-identification module that uses continuous visual adaptation techniques. This module ensures that the robot can seamlessly cooperate with the appropriate individual despite its appearance changes or partial or total occlusions. We used both a laboratory environment and an HRI scenario where the robot followed a person to test our framework. During the test, the targets were asked to change their appearance and disappear from the camera's field of view to test the module's ability to handle challenging cases of occlusion and outfit variations. We compared our framework with a state-of-the-art Multi-Object Tracking (MOT) method, and the results showed that our module, shortly named CARPE-ID, accurately tracked each selected target throughout the experiments in all cases except for two cases. In contrast, the MOT had an average of 4 tracking errors for each video.**

I. Introduction

To effectively interact with humans in environments where they coexist, robots must be able to identify individuals and adjust their actions accordingly. This personalized approach is often overlooked in traditional Human-Robot Interaction (HRI) settings which assume that the robot will collaborate with the nearest person and that there will be no other human interventions or distractions during task execution. However, this assumption isn't always realistic, particularly when tasks involve mobility and a larger workspace. Therefore, future collaborative robots need to recognize and identify their human counterparts to provide personalized assistance.

Previous studies aimed to address the problem of tracking humans using robots by employing an offline identification step followed by online tracking. For instance, one study focused on a mobile robot that was able to follow a person [12]. Another study introduced a re-identification application even in the presence of partial occlusions [17]. However, despite the progress made, these frameworks are not able to adapt to real-time changes in the target's appearance and cannot recover re-identification after total occlusions. These frameworks can efficiently follow a target on the image but often need human intervention to restart the tracking [5] [8].

We aim to minimize human intervention in robot recognition and re-identification. In a previous work, presented in [12], we faced the challenge of not being able to track the target person if they changed their appearance after the calibration step, such as by changing their outfit. To overcome this limitation, we introduce a novel re-identification module that utilizes a deep learning approach based on feature extraction and continual adaptation to accommodate changes in the target's appearance. During tracking, the robot constantly acquires new images and uses the new appearance information to update an ideal target representation model to re-identify the target when tracking fails.

We utilize a Multi-Object Tracking (MOT) algorithm known as yolo_tracking (MOT) based on the well-known YOLO framework [11] and the StrongSORT tracker [3] and provide an additional person re-identification layer that adapts itself based on target appearances. This layer enables a Re-identification system to handle the jumps in identification numbers (ID) that often occur because the object may change its ID because of occlusions or appearance differences.

The following is a summary of the contributions made by this work:

- A Continuous Adaptation framework for Re-identification (CARPE-ID) has been proposed to achieve personalized HRI tasks with a specific target;
- tests in a real Human-Robot collaborative scenario;
- limitations evaluation and potential solutions.

The following outlines the structure of the paper. First, in section II, we review the current state-of-the-art in Single and Multi-Object Tracking as well as HRI re-identification works and surveys. Our approach is then presented in section III. Following this, we show our experiments and results in section IV, including a discussion on our achievements, performance, and limitations. Lastly, we summarize the entire work in section V and present our conclusive statements.

II. Literature review

In the field of object tracking, various systems have been proposed. Multi-object tracking (MOT) and Single-object Tracking (SOT) are two commonly used methods for tracking objects in images. MOT algorithms are capable of detecting and tracking multiple classes of objects, while SOTs require only an initial guess of the object's bounding

This work was supported by the Leonardo S.p.A under Grant LDO/CTI/P/0020481/23

[1]Autonomous Systems & Robotics, Leonardo Labs, Genoa, Italy
Corresponding author: federico.rollo@leonardo.com
[2]HHCM & HRII, Istituto Italiano di Tecnologia, Genoa, Italy
[3]Industrial Innovation, DISI, Università di Trento, Trento, Italy
[4]Université de Lorraine, CNRS, Inria, LORIA, Villers-lès-Nancy, France

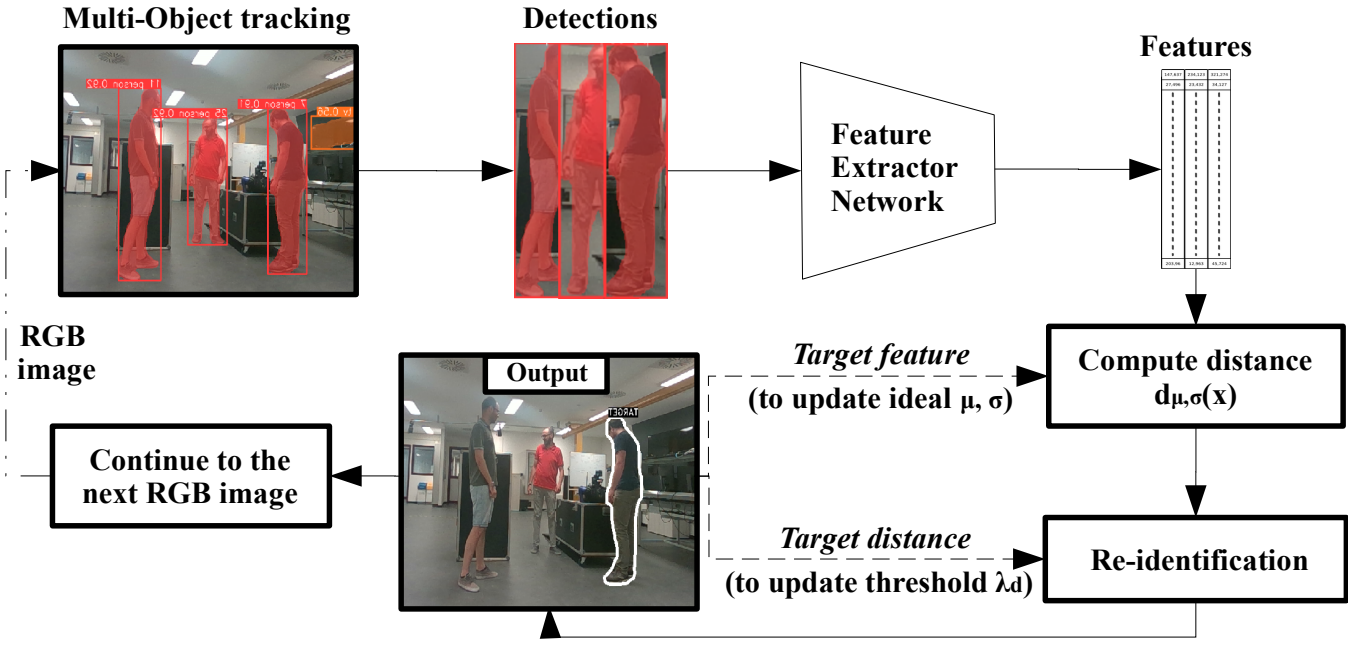


Fig. 1: The pipeline framework begins with an image input and ends with a re-identified target output. The first module is the MOT, where a neural network provides a rough tracking of objects present in the image. From the MOT module, the detections obtained are passed into a feature extractor that gives the output feature vectors $\boldsymbol{x}$. The feature vectors $\boldsymbol{x}$ allow us to compute the statistical distance $d_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{x})$ (as shown in equation (1)) which will be used by the re-identification module. If the target is successfully re-identified, the target statistical distance $d_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{x})$ and the target features $\boldsymbol{x}$ are used to update the re-identifier threshold λ_d and the ideal target representation $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ (represented by dashed lines).

box to track a selected object of interest in the image, regardless of its class type.

According to a recent survey [14], there are four categories of SOT techniques based on feature, segmentation, estimation, and learning. The researchers of the survey primarily studied the learning approaches and provided a summary of datasets and metrics used for the evaluation of such algorithms. In the same manner, in [18] the authors analyzed recent advancements in SOT, with a detailed focus on correlation and deep learning algorithms. Instead, the authors in [5] offered a more comprehensive analysis of object tracking using siamese networks and discriminative correlation filters. On the other hand, the review presented in [8] discusses the recent developments in MOT algorithms.

SOT and MOT algorithms do not meet the demands of personalizable HRI (Human-Robot Interaction) as they are often not robust enough to re-identify targets after partial or total occlusions. For robotics applications, more specialized methods are offered in the literature. The use of skeleton and face points is widely employed in re-identification for robotic assistance and person tracking. For instance, in [10], the authors proposed a method for creating a descriptor of the target person using soft biometric features extracted from the depth and colour information. They used AlphaPose [4] to extract 3D skeleton points and create a standard posture for the target person. This posture was then used to partition the target person using a grid, and the mean colour of each grid cell was used for re-identification. However, this approach has limitations, as it could lead to errors for people wearing comparable dresses. Another study [17] developed a robot person-following system that is robust to partial occlusion, which can be caused by the limited camera field of view. They use a predefined model and some skeleton points to re-identify the target. Specifically, they used some heuristics extracted from the target structure for person re-identification and tracking. However, this method assumes that the person always remains within the camera's field of view, which may not always be the case.

In [7], the authors propose a face recognition method for person re-identification and tracking. They first train a metric model offline using labelled data, and then use online face information to match the target and update the metric model. To merge appearance information with the skeleton, they introduce the feature funnel model (FFM). Another work [15] focuses on using a face re-identification approach to improve human-robot interaction. The authors use a pre-trained feature extractor CNN to re-identify target faces in an unsupervised manner. However, both these approaches

assume that the human collaborator always faces the robot, which is not always the case as observed in our experiments.

The study in [6] uses an RGB monocular approach for tracking human targets. The researchers first extract the skeleton using OpenPose [1] and then use an unscented Kalman filter to track the target. For the re-identification of the target, convolutional channel features are employed along with boosting techniques. However, the proposed method requires a calibration process at the beginning to learn the deep feature appearance, which will be used to re-identify the target. A different approach [2], used a thermal camera to re-identify the target and perform tracking. They trained a network using a custom dataset which is sampled using techniques founded on an entropy process to obtain a description for all the people. They use geometries extracted using descriptors to train a support vector machine classifier to distinguish people. Anyway, extensive training is required for the thermal camera integration, and the method cannot adjust to target changes, which is not suitable for our specifications.

It becomes evident from the above examination, that the restrictions highlighted in these works do not satisfy the specifications we proposed for a reliable tracking system. With the proposed approach in this paper, we solve the problem of person re-identification in the presence of target appearance shifts and after occlusions. In the human-robot interaction field, this is not enough studied for tracking systems as far as the authors know.

III. Method

Our approach for tracking people is presented in the pipeline illustrated in Figure 1. We have also proposed a pseudo-code implementation, which can be found in Algorithm 1. The tracking process begins when a user selects an ID. We initially collect the detected people's appearances obtained with the MOT in a database. This allows the user to choose the person they want to follow. Once the ID is selected, we feed the RGB image into the MOT algorithm to obtain a rough detection and tracking, assigning IDs to each person. We use these detections to crop each person's image, which is then fed into a deep neural network trained for person recognition (more details in section IV). This network extracts features, represented as a one-column vector, which we employ for re-identification and tracking of the target.

For each frame, the application performs two sequential steps: re-identification of the target and updating the ideal representation of it.

A. Re-identification

A statistical distance is computed from the features $\mathbf{x}$ extracted from the detected people as:

$$d_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{x}^*) = \sqrt{\frac{1}{D}\sum_{i=1}^{D}\left(\frac{\boldsymbol{x_i^*} - \boldsymbol{\mu}_i}{\boldsymbol{\sigma}_i}\right)^2}, \tag{1}$$

where $\boldsymbol{i}$ denotes the i-th index of a vector. The values of mean and standard deviation, denoted by $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ respectively, represent the model of the ideal target we aim to track. The mean $\boldsymbol{\mu}$ is initialized with the initial feature vector $\boldsymbol{x}$ while the standard deviation $\boldsymbol{\sigma}$ is initialized with a zeros-vector, both having dimensions D.

In the next step, we first check if there is any feature vector having the same ID as the one selected by the user. If there is, then we directly output the corresponding detection with its ID. If not, we proceed to the re-identification module. Here, the module compares the adaptive threshold λ_d with the feature vector with the smallest distance from the target model. If the module is unable to re-identify any person among the detected ones, then the framework continues to analyze the next RGB frame.

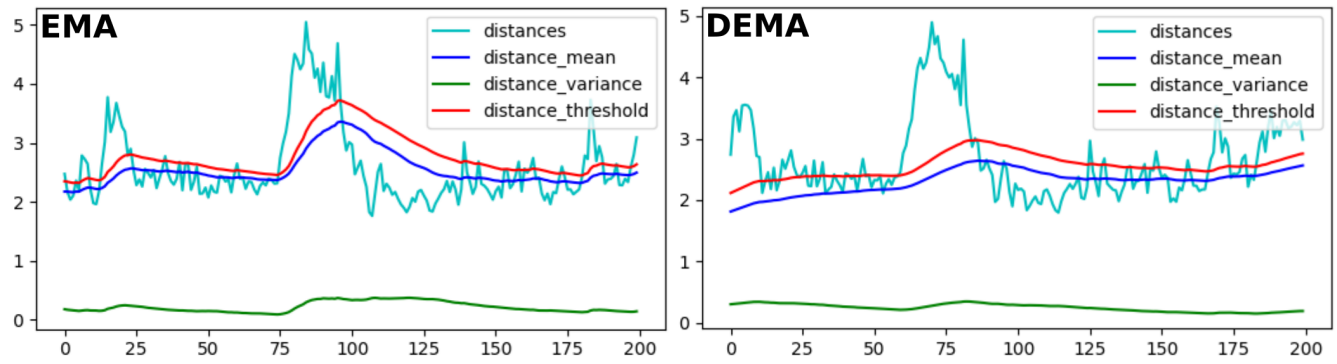
Algorithm 1 CARPE-ID framework

Input: $tracking\_id$
$\boldsymbol{\mu}, \boldsymbol{\sigma}, \mu_d, \sigma_d, \lambda_d \leftarrow initialize\_variables()$
while True **do**
  $\boldsymbol{I_{rgb}} \leftarrow cam.getRGB()$
  $dets \leftarrow mot.infer(\boldsymbol{I_{RGB}})$
  $feats \leftarrow reid.infer(dets)$
  $min\_dist \leftarrow MAX\_FLOAT\_NUM$
  $tracking\_feature \leftarrow Null$
  ▷ ***Re-identification***
  for $feature \in feats$ **do**
    if $tracking\_id = feat.id$ **then**
      $tracking\_feature \leftarrow feature$
      break
    else
      $d_{\boldsymbol{\mu},\boldsymbol{\sigma}} \leftarrow get\_distance(feature, \boldsymbol{\mu}, \boldsymbol{\sigma})$,
      if $d_{\boldsymbol{\mu},\boldsymbol{\sigma}} < min\_dist \ \ \& \ \ d_{\boldsymbol{\mu},\boldsymbol{\sigma}} < \lambda_d$ **then**
        $min\_dis \leftarrow d_{\boldsymbol{\mu},\boldsymbol{\sigma}}$
        $tracking\_feature \leftarrow feature$
        $tracking\_id \leftarrow feature.id$
      end if
    end if
  end for
  ▷ ***Target Model Update***
  if $target\_feature \neq Null$ **then**
    $var \leftarrow compute\_var(\boldsymbol{\mu}, tracking\_feature)$
    $\boldsymbol{\mu} \leftarrow DEMA(\boldsymbol{\mu}, tracking\_feature, \Delta_f)$
    $\boldsymbol{\sigma} \leftarrow DEMA(\boldsymbol{\sigma}, var, \Delta_f)$
    $var_d \leftarrow compute\_var(\mu_d, min\_dist)$
    $\mu_d \leftarrow DEMA(\mu_d, min\_dist, \Delta_{\lambda_d})$
    $\sigma_d \leftarrow DEMA(\sigma_d, var_d, \Delta_{\lambda_d})$
    $\lambda_d \leftarrow \mu_d + 2\sigma_d$
  end if
end while

B. Model update

After re-identifying the target, we use its feature vector and corresponding statistical distance to continually adjust the ideal target representation ($\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$), as well as the threshold λ_d used during the re-identification process using the following equation.

$$\lambda_d = \mu_d + 2\ \sigma_d, \tag{2}$$



(a) EMA function effects.

(b) DEMA function effects.

Fig. 2: This figure illustrates a comparison between DEMA and EMA effects on threshold filtering and the ideal target representation. On the left, they follow the light blue line (*i.e.*, the statistical distance) with a small delay. On the right, the damping guarantees that their values do not track the distance behaviour when peaks are present. Instead, they slow down, waiting for the adaptation of the target model to the newly acquired appearances.

where μ_d and σ_d are the mean and the variance of the statistical distances of the target. We update the variables $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, μ_d, and σ_d online using a Damped version of the Exponential Moving Average ($\boldsymbol{\chi}_{dema}$), which we refer to as DEMA. The formula for DEMA is expressed in the following equation.

$$\boldsymbol{\chi}_{dema}[k] = \alpha_{damp}\ \boldsymbol{\psi} + (1 - \alpha_{damp})\ \boldsymbol{\chi}_{dema}[k-1], \quad (3)$$

In this equation, $\boldsymbol{\psi}$ represents the new value at a particular point in time denoted by $[k]$ in the DEMA. The term α_{damp} is a weighting term that defines the importance of the newly acquired value ψ in comparison to the DEMA $\boldsymbol{\chi}_{dema}[k-1]$ of the previous time instant. The calculation of α_{damp} is based on several factors and depends on the specific implementation following the formula:

$$\alpha_{damp} = \frac{2}{N_{damp} + 1}, \quad (4)$$

To calculate the damping factor for the DEMA, we use the formula $N_{damp} = N \cdot \Delta$, where N represents the number of DEMA updates, and Δ depends on the type of information we are updating. It is a best choice to initialize $N = 0$ for faster convergence and set a fixed upper-bound N_{max} to ensure that the DEMA is affected by new values over time. This is because, as N increases, the damping factor α_{damp} tends towards 0.

There are two different damping factors for the DEMA. One is for the ideal target features represented by $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, which we call Δ_f (see (5)). The other is for the threshold λ_d, which is computed with μ_d and σ_d and is called Δ_{λ_d} (see (6)). Figure 2 compares the EMA with and without the damping factors. There we demonstrate that, adding the damping factor Δ_{λ_d}, the noise of the threshold λ_d (in red) at high frequencies is attenuated and the whole line is smoothed. In this way, we prevent wrong re-identifications by reducing λ_d peaks which appear in standard EMA.

$$\Delta_f = min(1,\ \frac{d_{\boldsymbol{\mu},\boldsymbol{\sigma}}}{2}), \quad (5)$$

$$\Delta_{\lambda_d} = max(1,\ 2\frac{d_{\boldsymbol{\mu},\boldsymbol{\sigma}}}{\lambda_d}). \quad (6)$$

When we apply the damping factor Δ_f from equation (5) in DEMA equation (3) with N_{damp}, we are avoiding situations where the distance in equation (1) from the current feature is small. In these cases, the target model and the feature vector are comparable, which would cause overfitting of the ideal representation, especially when the target is stationary in the same pose. The damping factor Δ_{λ_d} in equation (6) is used to prevent the threshold from growing excessively when the ratio between the distance $d_{\boldsymbol{\mu},\boldsymbol{\sigma}}$ and the threshold λ_d is too large. This helps to avoid wrong re-identification.

To enhance the algorithm's reliability, we implemented a *blacklist manager*. This manager assesses the IDs provided by the MOT and identifies which ones belong to distractors. By doing so, we can exclude them during the re-identification step. Whenever the target ID provided by the MOT remains unchanged during tracking, the manager adds the IDs associated with other individuals in the image (i.e. distractors) to the blacklist. This ensures that we minimize any re-identification errors that may occur in challenging situations.

IV. Validation

To test our framework, we carried out two experiments. Firstly, we validated the framework by capturing videos from different angles using a fixed camera in a laboratory setup (as explained in section IV-A). Secondly, to further validate the proposed method, we tested it in a human-robot interaction scenario, specifically a robot person-following scenario (as described in section IV-B). Finally, we analyze and discuss the results obtained from both experiments in section IV-C.

The system was run on a notebook comprising of an *IntelCore i9* processor and an *NVIDIA RTX 3080 Laptop* GPU. For image acquisition, we used an Intel Realsense

D455 camera and a Robotnik RB-Kairos+ was used as the assistant robot. To extract features (a one-column vector of dimensions 256), we used an IBN-ResNet-50 [9] trained on the dataset MSMT17 [16].

A. Individual experiments

We attempted to use the PersonPath22 dataset [13], which is a visual person-tracking dataset, to evaluate our framework. However, we discovered that this dataset is not suitable for our evaluation due to the following reasons:

- The videos in this dataset do not depict human-robot collaboration scenarios. Instead, they mainly comprise security camera footage or videos of crowded environments where people are far from the camera and appear in the image only for short periods.
- The videos are too brief to accurately validate the re-identification module.
- In most cases, the Multiple Object Tracking (MOT) algorithm we used in our framework already performed the correct tracking. People do not exit and re-enter in the image and our re-identification module is not necessary.

We faced certain limitations in validating the framework, which led us to acquire a custom dataset consisting of 18 videos for this purpose, consisting of 53 minutes. Additionally, for videos with multiple actors, the framework is evaluated separately for each person, bringing the total analyzed duration to 113 minutes. The videos were shot in two laboratory setups featuring single-person or group scenarios. Participants in the videos were asked to exit and re-enter the camera field of view and to change their appearance (for example, by varying clothes) to validate the framework's ability to re-identify totally or partially occluded individuals and to adjust the model following the newly acquired appearances. We analyzed the experiment and computed the following performance:

- The state-of-the-art MOT had min, mean, and max tracking lengths of 3.02, 21.2, and 52.2 seconds, respectively, with some outliers extending the max to 97.17 seconds (see figure 3a).
- The minimum, mean and maximum re-identification (Re-ID) delay, which is the time taken by the framework to re-identify the target, were found to be 0.06, 1.1 and 2.6 seconds respectively (as shown in figure 3c). However, there were two instances where the tracker took significantly longer, 6.5 and 12.1 seconds, for the re-identification process.
- The Multiple Object Tracking (MOT) algorithm had a minimum and maximum failure rate of 2 and 7 times, respectively, with an average failure rate of 4 times for each video, as shown in figure 3b. In all the videos we had only 2 errors. This shows the framework's reliability to totally and partially occluded targets.
- The minimum, mean, and maximum re-identifications made by our framework were 1, 4, and 7. (See Figure 3d).

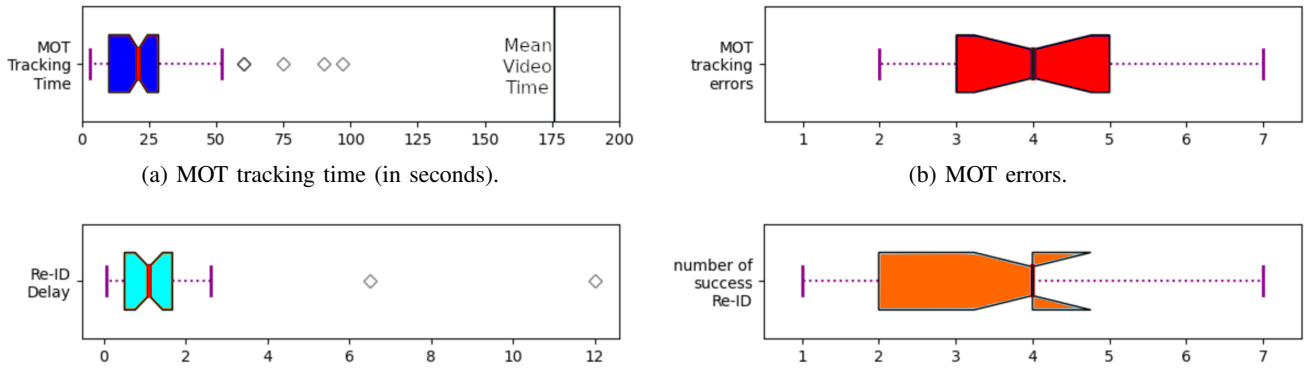
B. HRI task experiments

We conducted a real-world experiment to validate our framework using a robot person-following scenario, similar to the one described in [12]. In our setup, the target person was asked to move around while the robot followed them and avoided obstacles. However, in our experiment, the targets changed their appearances during the following and passed through obstacles to test the framework's robustness to partial or total occlusions. In all our experiments, the robot was able to correctly track, re-identify, and follow the target person, even in a laboratory setup where other people were working. An example of the setup of the experiments is represented in Figure 4. The robot (blue line) must follow the target which had to walk on the dashed line in red. We experimented ten times with five different people as targets, covering a total distance of 837 meters.

C. Discussion

We here analyze the evaluation test results to validate our contributions. The results achieved in section IV-A show that personalized approaches are necessary for re-identifying the target person in human-robot interaction tasks. The recognition capacities demanded by these scenarios are difficult to achieve because the target appears in the image at different times and spaces and with different appearances. We require a more reliable application because SOT and MOT have proven to be limited in our experiments. This statement is supported by the results presented in Figure 3a, where we can see that the MOT algorithm has a mean tracking time of 21.2 seconds, while the mean video time is 176 seconds and the minimum video time is 94 seconds. In the figure 3c, we have shown the time delay between when the target person appears in the camera field of view and the time they are re-identified. The mean delay we obtained (1.1s) is acceptable for HRI applications. However, there were two cases where the re-identifier could not re-identify the target in less than 6 seconds. This happened because the target person changed their appearance while outside the camera's field of view. For example, the target changed some clothes, fooling the re-identification module which wasn't able to re-identify him in an acceptable time, resulting in errors. In figure 3b and figure 3d, we compare the times when the MOT algorithm lost track of the target with the number of times the Re-ID framework can re-identify the target. Both have a mean of 4, indicating that the MOT errors are recovered by our framework. However, the figure reveals that MOT mistakes are more than re-identifications. This is because the MOT algorithm quickly switches between IDs, and the re-identification delay takes more time to identify the target. For example, in a video dataset, the MOT changed the target's ID three times in 1.4 seconds, and the re-identifier could only re-identify the last one in real-time.

In the person-following scenario, we present an example of how our framework can be used. Figure 4 demonstrates experiments where a robot follows a target in a laboratory environment with other people present. Despite some minor delays in re-identification, the robot completed the task of



(a) MOT tracking time (in seconds).

(b) MOT errors.

(c) Our delay (in seconds) for re-identification.

(d) Our number of re-identification.

Fig. 3: Individual experiments statistics

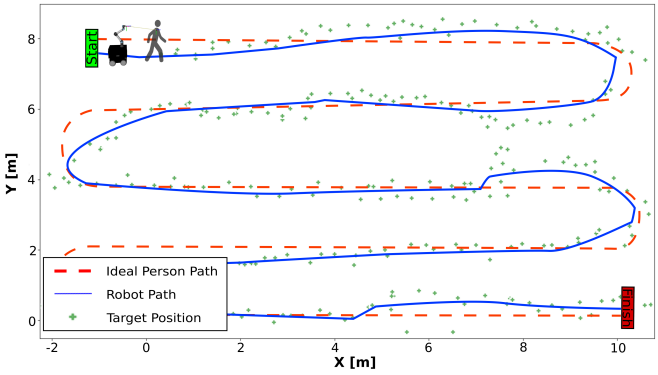


Fig. 4: This is an example of a person-following experiment. The person being followed must roughly follow a particular path (indicated by the red dashed line), while the robot (indicated by the blue line) follows them. The green plus signs represent the target positions, which are calculated using the CARPE-ID framework tracking output. The robot starts at the green position and needs to follow the person until they reach the red finish position.

following without any human intervention in all experiments.

V. Conclusions

We have developed a framework for person re-identification that can be adapted to different human-robot interaction scenarios. Our pipeline consists of a detection and preliminary person-tracking algorithm, a feature extraction network for appearance representation, and a re-identification structure based on statistical distance. We also incorporated an adaptable ideal target representation, threshold computation, and a custom version of the exponential moving average with a damping factor (DEMA). This setup enables us to track people even when they are partially or fully occluded or when their appearance changes during tracking (for instance, by wearing a sweatshirt). These are challenging situations that are difficult for traditional MOT algorithms to handle, as our experiments have shown.

It is worth noting that this work has some limitations. One of them is the issue of catastrophic forgetting, which means that the algorithm tends to forget the appearance of the target as time passes because it adapts to the new information it gathers. This can be faced with an online method that uses continuous learning techniques. The features extraction network can be trained using people detections extracted during *CARPE-ID* execution. This would specialize the feature extractor with the tracked person appearances forcing the outputted feature to be similar when belonging to the target and different if associated with a distractor. Proximity to the camera poses another limitation of our algorithm. However, this is not usually a problem in human-robot collaboration scenarios because the human is generally near to the robot and the camera. Lastly, the algorithm's time performance may suffer when more than 20 people are present in the image. Nevertheless, it is worth noting that such scenarios are rare, especially considering the robot camera's point of view.

In the future, we plan to improve the re-identification capabilities of our system when the target changes appearance outside of the camera's field of view. We intend to accomplish this by implementing a specialized network that can recognize the face of the target to stabilize the re-identifications. Additionally, we aim to improve *CARPE-ID* by introducing continual learning techniques as stated in the limitations. This technique involves using online self-supervised training to specialize the feature extraction network on the target appearance, and to differentiate more from the appearances of other people in the vicinity who may cause distractions.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[2] Serhan Coşar and Nicola Bellotto. Human re-identification with a robot thermal camera using entropy-based sampling. *Journal of Intelligent & Robotic Systems*, 2020.

[3] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 2023.

[4] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[5] Sajid Javed, Martin Danelljan, Fahad Shahbaz Khan, Muhammad Haris Khan, Michael Felsberg, and Jiri Matas. Visual object tracking with discriminative filters and siamese networks: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[6] Kenji Koide, Jun Miura, and Emanuele Menegatti. Monocular person tracking and identification with on-line deep feature selection for person following robots. *Robotics and Autonomous Systems*, 2020.

[7] Hong Liu, Liang Hu, and Liqian Ma. Online rgb-d person re-identification based on metric model update. *CAAI Transactions on Intelligence Technology*, 2017.

[8] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. *Artificial intelligence*.

[9] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[10] Cosimo Patruno, Roberto Marani, Grazia Cicirelli, Ettore Stella, and Tiziana D'Orazio. People re-identification using skeleton standard posture and color descriptors from rgb-d data. *Pattern Recognition*.

[11] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[12] Federico Rollo, Andrea Zunino, Gennaro Raiola, Fabio Amadio, Arash Ajoudani, and Nikolaos Tsagarakis. Followme: a robust person following framework based on visual re-identification and gestures. In *2023 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*, pages 84–89. IEEE, 2023.

[13] Bing Shuai, Alessandro Bergamo, Uta Buechler, Andrew Berneshawi, Alyssa Boden, and Joe Tighe. Large scale real-world multi person tracking. In *European Conference on Computer Vision*. Springer, 2022.

[14] Zahra Soleimanitaleb and Mohammad Ali Keyvanrad. Single object tracking: A survey of methods, datasets, and evaluation metrics. *arXiv preprint arXiv:2201.13066*, 2022.

[15] Yujiang Wang, Jie Shen, Stavros Petridis, and Maja Pantic. A real-time and unsupervised face re-identification system for human-robot interaction. *Pattern Recognition Letters*.

[16] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

[17] Hanjing Ye, Jieting Zhao, Yaling Pan, Weinan Cherr, Li He, and Hong Zhang. Robot person following under partial occlusion. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[18] Yucheng Zhang, Tian Wang, Kexin Liu, Baochang Zhang, and Lei Chen. Recent advances of single-object tracking methods: A brief survey. *Neurocomputing*, 455:1–11, 2021.