Following Closely: A Robust Monocular Person Following System for Mobile Robot
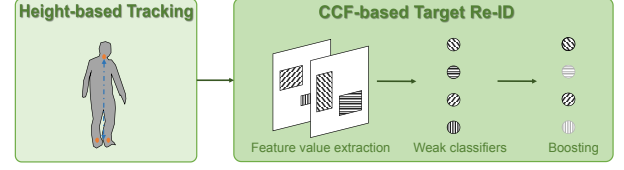
Hanjing Ye[1], Jieting Zhao[2], Yaling Pan[1], Weinan Chen[2] and Hong Zhang[2*]

Abstract— Monocular person following (MPF) is a capability that supports many useful applications of a mobile robot. However, existing MPF solutions are not completely satisfactory. Firstly, they often fail to track the target at a close distance either because they are based on visual servo or they need the observation of the full body by the robot. Secondly, their target Re-IDentification (Re-ID) abilities are weak in cases of target appearance change and highly similar appearance of distracting people. To remove the assumption of full-body observation, we propose a *width-based* tracking module, which relies on the target width, which can be observed even at a close distance. For handling issues related to appearance variation, we use a global CNN (convolutional neural network) descriptor to represent the target and a ridge regression model to learn a target appearance model online. We adopt a sampling strategy for online classifier learning, in which both long-term and short-term samples are involved. We evaluate our method in two datasets including a public person following dataset and a custom-built with challenging target appearance and target distance. Our method achieves state-of-the-art (SOTA) results on both datasets. The code and dataset of our work in this research are publicly available in **https://github.com/MedlarTea/MPF_GRR_SLT**.

(a) Height-based tracking with CCF [2]

(b) Our width-based tracking with GRR_SLT

Fig. 1. A MPF system consists of two key modules: tracking module and target Re-ID module. (a) is the SOTA method with a *height-based* tracking module and a CCF (convolutional channel feature) based target Re-ID module. (b) is the proposed method with a *width-based* tracking module and a GRR_SLT-based module (a global descriptor and a ridge regression combining with a short-long-term sample set)
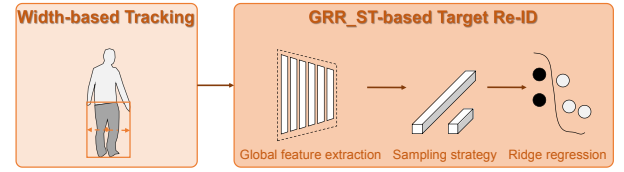
I. INTRODUCTION

Nowadays, mobile robotics is a fast-growing field of research. Due to its capability, many useful applications can benefit from the deployment of a mobile robot, such as surveillance, emergency rescue, entertainment, library guides, medical care, industry collaboration and so on. Some of these applications involve human-robot interaction, in which a mobile robot must have abilities of perception, localization, navigation, locomotion and even cognition about people in its working environment. Person following [1] is a capability that supports many useful applications of a mobile robot.

In order to perform person following, some proposed methods [3]–[9] track multiple people with the help of a distance measurement sensor such as LiDAR and RGBD camera. Once selecting one person in the field of view of the robot, it will follow the person based on the tracking position. To deploy person following on low-cost mobile robots, [10] uses a vision-based single object tracking (SOT) [11]–[13] to track the target, and relies on a visual servo to follow the target. [2] proposes a monocular-vision person following (MPF) system to track and follow the target by an assumption

of full-body observation. However, these MPF systems still suffer from challenging situations involving close observation, target appearance change and highly similar appearance between the target and distracting people.

To address the above problems, we propose a robust MPF system consisting mainly of a people tracking module and a target Re-ID module. Our people tracking module can obtain multiple people tracks even at a close distance to the target because our *width-based* people detection and position estimation make use of the width information of people as a prior without requiring the full-body observation of people. By using high-level global features of the target that are learned and adapted online, our target Re-ID module can re-identify the target even when the target is lost in difficult cases when it moves out of the view due to abrupt motion or distracting people of similar appearance appear in the scene. We rely on a sampling strategy to properly consider historic observations of the target used by the online classifier to mitigate an overfitting problem effectively.

In summary, in this paper, we propose a robust MPF system with the following two contributions:

- We design a *width-based* people tracking module to endow robots with the ability to follow the target reliably even at close distance; and
- we propose a robust target Re-ID module considering high-level target features and historic observations in

*corresponding author (hzhang@sustech.edu.cn).

[1]Hanjing Ye and Yaling Pan are with the Biomimetic and Intelligent Robotics Lab (BIRL), Guangdong University of Technology, Guangzhou, China.

[2]Jieting Zhao, Weinan Chen and Hong Zhang are with the Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, China.
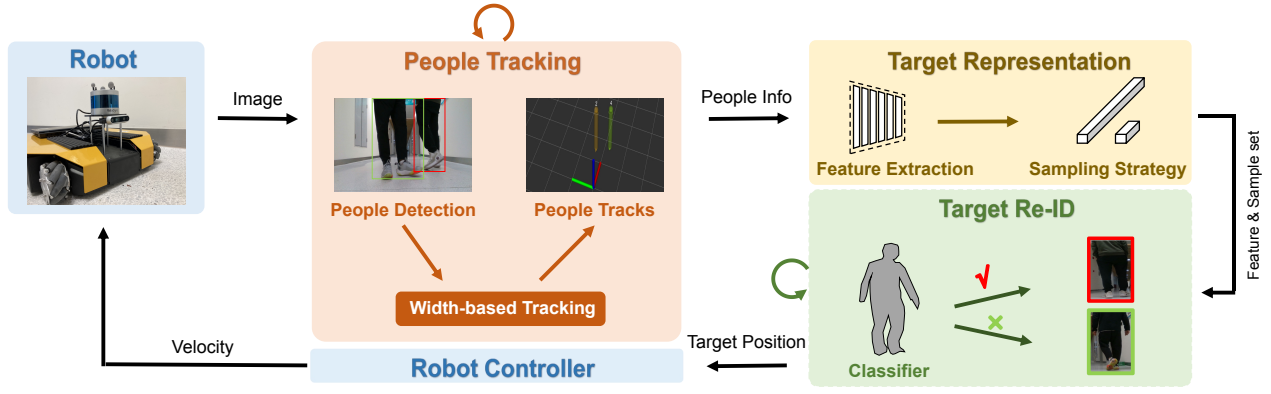
Fig. 2. The framework of the MPF system, where the arrow with a circle line means previous information is used. Our main contributions are: 1) **width-based** tracking, which can track people even at a close distance by utilizing the observale boxes; and 2) **target representation**, which is robust in challenging situations of distracting people of high similarity to the target by using high-level features and historic observations to construct a discriminative target appearance model.

constructing a robust target appearance model.

The rest of the paper is organized as follows. Section II presents related works and the motivation of our study. Section III introduces our MPF framework and details the proposed key modules involving people tracking module and target Re-ID module. Section IV is about experiments and implementation details. Section V shows the experimental results and discussion. Finally, section VI concludes this paper.

II. RELATED WORKS

A. Monocular-Vision Person Following Robot

Many existing works about the person following [3]–[9] use distance measurement sensors, which can be expensive and have difficulty in dealing with cluttered indoor environments. A few works are based on monocular vision. [10] tracks people in the image space and realizes the person following by visual servo. For estimating a person's position in the robot space, inspired by person position estimation with the hypothesis of known height and ground plane position [14]–[16], [2] proposes a *height-based* MPF framework. However, it would fail at a close distance to the target because it requires that people's full bodies of people including their necks are observable. In real robots applications, the close-range following commonly occurs, e.g., in robot dog following and service robot collaboration. To solve this problem, we design a *width-based* tracking method with a hypothesis of the known width of the target. Then we are able to track people at a close distance by using bounding boxes returned by a people detector that works reliably even at a close distance.

Target loss is a common occurrence in person following applications. To solve this problem, [2] includes a target Re-ID module in its design as shown in Figure 1(a). The Re-ID module uses features that are computed from convolutional channel feature (CCF) map of a rectangle region. Then a number of weak Bayes classifiers are initialized by randomly choosing a CCF map and a rectangle region. Lastly, online

boosting [17] is used to select weak classifiers to form a strong classifier. The main weakness of the proposed Re-ID module in [2] is that it cannot distinguish the target from distracting people of similar appearance due to the use of low-level features. To improve the Re-ID ability, we use a high-level global descriptor and a ridge regression model to learn a robust representation of the target online that is robust even in situations of high similarity of the target and distracting people.

B. MOT, SOT and Sampling Strategy

Much can be learned from the literature on multiple-object tracking (MOT) and single-object tracking (SOT) in order to design a learning strategy properly in our MPF problem. MOT [18]–[20] realizes multiple people tracking in the image space with boxes and pre-trained global descriptors. However, pre-trained descriptors are not adaptable, which would easily cause wrong Re-ID in situations of continuous appearance change of the target. Here, we combine the global descriptor with an online learning classifier to improve our target Re-ID ability by training with additional appearance samples generated through tracking.

SOT [12], [13], [21] tries to keep tracking a region of the target by searching for it in the neighborhood of the previously detected target region. It is a task including box regression and target classification. ATOM [13] first integrates online target classification learning into SOT to distinguish the target and its background. However, it cannot perform well when the target disappears and reappears. Here, we perform online target classification learning to distinguish the target and other distracting people instead of the target and the background.

In both ATOM [13] and exiting MPF works [2], the online classifier is trained by features computed from recent target observations. However, it will easily cause over-fitting, which would affect Re-ID performance in situations where appearance change frequently happens. The problem is also called *catastrophic forgetting* [22] in the deep learning literature. To alleviate it, *experience replay* has been proposed by randomly

adding historic samples to the recent sample set [23]–[25]. Inspired by this idea, we use a sampling strategy to construct a sample set consisting of the latest features and historic features of the target from past observations, to help build a training dataset that can effectively overcome the over-fitting problem.

III. METHODOLOGY

A. Main Components of Our MPF

The overall MPF framework is shown in Figure 2. Its modules and their effects are as follows upon the selection of the target, a step whose solution is application-dependent:

1) **People Tracking:** Detect and track people based on the current measurements and the tracks of the last timestep.
2) **Target Representation:** Extract people's features using a feature extractor and construct a training sample set by a sampling strategy.
3) **Target Re-ID:** Classify and re-identify the target based on a target Re-ID logic; export the target position and train the classifier if the target is found.
4) **Robot Controller:** Control the mobile robot to follow the target based on the target position.

B. Width-based People Tracking

In order to track people at a close distance, we design a *width-based* tracking module. Such a module is superior to the *height-based* people tracking because it can detect people and estimate their positions without requiring observation of the full body. Here, we use a Kalman filter to realize our *width-based* tracking. Our method takes advantage of the assumption of a known body width of people, which can be easily satisfied in practice.

Supposing extrinsic and intrinsic parameters are known as $\mathbf{R}_w^r, \mathbf{t}_w^r$ from world frame to robot frame, $\mathbf{R}_r^c, \mathbf{t}_r^c$ from robot frame to camera frame, and f_x, c_x of camera intrinsic parameters. The raw measurement is defined as: $\mathbf{b}_k = [u_{tl}\ v_{tl}\ u_{br}\ v_{br}]^T$, where tl and br mean top-left and bottom-right point of the bounding box respectively. Here, a person state in the world frame is defined as: $\mathbf{s}_k = [x_k\ y_k\ \dot{x}_k\ \dot{y}_k]^T$ consisting of position and velocity states. A constant velocity model is assumed to predict the state.

For estimating the distance between the person and the camera, we make a hypothesis that the target person is a cylinder with r radius in the direct front of the robot. We can then get the distance of a person z^c in the camera frame (detailed derivation and discussion are provided in Appendix VI-A):

$$z^c = f_x \cdot \frac{r}{u_{br} - u_{tl}} \tag{1}$$

Then supposing $\bar{\mathbf{s}} = [x\ y\ z]^T$, combining with Equation 1, and according to ridgy body transformation and projective transformation, we can get the observation equations that relate the box bounding variables of a tracked person and

the person's position as follows.:

$$f_x \cdot \frac{(\mathbf{R}_r^c(\mathbf{R}_w^r\bar{\mathbf{s}} + \mathbf{t}_w^r) + \mathbf{t}_r^c)|_x}{(\mathbf{R}_r^c(\mathbf{R}_w^r\bar{\mathbf{s}} + \mathbf{t}_w^r) + \mathbf{t}_r^c)|_z} + c_x = \frac{u_{tl} + u_{br}}{2} \tag{2a}$$

$$(\mathbf{R}_r^c(\mathbf{R}_w^r\bar{\mathbf{s}} + \mathbf{t}_w^r) + \mathbf{t}_r^c)|_z = f_x \cdot \frac{r}{u_{br} - u_{tl}}, \tag{2b}$$

where $|_x$ means the x value of the point. $|_z$ is similar.

This derivation results in a linear observation model, whose details are shown in Appendix VI-B. Supposing the expected obervation is $\mathbf{o}_k$. Through the linear observation model, we can get $\mathbf{o}_k$. To establish the data association between $\mathbf{o}_k$ and a raw measurement $\mathbf{b}_k$, we need to change $\mathbf{b}_k$ to the form as $\mathbf{o}_k$. Therefore, our *processed measurement* $\mathbf{y}_k$ is as follows:

$$\mathbf{y}_k = \begin{bmatrix} \frac{r \cdot (u_{tl,k} + u_{br,k} - 2c_x)}{2(u_{br,k} - u_{tl,k})} - (\mathbf{t}_r^c|_x)^2 - [1\ 0\ 0]\mathbf{R}_r^c\mathbf{t}_w^r \\ \frac{f_x \cdot r}{u_{br,k} - u_{tl,k}} - (\mathbf{t}_r^c|_z)^2 - [0\ 0\ 1]\mathbf{R}_r^c\mathbf{t}_w^r \end{bmatrix}, \tag{3}$$

where only bounding box information is required to be measured. Thus, we can obtain the measurements at a close distance without requiring full-body observation.

Due to inaccuracy of the detected bounding boxes when two people overlap, we keep bounding box of a person only if its largest IoU with other boxes is smaller than a threshold δ_{iou}.

$$\mathbf{B}_k = \{\mathbf{x}_i | f(\mathbf{x}_i, \bar{\mathbf{B}}_k) < \delta_{iou}, \mathbf{x_i} \in \mathbf{B}_k, \bar{\mathbf{B}_k} = \mathbf{B}_k \setminus \{\mathbf{x}_i\}\} \tag{4a}$$

$$f(\mathbf{c}, \mathbf{Q}) = \max_{\mathbf{q}_i \in \mathbf{Q}} IoU(\mathbf{c}, \mathbf{q}_i), \tag{4b}$$

where $\mathbf{B}_k$ is the set of raw measurements at k timestep.

After that, we can use Equation 3 to get the *processed measurement*. Here, we calculate the Euclidean distance between the *processed measurement* and the expected observation, so our distance metric is:

$$d(i, j) = ||\mathbf{o}^i - \mathbf{y}^j||_2^2 \tag{5}$$

Then, a GNN (global nearest neighbor) is used to match the processed measurements to the predicted Kalman states. After Kalman updates, we can get updated people's states.

The tracks information is then added to the people information for target Re-ID. Besides, the people information also contains corresponding image patches and boxes information.

C. Target Representation and Classifier

With the bounding boxes of detected people in the current view of the robot camera, we first extract their features by a pre-trained CNN. Here, we choose to use a global descriptor as in DeepSORT [18] in order to overcome the weakness of a local descriptor such as [2] so that our target Re-ID module can handle distracting people of similar appearance to our target.

In addition, we adopt the online learning module in [2] in order to handle the continuous changes of the target appearance with respect to viewpoint and lighting conditions as well as to take advantage of the additional appearance

samples generated through tracking. Instead of the Bayes classifier used in [2], we use a ridge regression model with L2 regularization as our online learning classifier. Such a regularization-based classifier is able to alleviate the overfitting problem caused by the limited numbers of the training set.

In the meantime, inspired by *experience replay* proposed in [23]–[25], we construct a training sample set containing the latest features and historic features to mitigate the overfitting problem caused by the lack of diversity in the latest samples. Historic features are selected based on the people information from the tracking module.

D. Target Re-ID and Robot Controller

For a complete MPF system, we also need to provide the Re-ID logic and the robot controller here. Our Re-ID logic is mainly based on [2]. In every frame, the classifier would predict the score of the target. If the score is lower than a threshold δ_{switch}, then the system will turn to *Re-ID* state for judging an id-switch is happening. If the target id is lost, it will also lead to *Re-ID* state. In *Re-ID* state, all candidates will be predicted by the classifier. The candidate will be judged as the target if its predicted score is larger than a threshold δ_{id} in N_{id} consecutive frames.

Subsequently, a proportional-integral-derivative controller is used for the robot control. Specifically, in the robot frame, we control the robot by maintaining a given x value and reducing y value to be zero for stable distance estimation by Equation 1.

IV. EXPERIMENTS SETUP

A. Datasets and Evaluation Metrics

Here, we use three datasets in the experiments. One is for the evaluation of our *width-based* tracking module, and the others two are used for the target Re-ID evaluation of the whole MPF system. The first dataset consists of sequential frames and the poses of the target person and the following robot in every frame whose poses are collected by a motion capture (MC) system. Five sequences are collected by walking in the front of the robot within 0.5 m - 7.0 m to allow us to improve the repeatability of the experimental results.

Two other datasets consist of only image sequences. One is a public person following dataset [7]. It contains 11 sequences that are captured by a stereo camera with challenging target Re-ID situations involving illumination change and clothes change. Another dataset is a custom-built dataset, which is designed to fill the gap of the public dataset, for its lack of challenging situations including long-term people occlusion, frequent distance change and similar clothes. It contains four sequences named as *corridor1, corridor2, lab_corridor* and *room*.

Their attributes and corresponding degrees are listed in Table I and some examples of these sequences are shown in Figure 3. *corridor1*, *corridor2* and *lab_corridor* are with dissimilar appearance of upper bodies and similar lower bodies,
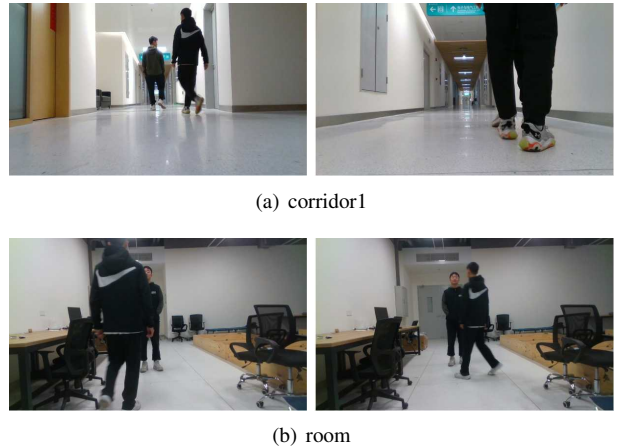

(a) corridor1

(b) room

Fig. 3. Examples of a custom-built dataset with challenging situations including long-term people occlusion, frequent distance change and similar clothes.

TABLE I

ATTRIBUTES OF THE CUSTOM-BUILT DATASET. MORE $+$ MEANS GREATER DEGREE, AND $-$ MEANS THAT THIS ATTRIBUTE IS NOT INVOLVED IN THE DATASET.

	corridor1	*corridor2*	*lab_corridor*	*room*
Similarity	+	+	+	++
Long-term occlusion	+++	+	+	++
Short-term occlusion	-	++	++	-
Distance change	+	++	+	+

while *room* is captured with totally similar appearance. *corridor2* and *lab_corridor* have only one long-term occlusion, and *corridor1* and *room* have two times, but the occlussion of *corridor1* is more serious. For short-term occlusion, mutual crossing exists in *corridor2* and *lab_corridor*, a situation that does not occur in *corridor1* and *room*. In addition, distance change occurs in all sequences.

In the last two datasets, we evaluate the Re-ID capability in terms of accuracy of target person localiztion in the image space. In each frame, if the distance between the center of the ground truth box and the center of the estimated target person region is smaller than a threshold, we regard Re-ID as being successful.

B. Baselines and Our Method

In the public dataset, we compare the proposed method with the SOTA method [2] named as *HEIGHT_CCF* consisting of a *height-based* tracking module and a *CCF* Re-ID module, and other methods reported in [7] including *OAB* [26], *ASE* [27], *SOAB* [28], *DS-KCF* [29], *CNN_v1*, *CNN_v2* and *CNN_v3* [7]. These reported methods are SOT-based methods (*OAB* and *ASE*) and SOT-based methods combining with stereo camera (*SOAB* and *DS-KCF*) or RGBD camera (*CNN_v1*, *CNN_v2* and *CNN_v3*). To compare *height-based* and *width-based* tracking method in a fair way, we evaluate a method called *WIDTH_CCF*, which consists of a *width-based* tracking module and a CCF Re-ID module. While the proposed *GRR* is involved in *WIDTH_GRR*.
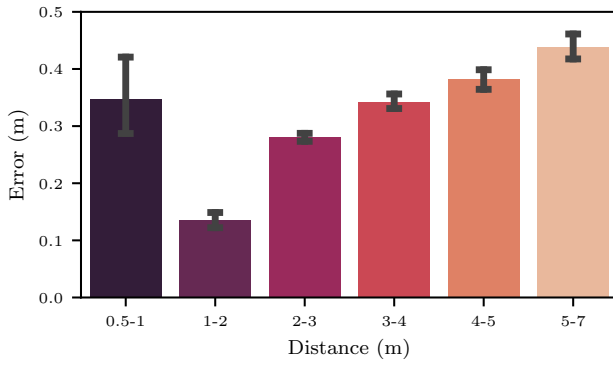
Fig. 4. Box plot of error respect to distance, where distance is the ground truth distance from the target to the robot captured by our MC system, and the error come from the ground truth distance and the estimated distance. In 0.5-1 m, tracking mean error is almost 0.35 m. From 1 m to 7 m, the error is getting larger.

The custom-built dataset is used to compare the effectiveness of the Re-ID modules including *CCF* (SOTA) and ours. The proposed Re-ID method and the *CCF* are evaluated combined with a *width-based* tracking module for a fair comparison. Furthermore, for revealing the influence of the size of the sample set on the effectiveness of the online learning Re-ID module, we make a study in terms of different sizes of the sample set. So *CCF* modules with 16, 32, 64 and 128 sizes of the sample set are named as *CCF_16*, *CCF_32*, *CCF_64* and *CCF_128* respectively. Similarly, *GRR_ST_16*, *GRR_ST_32*, *GRR_ST_64* and *GRR_ST_128* are corresponding to the different sizes setting of *GRR* with short-term sample set. The proposed method with both the long-term and short-term samples is named as *GRR_SLT_64*, which consists of *GRR* and the sampling strategy.

C. Implementation Details

Our people detection model is YOLOX [30] and feature extraction model[1] is similar to DeepSORT [18] whose global descriptor has a dimension of 512. In people tracking, $\delta_{iou} = 0.5$. In target identification, $N_{id} = 5$, $\delta_{switch} = 0.35$ and $\delta_{id} = 0.60$.

A Clearpath Dingo-O, a Realsense D435i with 1280×720 and 30Hz, and a laptop with Intel(R) Core(TM) i5-10200H CPU @ 2.40GHz and NVIDIA GeForce RTX 1650 are used in the person following procedure. All the datasets stored in rosbag format on a computer with Intel® Core™ i7-10700F CPU @ 2.90GHz and NVIDIA GeForce RTX 2060.

V. RESULTS & DISCUSSION

A. Effectiveness of Width-based Person Following

1) Accuracy of our width-based tracking module: As shown in Figure 4, the total estimation error of our tracking module is smaller than 0.5 m, which is appropriate for the MPF. The mean error when the following distance is between 0.5 and 1 meter is larger than that when the following distance is between 1 and 2 meters, 2 and 3

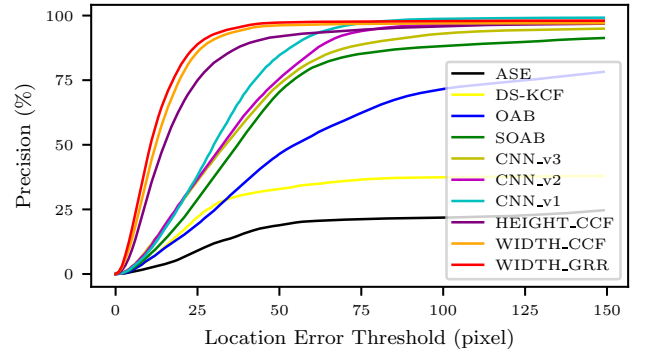[1]https://github.com/pmj110119/YOLOX_deepsort_tracker

Fig. 5. Plot of precision respect to different location error thresholds. *WIDTH_CCF* performs better than *HEIGHT_CCF*, which indicates *width-based* tracking could achieve better performance.
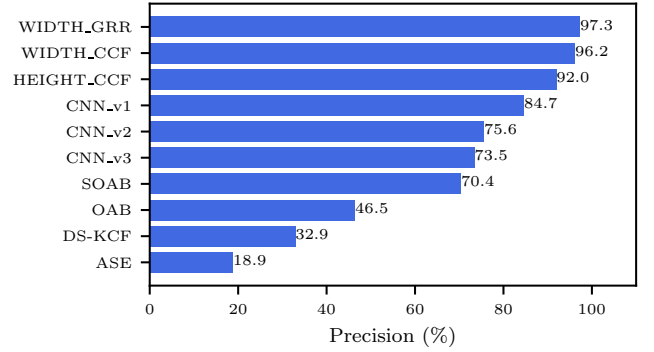

Fig. 6. Precision in location error threshold of 50 pixels. *WIDTH_CCF* is better than *HEIGHT_CCF* with 96.2% vs. 92.0%, and our proposed method *WIDTH_GRR* achieves the best result of 97.3% precision.

meters, and 3 and 4 meters, and its variance is larger than all other situations. This phenomenon could be attributed to the difference between the real person and our hypothetical model of a person. We assume that the person is a cylinder with radius r, where r is measured by the upper body; however, the radius of the upper body is different from the lower body. Thus, our distance estimation method could be further improved by a more precise person model, but the error below 0.5 m is enough for a simple application. When the following distance between 1 and 7 meters, the error gets larger as the distance gets larger. This could be attributed to the bias of our distance estimation, where the target is assumed to be directly in front of the robot, but our walking direction is random in the experiment. So the bias would be larger as the distance gets larger based on the projection theory. But this bias can be mitigated by our controller.

2) Effectiveness of our width-based tracking module: From Figure 5, we can observe that *WIDTH_CCF* surpasses *HEIGHT_CCF* and other reported methods with a larger precision. As shown in Figure 6, the precision of *WIDTH_-CCF* is 96.2% versus 92.0% of *HEIGHT_CCF* at 50 pixel threshold. This result is mainly attributed to the effect of the *width-based* tracking method. An example is shown in Figure 7. The left image is the image which is failed to be detected and tracked by *HEIGHT_CCF* without observation
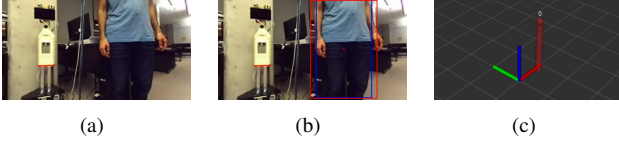
Fig. 7. (a) is the failed case of *HEIGHT_CCF* for its failed observation of full body at close distance. (b) is our result with a red box and the ground truth with a blue box. (c) is the result of our method in the robot frame.

of the person's full body or neck. The middle image is our detection result with a red box and the ground truth with a blue box. The right image is our tracking result of the target person in the robot frame. Thus, such a good performance in the public dataset depends to a large extent on the *width-based* tracking, which can utilize the boxes of the people to realize the people tracking even at a close distance. From the above observation and analysis, we can conclude that our *width-based* tracking module is beneficial for the MPF in the situations of distance change for its successful tracking even at a close distance.

And our *WIDTH_GRR* is a little better than *WIDTH_CCF* in terms of 97.3% versus 96.2% for its better Re-ID capability in situations of distance change between the people and the robot, which would be further discussed in the following.

B. Performance of our Re-ID Module

1) CCF vs. GRR_ST: Here, we compare the Re-ID ability of *CCF* and *GRR_ST* on our dataset and the best results of them are selected to be compared. From Table II, we can see that *GRR_ST* outperform *CCF* in all sequences, by 6.4% in *corridor1*, 9.4% in *corridor2*, 2.2% in *lab_corridor* and 62.7% in *room*. We can observe that high similarity is fatal to *CCF* for it only gets 34.5% mAP in *room*, where the target is wrongly re-identified when the first occlusion happened. Oppositely, *GRR_ST* acts well after two long-term occlusions. *CCF* performs badly because it only uses a sum value from a region of a low-level feature map as its feature value, which would cause confusion for the Naive Bayes classifier in the situations of high similarity. While *GRR_ST* uses a global descriptor as its feature vector, which not only integrates appearance information but also contains the spatial relation information of a person's parts, leading to a more discriminative representation of the target. So the classifier is able to distinguish the target from distracting people in high similarity cases.

In *corridor2* and *lab_corridor*, both *CCF* and *GRR_ST* perform well, where *GRR_ST* is a little better with 98.4% vs. 89.0% of *CCF* in *corridor2* and 92.7% vs. 84.4% in *lab_corridor*. *GRR_ST* can find the target once the target occurs without any hesitation, while *CCF* can find the target again after occlusion until the discriminative upper body appeared in the image.

In *corridor1* in which long-term occlusion is the most severe, both methods act well after the first occlusion for the reason that the discriminative part of the body could be observed. But both of them are failed to re-identify the target in the second occlusion. Before the occlusion happened, only

	corridor1	*corridor2*	*lab_corridor*	*room*
CCF_16	55.0	89.0	32.6	32.7
CCF_32	53.4	80.6	90.5	32.6
CCF_64	41.2	39.7	83.0	33.0
CCF_128	53.9	49.9	84.4	34.5
GRR_ST_16	61.4	32.3	89.7	97.0
GRR_ST_32	61.4	37.7	90.6	97.2
GRR_ST_64	61.4	98.3	90.6	97.2
GRR_ST_128	61.0	**98.4**	92.7	35.4
GRR_SLT_64	**99.3**	95.6	**92.8**	**97.2**

lower body parts that are not differentiated could be observed and this process lasts for a long time. So the sample set is full of these confusing samples, which leads to an over-fitting problem.

In conclusion, compared to *CCF*, *GRR_ST* is more discriminative for its better Re-ID ability in *room* of high similarity, and faster Re-ID speed in situations of distance change (*corridor2* and *lab_corridor*). From the above analysis, we can conclude that combining with a high-level global descriptor can help to improve the robustness of target Re-ID for its superior feature representation ability.

2) GRR_ST vs. GRR_SLT: From the sample set size study of *GRR_ST* in Table II, we can observe that the online training of a classifier is sensitive to the sample set. In *corridor2* of severe distance change, *GRR_ST* with only the latest samples (16 and 32 samples) perform badly, which could be attributed to the over-fitting problem. In *room*, the classifier performs poorly when too many old samples (128 samples) are added, while it performs well with the latest samples (16, 32 and 64 samples). These phenomenons indicate that historic observations could be contributed to the classifier. But it's important to answer *how historic* is the long-term samples and *how to select* them. *Experience replay* [23]–[25] randomly select historic samples from previously-seen examples from a "large dataset" that is different from the current samples. Our sampling strategy, as is mentioned in Section III-C, is similar to them. Our "large dataset" is built by the selection of historic observations instead of a large sample set that contains the latest observations.

With the proposed sampling strategy, we can attain 99.3% AP in *corridor1*, 92.8% AP in *lab_corridor* and 97.2% AP in *room* with a size of 64. The outstanding performance in *corridor_1* could be attributed to its historical memory, which can help to get rid of the over-fitting problem caused by the latest observations of the indiscernibility of the lower bodies. But it achieves 95.6% AP in *corridor2* vs. 98.3% AP of *GRR_ST_64*, where the Re-ID speed of *GRR_SLT_64* is slower. This means that our sampling strategy can be further improved.

Overall, our sampling strategy can help the classifier alle-

viate the over-fitting effects by adding historic observations. Similar to the effect of a global descriptor, samples with diversity can also help to construct a high-level representation of the target, which is useful for target Re-ID.

VI. CONCLUSION

In this paper, we propose a MPF system with a *width-based* tracking module and a robust target Re-ID module. The results of experiments about the *width-based* tracking module indicate that this module is accurate for the person following with an overall error lower than 0.5m. And most importantly, it can track the target even at a close distance because our *width-based* tracking module can track people without requiring full-body observation.

Also, our method achieves the best results in a custom-built dataset, which is beneficial from the discriminative ability of a high-level global descriptor. Besides, the historic samples selected by the sampling strategy also help to describe the target at a high-level representation to alleviate the over-fitting problem.

In the future, we will further improve the target Re-ID ability of our system by integrating the graph information of the target and the people or body parts of the target.

APPENDIX

A. Distance Estimation

Human boxes coordinates are defined by two endpoints: $(u_{tl}, v_{tl}), (u_{br}, v_{br})$, and the corresponding points in the camera frame are $(x_{tl}^c, y_{tl}^c, z_{tl}^c)$ and $(x_{br}^c, y_{br}^c, z_{br}^c)$ respectively. And we suppose the target is in the directly front of the camera (this can be realized by the robot controller module), so we have $|x_{tl}^c - x_{br}^c| = r$ and $z_{tl}^c = z_{br}^c = z^c$. With camera projection equations:

$$u_{tl} = f_x \cdot \frac{x_{tl}^c}{z_{tl}^c} + c_x \tag{6a}$$

$$u_{br} = f_x \cdot \frac{x_{br}^c}{z_{br}^c} + c_x, \tag{6b}$$

then subtracting Equation 6a to Equation 6b, and substituting above conditions, finally we can get Equation 1.

B. Linear Observation Model

Assembling Equation 2a to Equation 2b, and multiplying out, we can get:

$$(\mathbf{R}_r^c \mathbf{R}_w^r \bar{\mathbf{s}} + \mathbf{R}_r^c \mathbf{t}_w^r)|_x + (\mathbf{t}_r^c|_x)^2 = \frac{|u_{tl} + u_{br}|/2 - c_x}{f_x} \tag{7a}$$

$$(\mathbf{R}_r^c \mathbf{R}_w^r \bar{\mathbf{s}} + \mathbf{R}_r^c \mathbf{t}_w^r)|_z + (\mathbf{t}_r^c|_z)^2 = f_x \cdot \frac{r}{|u_{tl} - u_{br}|}, \tag{7b}$$

according to the hypothesis and definitions in Section III-B, setting $z = 0$ (supposing the person is a point on the x-y plane), we have:

$$\begin{bmatrix} r_{00} & r_{01} \\ r_{20} & r_{21} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$
$$= \begin{bmatrix} \frac{r \cdot (u_{tl} + u_{br} - 2c_x)}{2(u_{br} - u_{tl})} - (\mathbf{t}_r^c|_x)^2 - [1\ 0\ 0]\mathbf{R}_r^c \mathbf{t}_w^r \\ \frac{f_x \cdot r}{u_{br} - u_{tl}} - (\mathbf{t}_r^c|_z)^2 - [0\ 0\ 1]\mathbf{R}_r^c \mathbf{t}_w^r, \end{bmatrix}, \tag{8}$$

change $[x\ y]^T$ to $\mathbf{s} = [x\ y\ \dot{x}\ \dot{y}]^T$, fill 0 to the left matrix, and define $\mathbf{R}_w^r \mathbf{R}_r^c = [r_{00}\ r_{01}\ r_{02}; r_{10}\ r_{11}\ r_{12}; r_{20}\ r_{21}\ r_{22}]$, then we can get our linear observation model:

$$\begin{bmatrix} r_{00} & r_{01} & 0 & 0 \\ r_{20} & r_{21} & 0 & 0 \end{bmatrix} \mathbf{s}$$
$$= \begin{bmatrix} \frac{r \cdot (u_{tl} + u_{br} - 2c_x)}{2(u_{br} - u_{tl})} - (\mathbf{t}_r^c|_x)^2 - [1\ 0\ 0]\mathbf{R}_r^c \mathbf{t}_w^r \\ \frac{f_x \cdot r}{u_{br} - u_{tl}} - (\mathbf{t}_r^c|_z)^2 - [0\ 0\ 1]\mathbf{R}_r^c \mathbf{t}_w^r, \end{bmatrix} \tag{9}$$

REFERENCES

[1] M. J. Islam, J. Hong, and J. Sattar, "Person-following by autonomous robots: A categorical overview," *The International Journal of Robotics Research*, vol. 38, no. 14, pp. 1581–1618, 2019.

[2] K. Koide, J. Miura, and E. Menegatti, "Monocular person tracking and identification with on-line deep feature selection for person following robots," *Robotics and Autonomous Systems*, vol. 124, p. 103348, 2020.

[3] A. Leigh, J. Pineau, N. Olmedo, and H. Zhang, "Person tracking and following with 2d laser scanners," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 726–733.

[4] Y. Sung and W. Chung, "Hierarchical sample-based joint probabilistic data association filter for following human legs using a mobile robot in a cluttered environment," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 3, pp. 340–349, 2015.

[5] J. Yuan, S. Zhang, Q. Sun, G. Liu, and J. Cai, "Laser-based intersection-aware human following with a mobile robot in indoor environments," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 1, pp. 354–369, 2018.

[6] M. Wang, D. Su, L. Shi, Y. Liu, and J. V. Miro, "Real-time 3d human tracking for mobile robots with multisensors," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 5081–5087.

[7] B. X. Chen, R. Sahdev, and J. K. Tsotsos, "Integrating stereo vision with a cnn tracker for a person-following robot," in *International Conference on Computer Vision Systems*. Springer, 2017, pp. 300–313.

[8] 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.

[9] T. Linder, S. Breuers, B. Leibe, and K. O. Arras, "On multi-modal people tracking from mobile platforms in very crowded and dynamic environments," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 5512–5519.

[10] M. Zhang, X. Liu, D. Xu, Z. Cao, and J. Yu, "Vision-based target-following guider for mobile robot," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9360–9371, 2019.

[11] L. Zheng, M. Tang, Y. Chen, G. Zhu, J. Wang, and H. Lu, "Improving multiple object tracking with single object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2453–2462.

[12] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "Siamrpn++: Evolution of siamese visual tracking with very deep networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4282–4291.

[13] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "Atom: Accurate tracking by overlap maximization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4660–4669.

[14] W. Choi and S. Savarese, "Multiple target tracking in world coordinate with single, minimally calibrated camera," in *European Conference on Computer Vision*. Springer, 2010, pp. 553–567.

[15] I. Ardiyanto and J. Miura, "Partial least squares-based human upper body orientation estimation with combined detection and tracking," *Image and Vision Computing*, vol. 32, no. 11, pp. 904–915, 2014.

[16] D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," *International Journal of Computer Vision*, vol. 80, no. 1, pp. 3–15, 2008.

[17] H. Grabner and H. Bischof, "On-line boosting and vision," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, vol. 1. Ieee, 2006, pp. 260–267.

[18] 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.

[19] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "Fairmot: On the fairness of detection and re-identification in multiple object tracking," *International Journal of Computer Vision*, vol. 129, no. 11, pp. 3069–3087, 2021.

[20] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "Mot16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831*, 2016.

[21] C. Mayer, M. Danelljan, D. P. Paudel, and L. Van Gool, "Learning target candidate association to keep track of what not to track," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 444–13 454.

[22] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[23] M. Knowles, V. Peretroukhin, W. N. Greene, and N. Roy, "Toward robust and efficient online adaptation for deep stereo depth estimation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 12 192–12 198.

[24] Y. Kuznietsov, M. Proesmans, and L. Van Gool, "Comoda: Continuous monocular depth adaptation using past experiences," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2907–2917.

[25] Z. Zhang, S. Lathuiliere, E. Ricci, N. Sebe, Y. Yan, and J. Yang, "Online depth learning against forgetting in monocular videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4494–4503.

[26] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting." in *Bmvc*, vol. 1, no. 5. Citeseer, 2006, p. 6.

[27] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *British Machine Vision Conference, Nottingham, September 1-5, 2014*. Bmva Press, 2014.

[28] 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.

[29] M. Camplani, S. L. Hannuna, M. Mirmehdi, D. Damen, A. Paiement, L. Tao, and T. Burghardt, "Real-time rgb-d tracking with depth scaling kernelised correlation filters and occlusion handling." in *BMVC*, vol. 3, 2015.

[30] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.