

# Short Papers

---

## A Robust Visual Person-Following Approach for Mobile Robots in Disturbing Environments

Lei Pang<sup>ID</sup>, Zhiqiang Cao<sup>ID</sup>, Senior Member, IEEE, Junzhi Yu<sup>ID</sup>, Senior Member, IEEE, Peiyu Guan, Xuechao Chen<sup>ID</sup>, and Weimin Zhang<sup>ID</sup>

**Abstract**—This article proposes a robust visual following approach with a deep learning-based person detector, a Kalman filter (KF), and a reidentification module. The KF is introduced to predict the position of the target person, and its state is updated by the associated detection result. To deal with severe distractions and even full occlusion, the reidentification module with an identification model, a verification model, and an appearance gallery is employed in multi-person disturbing environments. Without any customized markers, the proposed approach can follow the target person steadily, and it is robust to occlusion and posture changes of the target person. Experiments results validate the effectiveness of the proposed approach.

**Index Terms**—Kalman filter (KF), mobile robot, person detector, person-following, reidentification.

### I. INTRODUCTION

The person following is an important ability for mobile robots to cooperate with human. It has many potential applications, such as elder person assistance [1], medical service support [2], robotic suitcase [3], robotic wheelchair [4], robotic shopping cart [5], etc.

The researchers have utilized different sensors to the task of person following. Compared with the solutions employing pyroelectric sensor [6], two-dimensional (2-D) laser scanner [4], ultrasonic sensor [7], LIDAR [8], vision-based solutions [9], [10] are more effective and promising. Many visual trackers, such as kernelized correlation filters (KCF) [11], tracking-learning-detection (TLD) [12], and long-term correlation tracking (LCT) [13], are proposed and they endeavor to

Manuscript received March 8, 2019; revised July 28, 2019; accepted September 10, 2019. Date of publication October 8, 2019; date of current version June 3, 2020. This work was supported in part by the Beijing Advanced Innovation Center for Intelligent Robots and Systems under Grant 2018IRS21, in part by the National Natural Science Foundation of China under Grants 61633017, 61633020, and 61836015, in part by the Key Research and Development Program of Shandong Province under Grant 2017CXGC0925, and in part by the Open Foundation of the State Key Laboratory of Management and Control for Complex Systems, CASIA under Grant 20190106. (Corresponding author: Zhiqiang Cao.)

L. Pang, Z. Cao, and P. Guan are with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100049, China (e-mail: panglei2015@ia.ac.cn; zhiqiang.cao@ia.ac.cn; guanpeiyu2017@ia.ac.cn).

J. Yu is with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the State Key Laboratory for Turbulence and Complex System, Department of Mechanics and Engineering Science, BIC-ESAT, College of Engineering, Peking University, Beijing 100871, China (e-mail: junzhi.yu@ia.ac.cn).

X. Chen and W. Zhang are with the Beijing Advanced Innovation Center for Intelligent Robots and Systems, Beijing 100081, China, and also with the Intelligent Robotics Institute, Beijing Institute of Technology, Beijing 100081, China (e-mail: chenxuechao@bit.edu.cn; zhwm@bit.edu.cn).

Digital Object Identifier 10.1109/JSYST.2019.2942953

establish an appearance model of the target. In recent years, the trackers based on convolutional neural networks (CNN) [14]–[17] become a research hotspot. In the short run, visual trackers can track the target with high accuracy. However, they tend to fail in disturbing environments with partial/full occlusion of the target.

By analyzing the depth distribution and the response of the KCF tracker, occlusion is detected and handled in [18] and [19]. A problem of such solution is that its robustness depends on the accuracy of the captured depth map. Also, the response of KCF with HoG feature is not reliable. Human body information including facial feature [20], [21] offers abundant solutions. Sun *et al.* proposed a method to recognize the lost target person based on the soft biometrics features including color of clothes and body size [22]. Koide *et al.* identified the target person by the integration of height, gait, and color features [23]. These methods with body information suffer from the postures, gaits of the target person as well as illumination conditions. With the customized markers [8] and radio-frequency identification (RFID) tags [24] sticking on the target, a better identification can be achieved; however, these markers are also constrained by the postures of the target. How to identify the target person in disturbing environments still remain unsolved.

In this article, a vision-based person-following approach is proposed for mobile robots. The proposed approach combines a person detector, a Kalman filter (KF) with a reidentification module using a single stereo camera fixed on the mobile robot. The KF is introduced, and its state is updated by the associated detection result without feature extraction and online model learning. More importantly, the reidentification is employed to determine the identities of the detection results in multi-person disturbing scenarios. The proposed approach does not require any customized markers, and it possesses a good adaptability to disturbing environments.

### II. PROPOSED VISUAL FOLLOWING APPROACH

#### A. Overview

In this article, a stereo camera is employed as the only sensor to simultaneously provide RGB image and depth image, which is installed along the heading direction of the mobile robot. The robot coordinate system is considered to be the same as the left camera coordinate system  $OX^C Y^C Z^C$ . For  $OX^C Y^C Z^C$ , its original point  $O$  is the optical center of the left camera, and its  $z$ -axis is along the optical axis of the left camera.

Let  $B_t$  be the target's bounding box at frame  $t$ . The distance  $D_t$  between the target person with the mobile robot can be estimated according to the position of  $B_t$  on the depth image attached to the left RGB image. The horizontal central coordinate  $P_t$  of  $B_t$  is regarded as the position of the target person, and the horizontal direction angle  $\theta_t$  of the target person with respect to the  $Z^C$ -axis in  $OX^C Y^C Z^C$  is

computed by  $\theta_t = \frac{(P_t - R_h/2)}{R_h} \Theta$ , where  $R_h$  and  $\Theta$  are the resolution and the field of view of the left camera in horizontal direction, respectively. With the angle  $\theta_t$  and the distance  $D_t$ , the mobile robot is then controlled to keep the following to its specified person, where the target's bounding box  $B_t$  plays a crucial role.

### B. Person Detector

In this article, YOLOv2 [25] is applied as the person detector, and it outputs the bounding boxes  $B^d = \{B_i^d | i = 1, 2, \dots, N^d\}$  of all persons at frame  $t$ , where  $N^d$  is the number of detected persons. YOLOv2 employs a deep network of Darknet-19 with 19 convolutional layers and five maxpooling layers. Besides, some strategies such as batch normalization, anchor boxes, dimension cluster, and multi-scale training are adopted to improve the performance of object detection. To accomplish person following using visual approach, a primary task is to locate the target person in images. By a well-trained person detector, all persons can be accurately located in an image, which provides the basis to find the target person.

Although the persons can be detected without ego-motion information, the identity of each detected person cannot be obtained. The position of the target person can be located by considering the data association of the detection results between consecutive frames. Hence, besides the person detector, a KF coupled with a reidentification module is designed to solve the problem of data association.

### C. Target Localization With KF and Reidentification Module

**1) KF Prediction:** In order to associate the results of the person detector as well as alleviate the influence of the short-time missing detections, a KF is employed to predict the position of the target person. The state of KF is modeled as  $x = [\beta, \dot{\beta}]^T$ , where  $\beta = [c_x, c_y, a, r]^T$  denotes the information of the target's bounding box,  $\dot{\beta}$  indicates the changing rate of  $\beta$ ,  $(c_x, c_y)$  is the center coordinate of the bounding box,  $a$  and  $r$  represent the area and the aspect ratio of the bounding box, respectively.

A constant velocity model independent of robot motion is adopted to predict the new position of the target person at frame  $t$ , and KF predicts the new state  $x^p$  by

$$x^p = \begin{bmatrix} \beta_t \\ \dot{\beta}_t \end{bmatrix} = \begin{bmatrix} \beta_{t-1} \\ \dot{\beta}_{t-1} \end{bmatrix} + \begin{bmatrix} \Delta t \times \dot{\beta}_{t-1} \\ O_{4 \times 1} \end{bmatrix} \quad (1)$$

where  $\Delta t$  is the interval of frames between two predictions. Then, the predicted target's bounding box can be computed according to  $\beta_t$ , which is labeled as  $B^p$ . Note that the initial changing rate is set to zero, i.e.,  $\dot{\beta}_0 = [0, 0, 0, 0]^T$ , and its corresponding covariance is initialized with large values.

**2) Distraction Judgment:** A rectangle region  $R^{dp}$  around  $B^p$  is set to judge the occurrence of the disturbing pedestrians, where  $R^{dp}$  has the same height as  $B^p$  with a larger width  $W$ . Then, the number  $N^{dp}$  that the detection results intersect with  $R^{dp}$  is counted. Once  $N^{dp} > 1$  is satisfied during successive  $F_1^{dp}$  frames, we conclude that the distraction occurs; afterward, if  $N^{dp} \leq 1$  lasts successive  $F_2^{dp}$  frames, the disturbance from pedestrians are considered to be disappeared.

**3) Association Without Disturbing Pedestrians:** For the case where there is no distraction, it is reasonable to judge the association between the detections and prediction of KF with their spatial relationship. Hence, at frame  $t$ , a spatial condition is considered to find the matched measurement, and the spatial score  $S_i^S$  of each  $B_i^d$ ,  $i = 1, 2, \dots, N^d$  is calculated by

$$S_i^S = \frac{\text{area}(B^p \cap B_i^d)}{\text{area}(B^p \cup B_i^d)}, i = 1, 2, \dots, N^d. \quad (2)$$

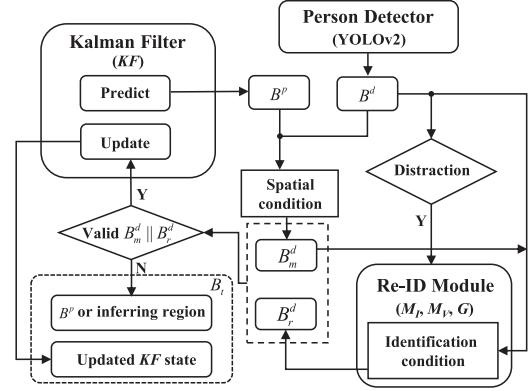


Fig. 1. Schematic flowchart of the proposed approach.

The bounding box with maximum spatial score  $S_m^S = \max\{S_i^S | i = 1, 2, \dots, N^d\}$  and  $S_m^S \geq T^S$  is considered as the matched measurement, which is labeled as  $B_m^d$ , where  $T^S$  is a given threshold.

**4) Distraction Handling With Reidentification Module:** It is not enough to rely only on the criterion of spatial condition when the distraction occurs. Person reidentification technique has been widely applied in the field of video surveillance to find the target person in the scenarios. In this article, a reidentification module is introduced to handle the distraction. Note that in the process of identity verification, the detection matched to the prediction of KF, i.e.,  $B_m^d$ , will be preferentially verified to reduce the computation cost.

Firstly, an identification model  $M^I$  is adopted to extract the features of detected persons, which employs a VGG-16 as the base network [26] and learns discriminative pedestrian descriptors by considering multiple loss functions. The model  $M^I$  is trained with the Market1501 dataset. With  $M^I$ , the identities of detected persons are verified by comparing the cosine similarity of fully connected features.

In practice, relying on a single identification model  $M^I$ , the reliability of retrieving the target person cannot be always guaranteed in complex environments. Herein, a complementary verification model  $M^V$  is introduced to add further valuable information. The model  $M^V$  is a well-designed network [27], which simultaneously learns feature representation and the corresponding similarity metric. It receives a pair of images with the size of  $60 \times 160$  pixels and outputs the similarity probability. In this article,  $5 \times 5$  convolution filters are used to replace the  $3 \times 3$  convolution filters in [27], and batch normalization operations are added after each convolutional layer and the full connected layer. The  $M^V$  is trained with the large-scale person reidentification dataset CUHK03.

In order to improve the robustness of target reidentification, a dynamic appearance gallery  $G$  is set to record multiple appearances of the target person.  $G$  is comprised of  $N$  clipped images  $\{m_1, m_2, \dots, m_N\}$  with corresponding fully connected features  $f = \{f_1, f_2, \dots, f_N\}$ . Note that the gallery shall be updated every  $F^G$  frames only in the cases without distraction.

With the combination of these two models and the gallery, the detected results  $B^d$  are identified. Each bounding box  $B_i^d$ ,  $i = 1, 2, \dots, N^d$  is first fed forward to  $M^I$  to compute corresponding full connected features  $f^I$ . According to the cosine similarities between  $f^I$  and each element in  $f$ , we get the identification score  $S^I$  as follows:

$$S^I = \frac{1}{N} \sum_{j=1}^N \frac{f^I \cdot f_j}{\|f^I\| \|f_j\|}. \quad (3)$$

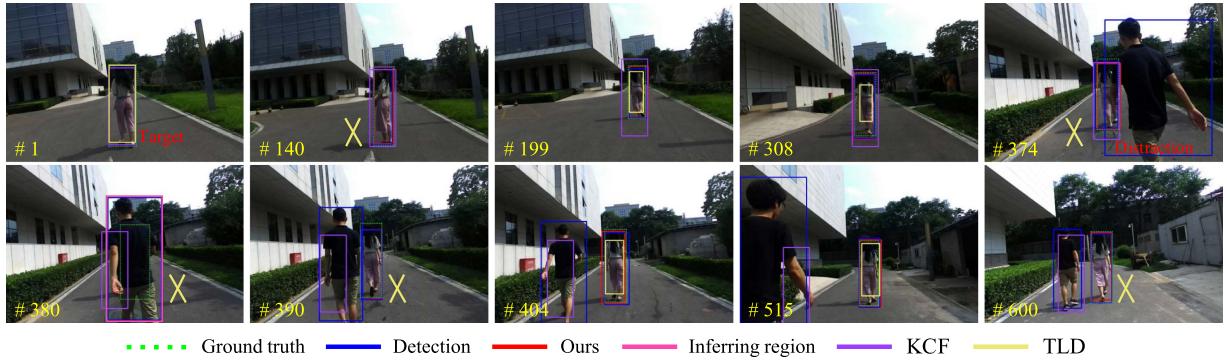


Fig. 2. Results with different frames for our approach, KCF [11] and TLD [12] in the experiment 1, where the label  $\times$  means that there is no output for TLD.

The image corresponding to a bounding box in  $B^d$  and each image in  $G$  are also inputted to  $M^V$  to acquire the similarity probability  $S_j^V$ ,  $j = 1, 2, \dots, N$ . Then, we have the verification score  $S^V$ , and  $S^V = \max\{S_j^V | j = 1, 2, \dots, N\}$ . The one in  $B^d$  that satisfies the following identification condition (4) will be labeled as  $B_r^d$

$$S^V \geq T_2^V \cup S^I \geq T_2^I \cup (S^V \geq T_1^V \cap S^I \geq T_1^I) \quad (4)$$

where  $T_1^I$ ,  $T_2^I$ ,  $T_1^V$ , and  $T_2^V$  are thresholds, and they are set to 0.7, 0.8, 0.4, and 0.6, respectively. Notice that other bounding boxes in  $B^d$ ,  $B_i^d \neq B_m^d$ ,  $i = 1, 2, \dots, N^d$ , shall not be concerned once  $B_r^d = B_m^d$ .

If there is no bounding box in  $B^d$  satisfying (4), the target person usually locates behind the distractive pedestrians. In this case, an inferring region  $B^s$  is designed to serve as the target bounding box  $B_t$ , where  $B^s$  is considered as a rectangle region containing all bounding boxes in  $B^d$  that intersect with  $B^p$ .

5) *KF Update and Target Localization:* If there exists  $B_m^d$  or  $B_r^d$ , KF will be updated using the standard KF equations. Then, the 4-D measured state  $\beta$  is calculated according to the matched bounding box, and the measurement function is shown as follows:

$$z_t = \beta + \eta \quad (5)$$

where  $\eta$  is the  $4 \times 1$  measurement noise. The updated state of KF is used to compute  $B_t$ . When there is no bounding box in  $B^d$  matched with  $B^p$ , the state of KF remains unchanged, and  $B_t = B^p$ .

Note that if KF cannot be updated during consecutive  $T^F$  frames, KF terminates prediction, and it will be reinitialized with the detection result assigned to the target person. The schematic flowchart of the proposed approach is illustrated in Fig. 1.

### III. EXPERIMENTS

The experiments consider a single ZED stereo camera with  $\Theta = 72^\circ$  and  $R_h$  is 1080 pixels.  $F_1^{dp} = 3$ ,  $F_2^{dp} = 20$ ,  $T^F = 3$ , and  $T^S = 0.2$ .  $G$  is updated every  $F^G = 50$  frames with  $N = 3$ .

#### A. Analyses of the Proposed Approach

Fig. 2 gives the results of the experiment 1 where full occlusion and neighboring distraction occur. One can see that a pedestrian causes full occlusion (see frame 380 in Fig. 2). At this point, the inferring region is calculated and its horizontal central coordinate is very close to that of the ground truth. At frame 390, the target person is reconfirmed according to (4) where  $S^V$  and  $S^I$  are 0.65 and 0.84, respectively. Next, another pedestrian walks aside the target person, and the results show that the target person can be correctly tracked.

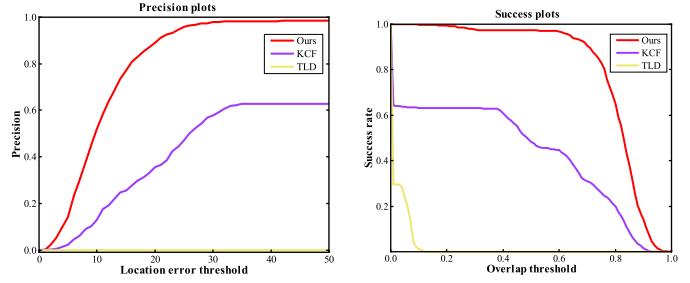


Fig. 3. One-Pass evaluation in the experiment 1 of our approach with KCF [11] and TLD [12].

Besides, Fig. 2 also provides the comparison of our approach with KCF [11] and TLD [12]. The comparison results with the metrics of precision plot and success plot are presented in Fig. 3. The precision plot shows the percentage of successfully tracked frames on which the center location error of a method is within a given threshold. The success plot describes the percentage of successfully tracked frames on which the intersection over union is within a given overlap threshold. For TLD, it fails to detect the target person at about 63% of 600 frames, whereas KCF fails after the occlusion without recovery from the tracking failure. Our approach can always locate the target person with an average speed of 32.4 frame/s.

#### B. Experiment on a Mobile Robot

Experiment 2 is conducted on a mobile robot where disturbing pedestrians exert challenges. Fig. 4 demonstrates snapshots of the experiment 2. The comparison for our approach, KCF [11], and TLD [12] is given in Fig. 5. In the following process, a pedestrian causes partial and full occlusion, as shown in Fig. 4(b). From frame 145 in Fig. 5, the inferring region is regarded as the output. At frame 151, once the target person appears, it is redetected and identified immediately, where  $S_V = 0.999$  and  $S_I = 0.105$  for the target person. Meanwhile, the KF is reinitialized with the verified detection result. After that, another pedestrian with similar wearing exerts a new disturbance as shown in Fig. 4(d). During this process, the target person is still effectively tracked (see frame 610 in Fig. 5). Fig. 5 demonstrates that the proposed approach is superior to the compared methods. With our proposed approach, the robot calculates the localization result and achieves a stable following of the target person with possible posture changes.



Fig. 4. Snapshots of the experiment 2.

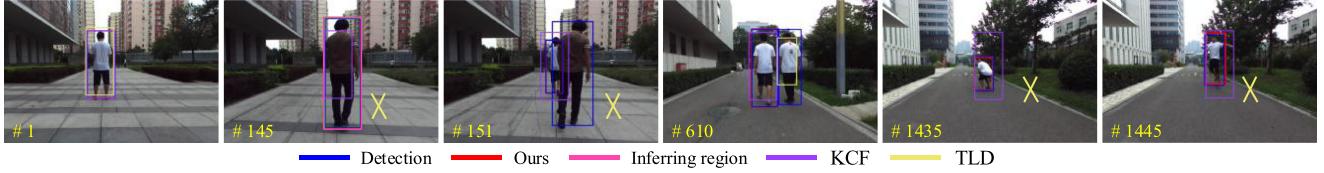


Fig. 5. Comparison results for our approach, KCF [11] and TLD [12] in the experiment 2, where the label  $\times$  means that there is no output for TLD.

#### IV. CONCLUSION

In this article, a vision-based person-following approach for mobile robot was proposed. The combination of a person detector, a KF, and a reidentification module guarantees a reliable following. The experimental results demonstrate the effectiveness of the proposed approach. In the future, we shall combine multiple sensors to further improve the robustness of the proposed approach.

#### REFERENCES

- [1] A. Tomoya, S. Nakayama, A. Hoshina, and M. Sugaya, "A mobile robot for following, watching, and detecting falls for elderly care," *Procedia Comput. Sci.*, vol. 112, pp. 1994–2003, 2017.
- [2] R. Tasaki, H. Sakurai, and K. Terashima, "Moving target localization method using foot mounted acceleration sensor for autonomous following robot," in *Proc. IEEE Conf. Control Technol. Appl.*, 2017, pp. 827–833.
- [3] B. Q. Ferreira, K. Karipidou, F. Rosa, S. Petrisca, P. Alves-Oliveira, and A. Paiva, "A study on trust in a robotic suitcase," in *Proc. Int. Conf. Social Robot.*, 2016, pp. 179–189.
- [4] A. Leigh, J. Pineau, N. Olmedo, and H. Zhang, "Person tracking and following with 2D laser scanners," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2015, pp. 726–733.
- [5] E. Ackerman, "Walmart and five elements robotics working on robotic shopping cart," *IEEE Spectrum*, 2016. [Online]. Available: <https://spectrum.ieee.org/automaton/robotics/industrial-robots/walmart-and-five-elements-robotics-working-on-robotic-shopping-cart>.
- [6] Q. Hao, F. Hu, and Y. Xiao, "Multiple human tracking and identification with wireless distributed pyroelectric sensor systems," *IEEE Syst. J.*, vol. 3, no. 4, pp. 428–439, Dec. 2009.
- [7] D. Su and J. V. Miro, "An ultrasonic/RF GP-based sensor model robotic solution for indoors/outdoors person tracking," in *Proc. Int. Conf. Control Autom. Robot. Vis.*, 2014, pp. 1662–1667.
- [8] M. Perdoch, D. M. Bradley, J. K. Chang, H. Herman, P. Rander, and A. Stentz, "Leader tracking for a walking logistics robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 2994–3001.
- [9] E. Babaian, N. K. Korghond, A. Ahmadi, M. Karimi, and S. S. Ghidary, "Skeleton and visual tracking fusion for human following task of service robots," in *Proc. RSI Int. Conf. Robot. Mechatronics*, 2015, pp. 761–766.
- [10] L. Pang, L. Zhang, Y. Yu, J. Yu, Z. Cao, and C. Zhou, "A human-following approach using binocular camera," in *Proc. IEEE Int. Conf. Mechatronics Autom.*, 2017, pp. 1487–1492.
- [11] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.
- [12] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.
- [13] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Adaptive correlation filters with long-term and short-term memory for object tracking," *Int. J. Comput. Vis.*, vol. 126, no. 8, pp. 771–796, 2018.
- [14] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1144–1152.
- [15] Y. Song, C. Ma, L. Gong, J. Zhang, R. W. H. Lau, and M. Yang, "CREST: Convolutional residual learning for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2574–2583.
- [16] S. Pu, Y. Song, C. Ma, H. Zhang, and M.-H. Yang, "Deep attentive tracking via reciprocative learning," in *Proc. Adv. Neural Inform. Process. Syst.*, 2018, pp. 1931–1941.
- [17] Y. Song *et al.*, "VITAL: Visual tracking via adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8990–8999.
- [18] M. Camplani *et al.*, "Real-time RGB-D tracking with depth scaling kernelised correlation filters and occlusion handling," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 145.1–145.11.
- [19] L. Zhang, Z. Cao, X. Meng, C. Zhou, and S. Wang, "Real-time depth-based tracking using a binocular camera," in *Proc. 12th World Congr. Intell. Control Autom.*, 2016, pp. 2041–2046.
- [20] C. Fahn and Y. Lin, "Real-time face recognition techniques used for the interaction between humans and robots," in *Proc. Int. Comput. Symp.*, 2010, pp. 234–239.
- [21] D. M. Vo, L. Jiang, and A. Zell, "Real time person detection and tracking by mobile robots using RGB-D images," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, 2014, pp. 689–694.
- [22] S. Sun, N. An, X. Zhao, and M. Tan, "Human recognition for following robots with a Kinect sensor," in *Proc. IEEE Int. Conf. Robot. Biomim.*, 2016, pp. 1331–1336.
- [23] K. Koide and J. Miura, "Identification of a specific person using color, height, and gait features for a person following robot," *Robot. Auton. Syst.*, vol. 84, pp. 76–87, 2016.
- [24] T. Germa, F. Lerasle, N. Ouadah, and V. Cadenat, "Vision and RFID data fusion for tracking people in crowds by a mobile robot," *Comput. Vis. Image Und.*, vol. 114, no. 6, pp. 641–651, 2010.
- [25] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6517–6525.
- [26] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Trans. Multimedia Comput. Commun.*, vol. 14, no. 1, pp. 13:1–13:20, 2017.
- [27] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3908–3916.