Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

An Efficient Human-Following Method by Fusing Kernelized Correlation Filter and Depth Information for Mobile Robot

Shijian Su, Shuying Cheng, Houde Dai, Mingqiang Lin, Heng Yu, and Jianwei Zhang

***Abstract*—Human-following ability enables mobile robots to cooperate with human beings smartly. The kernelized correlation filter (KCF) was typically adopted as a human tracker during the human-following process due to its advantages of fast speed and high precision. However, it is easy to drift away owing to the changes of human pose. In this study, we proposed an efficient human-following method by fusion of KCF and depth information for the mobile robot. The target human was separated from the background regions based on the seeded region growing. KCF tracker outputs and segmentation results were fused by using the adaptive weighted fusion method. The weighting factors were adjusted based on the tracking quality indicators, which reflect the tracking reliability of KCF and the distribution of depth information. Experimental results verify the effectiveness of the proposed method in terms of tracking speed and stability. The center location error of all frames was less than 48 pixels and the bounding box overlap rate of all frames was larger than 0.62.**

***Index Terms*—Human-following, mobile robot, kernelized correlation filter, human segmentation, sensor fusion.**

I. Introduction

ROBOTS tend to be coexisting, cooperative, and cognitive with human beings in some application scenarios. For instance, the capability of human-following enables mobile robots to work together with humans efficiently [1-2]. Lee *et.al* [3] proposed a construction robot to assist humans in handling heavy construction materials. Similarly, logistics assistant robots in an airport [4] could help people to carry luggage.

This work was supported in part by National Natural Science Foundation of China under Grant 61973293, in part by the Chinese Academy of Sciences under Grant 121835KYSB20190069, in part by the Science and Technology Department of Fujian Province under Grants 2018H2001, 2018Y0036 and 2019T3010, and in part by Quanzhou Science and Technology Project under Grant 2019C012R. (*Corresponding author: Houde Dai*)

S. Su is with Quanzhou Institute of Equipment Manufacturing, Haixi Institutes, Chinese Academy of Sciences, Jinjiang 362216, China, and also with Institute of Micro/Nano Devices and Solar Cells, College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China (e-mail: shijiansu@fjirsm.ac.cn).

S. Cheng is with Institute of Micro/Nano Devices and Solar Cells, College of Physics and Information Engineering, Fuzhou University, Fuzhou 350116, China (e-mail: sycheng@fzu.edu.cn).

H. Dai, M. Lin, and H. Yu are with Quanzhou Institute of Equipment Manufacturing, Haixi Institutes, Chinese Academy of Sciences, Jinjiang 362216, China (e-mail: dhd@fjirsm.ac.cn).

J. Zhang is with Department Informatics, University of Hamburg, D-22527 Hamburg (e-mail: zhang@informatik.uni-hamburg.de).

Researchers have proposed a series of human-following solutions based on various sensing modules, such as light detection and ranging (LiDAR) [5-6], camera (monocular [7-8] or stereo [9-10]), ultra-wideband (UWB) [11], and radio-frequency identification (RFID) [12].

In general, the stereo vision approach has advantages of low hardware cost and abundant information acquisition ability, i.e., including both color images and depth information. Thus, a stereo vision module was adopted to perform objection detection and correlation filter in this study.

The human-following robot based on object detection generally detects the human in each frame, and then obtains the position of the target directly. The main work of this method is feature extraction and classification. Various solutions like the combination of speeded-up robust feature (SURF) and k-dimensional (K-D) tree [7], and the combination of the histogram of oriented gradients (HOG) with the support vector machine (SVM) [13] have been investigated for human detection. Especially, a hybrid gait feature was proposed by Chi *et al.* [8] to capture the static, dynamic, and trajectory features for each gait cycle. The advantage of these methods is that they could detect and track the target in all video sequences. However, it had high computational complexity and the detection result based on the artificial design features was disappointing for partially occluded humans [14]. In recent years, object detection using deep learning has made significant breakthroughs in detection speed and accuracy. The latest version YOLOv3 (you only look once) predicts human beings with a single network evaluation, which makes it 100 times faster than Fast R-CNN [15-17]. The primary drawback of YOLOv3 is its high computational complexity for the robot controller.

Correlation tracking method has attracted considerable attention due to its state-of-the-art performance in accuracy and time consumption. The correlation filter and minimum output sum of squared error (MOSSE) filter were first introduced by Bolme *et al*. [18] for object tracking. Henriques *et al*. [19] proposed a kernelized correlation filter (KCF) with high-speed tracking, which exploited Ridge Regression with cyclic shifts to reduce both storage and computation. KCF is based on the tracking-by-detection framework. To tackle the appearance changes of non-rigid targets such as the human body, the tracking model should be updated online, whereas the online approach results in the drifting problem. For a

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

human-following robot, the drifting problem is mainly caused by the variations of human pose in long-term tracking.

To overcome the drifting problem, Lin *et al*. [20] improved the KCF by adaptively changing the target response and integrating the color feature, and Pei *et al*. [21] automatically adjusted the learning factor to reconcile the stability-versus-plasticity dilemma. The drifting problem could be alleviated through these improvements, but it cannot be overcome entirely in long-term tracking. Some solutions utilized re-detection mechanisms to enhance the robustness, such as Zhang *et al*. [22] detected the occlusion based on the depth evaluation and the maximum response from KCF, and Qu *et al*. [23] introduced the confidence of candidate patches to measure the tracking reliability. However, it was only valid when the target was occluded or the tracking fails. Zhou *et al*. [24] initialized multiple KCF on a target in different time points and fused them via the maximum likelihood, while it required more memory and computing resources.

KCF should initiate a tracking window centered on the target in the start frame. In this study, YOLOv3 and KCF were adopted as the human detector and tracker respectively to function as a human-following robot. In addition, we proposed a tracking method based on adaptive weighted fusion of KCF and depth information to overcome the drifting problem caused by the changes of human pose. The weighting factors were adjusted based on the tracking quality indicators, which reflect the tracking reliability of KCF and the distribution of depth information.

In Section II, we present the overall framework of the proposed tracking method. Section III describes the implementation of all related modules. Experimental evaluations and results are provided in Section IV. Finally, we concluded this study in Section V.

II. System Overview

The main task for the human-following robot is to detect and track a certain human, and then generate necessary control commands to keep the human within the field of view (FOV) of the robot camera. The overall framework of the proposed method for the human-following robot is shown in Fig. 1.

The Kinect camera could provide color image, depth image, and three-dimension (3D) point clouds simultaneously. The depth image and 3D point clouds have been calibrated to match the correct pixel location in the color image. Depending on the camera model, the depth image and 3D point clouds can be converted to each other.

Firstly, the human detector based on YOLOv3 performs repeatedly until a human is detected. Then the detected human is selected by drawing a rectangular tracking window. In the subsequent frames, the KCF tracker is adopted to track the selected human. At the same time, the target human is separated from the background regions based on the seeded region growing algorithm. Segmentation results and KCF tracker outputs are fused to overcome the drifting problem in long-term tracking. All related modules are described in detail in the next section.

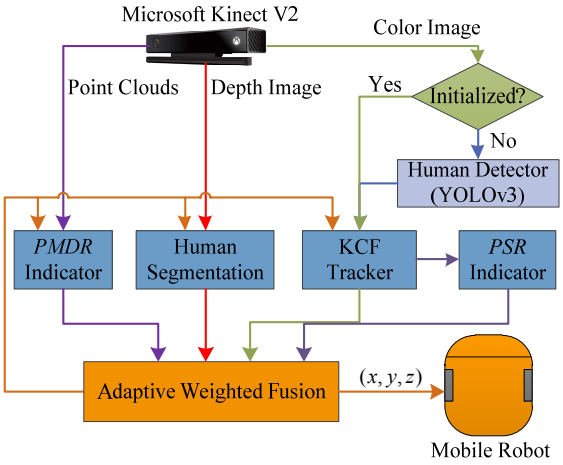


Fig. 1. Framework of the proposed tracking method. A Microsoft Kinect V2 was adopted for image acquisition. The mobile robot tracks and then follows a person with the proposed tracking method.

III. Methods

A. Human-Following Method based on YOLOv3 and KCF

YOLOv3 employs features from the entire image to predict what the objects are present and where they are, which has high detection speed and accuracy. YOLOv3 divides the image into regions and predicts bounding boxes and probabilities for each region. Each bounding box consists of 5 predictions: b_x, b_y, b_w, b_h, and b_c. (b_x, b_y) represents the top-left coordinates of the bounding box relative to the whole image. (b_w, b_h) indicates the width and height of the bounding box. $b_c \in [0,1]$ is a confidence score that reflects the relationship between the box and its estimation.

When more than one person is detected in the camera's FOV, the depth information provided by the Kinect camera and the above predictions is exploited to select the optimal tracking target. The evaluation function is defined as

$$G(b_c, x, y, z) = \alpha \cdot (1 - b_c) + \beta \cdot dist(x, y, z), \quad (1)$$

where α and β are the weighting parameters, (x, y, z) is the central position of the detected human in the Kinect coordinate, the function $dist(x, y, z)$ represents the distance to the expected position (x_e, y_e, z_e). The target with the minimum evaluation value is chosen as the final tracking target.

Depending on the YOLOv3 detector, the target human is initially selected by drawing a tracking window. Then the human is tracked by correlating the filter over the tracking window in the subsequent frames.

The correlation is computed when the shifting model is robust to the shift of the tracking target. The computational cost of this process is very low by using the circulant matrix. The image patch $\mathbf{X}$ from the tracking window is a 2D matrix with $M \times N$ pixels. They are transformed into a $1 \times n$ ($n = M \times N$) vector $\mathbf{x}$. All the circular shifts of x_i, $i \in (0,1,...,M \times N - 1)$, are generated as training samples with training target y_i. The correlation filter $\mathbf{w}$ could be trained by minimizing the squared error

$$\mathbf{w} = \arg\min_{\mathbf{w}} \sum_i (\varphi(x_i)\cdot\mathbf{w} - y_i)^2 + \lambda\|\mathbf{w}\|^2 \ , \quad (2)$$

where φ denotes the mapping to high-dimensional feature space and λ is a regularization parameter.

w could be expressed as a linear combination of the samples, *i.e.*, $\mathbf{w} = \sum_i \alpha_i \varphi(x_i)$, where α_i is the coefficient in the dual space, as opposed to the primal space **w**. The closed-form solution of (2) is

$$\boldsymbol{\alpha} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y} \ , \quad (3)$$

where **I** is an identity matrix, and **K** is the kernel matrix whose elements are $K_{ij} = k(x_i, x_j) = \varphi^T(x_i)\varphi(x_j)$.

Given a circulant data C(**x**), the corresponding kernel matrix **K** is a circulant matrix if the Gaussian kernel is adopted as the kernel function. Circulant data C(**x**) is obtained from the vector **x** by concatenating all possible cyclic shifts of **x**. Then (3) could be diagonalized, obtaining

$$\hat{\boldsymbol{\alpha}} = \hat{\mathbf{y}} / (\hat{\mathbf{k}}_{xx} + \lambda) \ , \quad (4)$$

where $\mathbf{k}_{xx}$ is the first row of the kernel matrix $\mathbf{K} = C(\mathbf{k}^{xx})$, and a hat ^ denotes the DFT of a vector. $\hat{\boldsymbol{\alpha}}$ is the vector of coefficients in the Fourier domain, which is obtained in the training progress.

To detect the target human in the tracking progress, we calculate the output response at several locations, *i.e.*, for candidate patches **z**. These patches can be modeled by cyclic shifts. The output response at all locations can be obtained

$$\mathbf{y} = F^{-1}(\hat{\boldsymbol{\alpha}} \odot \hat{\mathbf{k}}_{xz}) \ , \quad (5)$$

where $\mathbf{k}_{xz}$ is the kernel correlation of **x** and **z,** $\odot$ denotes element-wise product and F^{-1} is inverse DFT. The response reflects the detection score for each location, and the position of the maximum score is the new location of the target human.

To adapt the tracker to different target appearance, the tracking model needs to be updated online based on the new location at every frame.

$$\begin{cases} \hat{\boldsymbol{\alpha}}_{t+1} = (1-\rho)\hat{\boldsymbol{\alpha}}_t + \rho\hat{\boldsymbol{\alpha}} \\ \hat{\mathbf{x}}_{t+1} = (1-\rho)\hat{\mathbf{x}}_t + \rho\hat{\mathbf{x}} \end{cases} , \quad (6)$$

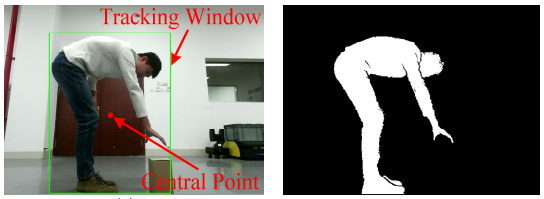
where ρ is the learning rate, $\hat{\boldsymbol{\alpha}}_{t+1}$ and $\hat{\mathbf{x}}_{t+1}$ are the vector coefficient and tracking model at frame t+1, respectively. Finally, the position of target human $\boldsymbol{c}_1 = (c_{1x}, c_{1y}, c_{1w}, c_{1h})$ can be obtained from the KCF tracker.

B. Human Segmentation based on Seeded Region Growing

Although the online model update makes the KCF tracker adapt to the variations of target features, it often leads to the drifting problem. Therefore, we adopted the depth information to restrain the drifting problem.

The intensity values in the depth image represent the distance between the object and the camera, which makes it suitable for segmentation. The target human is segmented based on depth information, which has the following advantages: 1) Solving the drifting problem by fusing the segmentation results with KCF tracker outputs; 2) Estimating the distance between the robot and the human more accurately. If we adopt the depth information in the entire tracking window to estimate the distance, the result is inaccurate because of the background involved.

The neighboring pixels within the human region are with similar values in the depth image. Hence the seeded region growing (SRG) [25] is applied in our study. SRG starts with assigned seeds and grows regions by adding a pixel to its nearest neighboring seed region depending on a region similarity criterion. How to generate the initial seed is the primary concern for the performing of the segmentation based on region growing. Huang [26] utilized a deep classification network to locate discriminative regions as seed cues. Similarly, we also have a tracking window containing the target human in the tracking process. The tracking window is determined by YOLOv3 detector and KCF tracker. However, there still are some background regions besides the target human in the tracking window, as shown in Fig. 2 (a).



(a) (b)

Fig. 2. Object segmentation based on SRG. (a) A tracking window containing the target human; (b) Removing background using seeded region growing method.

To select a suitable initial seed, a simple method is to randomly select several points near the center of the tracking window and then take the average value. This method assumes that the center of the tracking window belongs to the target. However, the center of the tracking window might be the background when the human is in the posture as shown in Fig. 2 (a). We propose a more robust method inspired by the random sample consensus (RANSAC) [27]. Our method is stated as follows, where that sample data points M are chosen evenly in the tracking window:

Step 1. Select the i-th point from M, where $i \in \{1, ..., M\}$.

Step 2. Calculate the difference between the value of the i-th point and the j-th point, where $j \in \{1, ..., M\}$ and $j \neq i$. Add the j-th point to subset S_i if the difference is less than the threshold R. Count the number of points in S_i.

Step 3. Repeat Step 1 and Step 2 until all sample points have been traversed. The point with the maximum points in S_i is regarded as the initial seed for SRG.

Compared to RANSAC, this method without instantiating a model is simpler and quicker. In the experiment, we took one sample data point every 5 pixels. The threshold R was set to 0.3 meters. The target human was segmented from the depth image as shown in Fig. 2 (b).

Therefore, the location of the target human $\boldsymbol{c}_2 = (c_{2x}, c_{2y}, c_{2w}, c_{2h})$ can be obtained from the human segmentation.

C. Adaptive Weighted Fusion of KCF and Depth Information

The results obtained from KCF and human segmentation are of the same type. Hence the adaptive weighted fusion algorithm is employed to fuse the KCF tracker outputs and human segmentation results to obtain the optimal position of the target

person. The fused position is

$$\begin{cases} \boldsymbol{c}_f = w_1 \cdot \boldsymbol{c}_1 + w_2 \cdot \boldsymbol{c}_2 \\ s.t.\ \ w_1 + w_2 = 1 \end{cases}, \tag{7}$$

in which w_1 and w_2 are the weighting factors. How to adjust the w_1 and w_2 adaptively is the key to estimating the $\mathbf{c}_f$.

The peak to side-lobe ratio (*PSR*) is a tracking quality indicator, which measures the strength of the correlation peak. It reflects the correlation between the tracking target and the tracking model.

$$PSR = (g_{\max} - \mu_s) / \sigma_s \ , \tag{8}$$

in which $g_{\max}$ is peak value, μ_s and σ_s are the mean and standard deviation of the side-lobe respectively. The *PSR* would decrease in two cases: 1) The appearance of the target is changed significantly due to pose variations, such as out-of-plane rotation; 2) The target is partially or fully occluded by other objects in the environment.

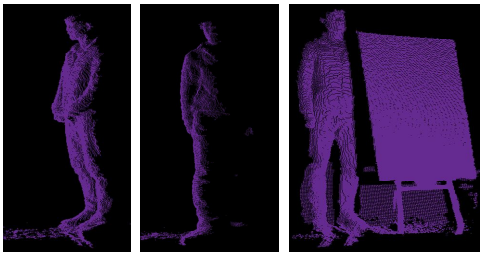


Fig. 3. Distribution of point clouds under different conditions. (a) Normal pose; (b) Pose change; (c) Occluded by other objects.

Unlike the KCF tracker, human segmentation based on SRG will work well if the pose of target human changes significantly. Thus the point clouds provided by the stereo vision system, as shown in Fig. 3, are adopted to evaluate the depth distribution. Since the quality of human segmentation needs to be calculated on each frame, the mean and standard deviation of point clouds is adopted, which have low computational complexity. The quality indicator *PMDR* is defined as

$$PMDR = \mu_p / \sigma_p \ , \tag{9}$$

where μ_p and σ_p are the mean and standard deviation of the point clouds. *PMDR* will decrease obviously under the condition that the target human is partially occluded.

PSR and *PMDR* are normalized, and then we can obtain δ_1 and δ_2 which indicate the tracking quality of the KCF tracker and the quality of human segmentation, respectively. The weighting factors w_1 and w_2 can be adjusted adaptively depending on δ_1 and δ_2 as follows:

$$\begin{cases} w_1 = \delta_1 / (\delta_1 + \delta_2) \\ w_2 = \delta_2 / (\delta_1 + \delta_2) \end{cases}. \tag{10}$$

If δ_2 is larger than δ_1 during the fusion process, the fusion result tends to be the result of human segmentation. When both δ_1 and δ_2 are lower than a certain threshold, it can be determined that the target human is occluded, which is very likely to result in tracking failure. In this condition, approaches such as restarting the human detector can be adopted.

IV. Experimental Evaluation

The proposed tracking method employs color images, depth images, and point clouds. It is difficult to find an available benchmark dataset to test our proposed method. Therefore, experiments were carried out on an actual mobile robot.

A. Experimental Platform and Real-Time Evaluation

Fig. 4 shows the experimental platform based on our custom-developed self-designed mobile robot. Robot operating system (ROS) and Ubuntu 16.04 are installed on the Jetson TX2 module (NVIDIA Corp., U.S.), which is a power-efficient (7.5 W) embedded artificial intelligence computing device. The color images with a resolution of 480×270, which were obtained from the Kinect v2 (Microsoft Corp., U.S.), were adopted in the experiment.

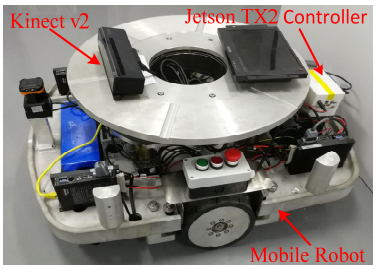


Fig. 4. Self-developed experimental platform for human following. The Kinect v2 was mounted on the top of the mobile robot to capture the target human. The Jetson TX2 was used to run the detecting and tracking algorithm.

YOLOv3-based human detector method and the proposed tracking method based on the fusion of KCF tracker and human segmentation were tested on the Jetson TX2. YOLOv3 could run at 0.3 frames per second (FPS) with only CPU, and 24 FPS after GPU acceleration. The proposed tracking method could run up to 17 FPS when only CPU was used. Thus the proposed tracking method is efficient, which can run on the embedded controller of the mobile robot in real-time.

B. Test of Quality Indicators

The human-following robot will encounter various challenges in the actual application scenario. Some complex scenes are represented in Fig. 5 (c). The human poses changed in situations B, D, E, and F, and the target human got part occlusion in situations C and G. In situation H, there were no effective point clouds in the tracking window when the human was out of view. Meanwhile, Fig. 5 (a) and (b) show the tracking quality of KCF tracker and the quality of human segmentation under the corresponding situations respectively.

According to Fig. 5 (a) and (c), the value of *PSR* will decrease in both cases where the target human has occluded and his pose changes. Notably, the *PSR* value in situation G is less than that in situation D, *i. e.*, we cannot make a distinction between pose change and occlusion. The main difference against *PSR* is that *PMDR* value declines rapidly only when the human is occluded, as shown in Fig. 5 (b). Changes in *PMDR* value correspond to the fact that the effect of human segmentation is mainly affected by occlusion.

As mentioned above, the *PSR* and *PMDR* could reflect the tracking quality of the KCF tracker and the quality of human

segmentation. The target human tends to be lost when both PSR and PMDR are decreased significantly.

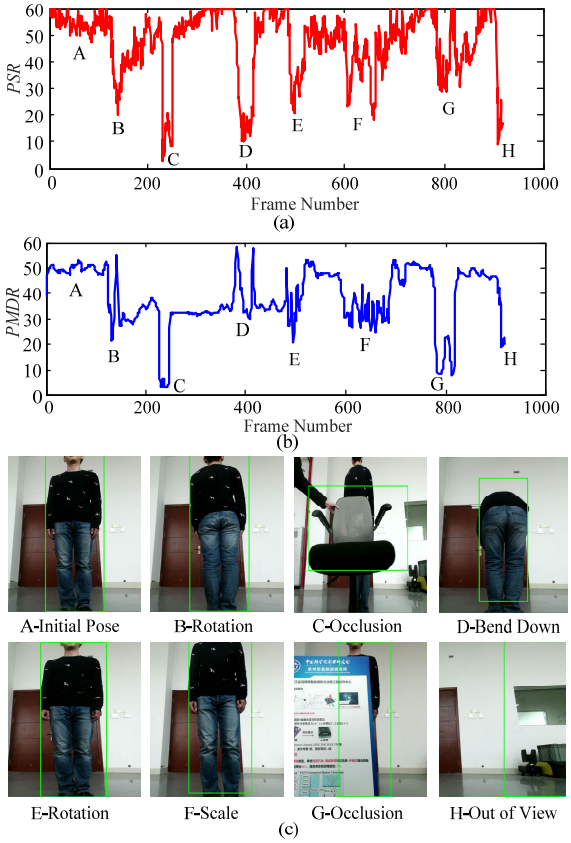


Fig. 5. Tracking quality in different situations. (a) Curve of *PSR*. The two minimum values appear at points C and D, which are caused by pose change and occlusions, respectively. (b) Curve of *PMDR*. The two minimum values appear at point C and G, both of which are caused by occlusion. (c) Snapshots of some challenging situations.

C. *Performance of the Proposed Tracking Method*

The *rosbag* is a command-line tool provided by ROS to record and playback a bag file with the contents of specified topics. We used the tool to record color and depth image sequences, which depicted the situations about pose changes. The image sequences were adopted to test the performance of the original KCF and the proposed tracking method, respectively.

Fig. 6. Comparison between the original KCF and the proposed tracking method. The red window comes from the original KCF and the green window comes from the proposed tracking method.

Fig. 6 indicates that the red tracking window becomes larger after pose changes between frames 520 and 1400. The background would be updated into the tracking model owing to the online model update mechanism. According to frames 1722 and 2123, we can observe that the red tracking window shifts from the human body when the human underwent pose changes, and then it locks to a new center point. With the accumulation of drifting, it tends to fail in long-term tracking. As the green tracking window shows, the improved KCF tracker could track the human target well without the drifting problem.

Then we use the precision and success rate to evaluate the performance of the tracker quantitatively. Center location error (CLE) is defined as the average pixel distance between the center location of the tracked human and the ground truth. The precision plot Fig. 7 (a) shows the percentage of frames whose CLE is within the given threshold t_p. Another evaluation metric is the bounding box overlap. The overlap score is defined as

$$S = \left|R_t \cap R_g\right| \Big/ \left|R_t \cup R_g\right| , \qquad (11)$$

where R_t and R_g denote the tracked and ground truth bounding box, respectively; $\cap$ and $\cup$ represent the intersection and union of the two bounding boxes, respectively. The success plot Fig. 7 (b) shows the ratio of frames whose S is larger than the given threshold t_s.

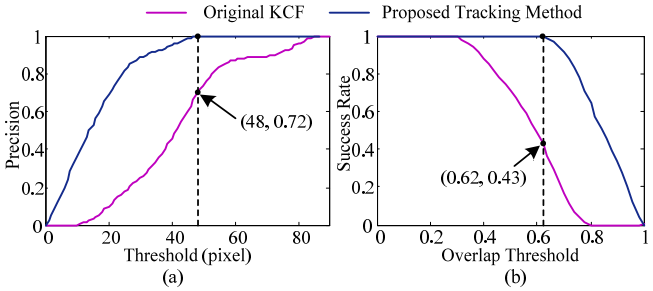


Fig. 7. Quantitative comparison between the original KCF and the proposed tracking method. (a) The precision of original KCF and the proposed tracking method are 0.72 and 1 respectively at the threshold of 30 pixels; (b) The success rates of original KCF and the proposed tracking method are 0.43 and 1 respectively at the overlap threshold of 0.62.

For the proposed tracking method, the center location error of all frames is less than 48 pixels, and the overlap rate of all frames is larger than 0.62. Results indicate that the combination of KCF tracker outputs and human segmentation results can significantly alleviate the drifting problem caused by the variations of human pose.

D. *Human-Following Robot Evaluation*

To further test the performance of the human-following method, we deploy the tracking method on the mobile robot. The Vicon motion system (Oxford Metrics Ltd., UK) was used to record the trajectories of both the mobile robot and human. One optical maker was mounted on top of the mobile robot, and another optical maker was held by the human. The expected distance z_e between the mobile robot and the human target was set to 1.1 meters.

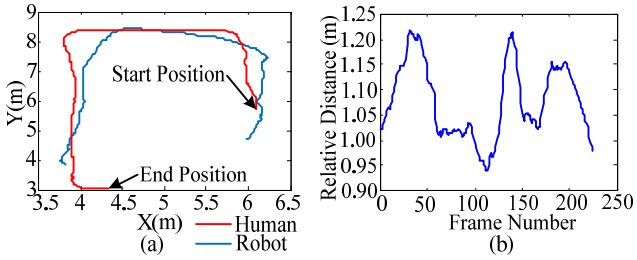


Fig. 8. Experimental results of a human-following robot. (a) Trajectories of the robot and human; (b) Relative distance between the robot and the human.

From the start position to the end position, the total displacement of human walking was about 11 meters, which took about 56 seconds. According to Fig. 8 (a), the trajectories of the robot, and the human are similar. The relative distance between the robot and the human could be maintained between 0.9m and 1.25m, which is closed to the expected distance, as shown in Fig. 8 (b). The results demonstrate that the robot could follow the target human effectively.

V. CONCLUSION

For a human-following robot, the drifting problem of the KCF tracker is mainly caused by the variations of human pose in long-term tracking. In this study, we take full advantage of depth information provided by Kinect v2 to overcome this limitation. The depth information was introduced to separate the target human from background regions and to determine whether the target human is occluded or not. Segmentation results and KCF tracker outputs were fused by using the adaptive weighted fusion method. Experiments demonstrate the effectiveness of the proposed tracking method, which can be conducted on an embedded controller in real-time and overcome the drifting problem. In addition, only a single person was tracking in the experiment, although the tracking of a person in a complex environment is feasible. In the next study, this method will be verified in some realistic scenarios, such as the human robot interaction task in an indoor cargo storage warehouse. However, one limitation of this method is that its performance can be influenced by the external ambient light. Hence, the algorithm need improved for the outdoor applications or the environment with high luminosity.

In the future, we will further improve the accuracy of the tracking quality indicator. For instance, the feature model obtained from the most reliable tracked targets could be employed to measure the confidence of the tracking results. In addition, other performances such obstacle avoidance ability and motion control strategy should be further evaluated and optimized.

REFERENCES

[1] K. Morioka, J. H. Lee, and H. Hashimoto, "Human-following mobile robot in a distributed intelligent sensor network," *IEEE Transactions on Industrial Electronics*, vol. 51, no. 1, pp. 229-237, 2004.

[2] M. Chueh, Y. L. W. Au Yeung, K. P. C. Lei, and S. S. Joshi, "Following controller for autonomous mobile robots using behavioral cues," *IEEE Transactions on Industrial Electronics*, vol. 55, no. 8, pp. 3124-3132, 2008.

[3] S. Y. Lee, K. Y. Lee, S. H. Lee, J. W. Kim, and C. S. Han, "Human-robot cooperation control for installing heavy construction materials," *Autonomous Robots*, vol. 22, no.3, pp. 305-319, 2007.

[4] M. Gupta, S. Kumar, L. Behera, and V. K. Subramanian, "A novel vision-based tracking algorithm for a human-following mobile robot," *IEEE Transactions on Systems Man & Cybernetics Systems*, vol. 47, no. 7, pp. 1415-1427, 2017.

[5] W. Chung, H. Kim, Y. Yoo, C. B. Moon, and J. Park, "The detection and following of human legs through inductive approaches for a mobile robot with a single laser range finder," *IEEE Transactions on Industrial Electronics*, vol. 59, no. 8, pp. 3156-3166, 2012.

[6] 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*, pp. 1-16, 2018.

[7] M. Gupta, S. Kumar, L. Behera, and V. K. Subramanian, "A novel vision-based tracking algorithm for a human-following mobile robot," *IEEE Transactions on Systems Man & Cybernetics Systems*, vol. 47, no. 7, pp. 1415-1427, 2017.

[8] W. Chi, J. Wang, and Q. H. Meng, "A gait recognition method for human following in service robots," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*," vol. 48, no. 9, pp. 1429-1440, 2017.

[9] J. S. Hu, J. J. Wang, and D. M. Ho, "Design of sensing system and anticipative behavior for human following of mobile robots," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 4, pp. 1916-1927, 2014.

[10] A. Jevtić, G. Doisy, Y. Parmet, and Y. Edan, "Comparison of Interaction Modalities for Mobile Indoor Robot Guidance: Direct Physical Interaction, Person Following, and Pointing Control," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 6, pp. 653-663, 2015.

[11] T. Feng, Y. Yu, L. Wu, Y. Bai, Z. Xiao, and Z. Liu, "A human-tracking robot using ultra wideband technology," *IEEE Access*, vol. 6, pp. 42541-42550, 2018.

[12] M. Kim, N. Y. Chong, H.-S. Ahn, and W. Yu, "RFID-enabled target tracking and following with a mobile robot using direction finding antennas," *in* IEEE Conference Automation Science and Engineering, pp. 1014-1019, Sep. 22-25, 2007.

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

[14] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743-761, 2012.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *in* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[16] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *in* IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017.

[17] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv, 2018.

[18] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Liu, "Visual Object Tracking Using Adaptive Correlation Filters," *in* IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

[19] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," I*EEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583-596, 2015.

[20] Z. Liu, Z. Lian, and Y. Li, "A novel adaptive kernel correlation filter tracker with multiple feature integration," *in* IEEE International Conference on Image Processing (ICIP), pp. 2572-2576, 2017.

[21] M. Pei, W. Li, Z. Ke, and Q. Gao, "Improved kernelized correlation filters tracking algorithm with adaptive learning factor," *in* IEEE Control Conference, 2016.

[22] L. Zhang, Z. Cao, X. Meng, C. Zhou, and S. Wang, "Real-time depth-based tracking using a binocular camera," *in* 12th World Congress on Intelligent Control and Automation (WCICA), pp. 2041-2046, 2016.

[23] Z. Qu, X. Lv, J. Liu, L. Jing, *et al*., "Long-Term Reliable Visual Tracking with UAVs," *in* IEEE International Conference on Systems, and Cybernetics (SMC), pp. 2000-2005, 2017.

[24] Y. Zhou, T. Wang, R. Hu, H. Su, *et al.*, "Multiple Kernelized Correlation Filters (MKCF) for Extended Object Tracking Using X-band Marine Radar Data," IEEE Transactions on Signal Processing, pp. 3676-3688, 2019.

[25] R. Adams and L. Bischof, "Seeded region growing," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 16, no. 6, pp. 641–647, 1994.

[26] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang, "Weakly-supervised semantic segmentation network with deep seeded region growing," *in* IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2018.

[27] M.A. Fischler and R.C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Readings in Computer Vision*, vol. 24, no. 6, pp. 381- 395, 1981.