2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Macau, China, November 4-8, 2019

Person-following for Telepresence Robots Using Web Cameras

Xianda Cheng, Yunde Jia, *Member, IEEE,* Jingyu Su, and Yuwei Wu

***Abstract*— Many existing mobile robotic telepresence systems have equipped with two web cameras, one is a forward-facing camera (FF camera) for video communication, and the other is a downward-facing camera (DF camera) for robot navigation. In this paper, we present a new framework of autonomous person-following for telepresence robots using the two web cameras. Based on correlation filters tracking methods, we use the FF camera to track the upper body of a person and the DF camera to localize and track the person's feet. We improve the robustness of feet trackers, consisting of a left foot tracker and a right foot tracker, by making full use of the spatial constraints of the human body parts. We conducted experiments on tracking in different environmental situations and real person-following scenario to evaluate the effectiveness of our method.**

I. INTRODUCTION

A telepresence robot is a mobile videoconferencing system which can be teleoperated by a remote operator [1], and has many applications in terms of telepresence in local environments such as home, office, school, exhibitions, etc. A remote operator can teleoperate a telepresence robot, as his/her embodiment, to interact with local persons. In many situations, a telepresence robot needs to follow a local person to a new destination or to walk together with a local person whom a remote operator is interacting with. Therefore, a telepresence robot should have the ability of following a local person to reduce the burden on a remote operator to drive the robot, which makes you have more energy to communicate and interact with the local person.

The most popular approaches of person-following for social robots are to use laser range finders (LRFs) [2], [3], [4], [5], [6] or stereo vision [7], [8] to detect, localize, and track a person. Telepresence robots are a kind of social robots, which have commonly equipped with web cameras for video communication and navigation. In order to make a robot have the ability of following a person, researchers usually add distance imaging sensors, such as LRF or Kinect, to the robot [9]. In this paper, we present a person-following framework only using the web cameras of a telepresence robot without adding distance image sensors. Specifically, since many mobile robotic telepresence systems have a FF camera for video communication and a DF camera for robot navigation, we use the FF camera to track the person's upper body and the DF camera to localize and track the person's feet, as shown in Fig. 1.

For upper body tracking in our system, scale change is a major issue that often affects tracking results. We use the fast discriminative scale space tracking (FDSST) algorithm [10], which is designed to estimate the scale by additionally training a scale filter, to cope with scale changes. For feet tracking, factors affecting the results are illumination changes, fast motion, other people wearing similar shoes or pants, etc. One foot tracker might shift to the other foot because they are very similar in appearance [11]. Another problem we have to face is that self-occlusion during turning around will cause tracking failure. In order to alleviate the above issues, we use the spatial constraints of the body parts to improve the robustness of the feet tracker.

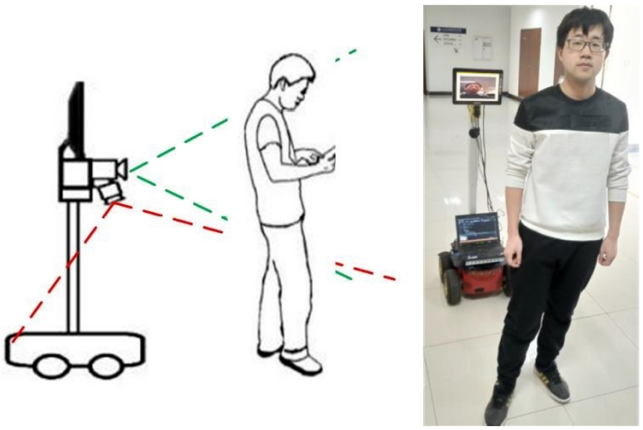

Fig. 1. System overview. The telepresence robot automatically follows a person using two web cameras, which can reduce the workload of controlling the robot.

II. RELATED WORK

One of the most common frameworks of person-following is to use laser range finders (LRFs) [9], [12]. For example Leigh *et al.* [12] proposed a tracking method with good portability to track both legs by using an LRF. Local occupancy grid maps is also integrated to improve data association. Utilizing stereo vision is another popular approach to achieve person-following through depth maps [7], [15]. For example, Chen *et al.* [7] proposed an online Ada-Boosting tracking algorithm with integrated depth information from a stereo camera. Pang *et al.* [15] used a binocular camera to obtain depth maps and integrated local depth pattern features into the kernelized correlation filter (KCF).

There has been few work on person-following for robotic systems only using web cameras. Anezaki *et al.* [16] proposed a human-tracking robot using QR code recognition and shape-based pattern matching. Shukor *et al.* [17] used a color-based detection methond for object tracking and following robot in library environment. Hu *et al.* [18] proposed

X. Cheng, Y. Jia, J. Su, and Y. Wu are with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 10081, P.R. China, e-mails: {chengxianda, jiayunde, sujingyu, wuyuwei}@bit.edu.cn.

a person tracking and following method combining clothes color features of the person's upper body and the contour of the head-shoulder. All the above work employs the change of person's shape scale (pixel number) to control the movement of the robot. This control strategy isn't robust, especially when there is an out-of-plane rotation of the human body. Different from the work mentioned above, we propose to calculate the distance directly in the two-dimensional space of two-dimensional ground plane image from the DF camera.

In our system, we use the FF camera to track the upper body of a person and the DF camera to localize and track the persons feet. A similar work was done by Koide and Miura [20]. They proposed an online boosting based person-following approach using the combination of color, height and gait features to identify the followed person. They used a face-forward camera to capture the followed person's image, and then identify to identify the color feature of clothes and estimate the height based on imaging geometry. However, without the help of LRFs, it is difficult to estimate the person's height to meet the needs of applications. Besides, gait features were calculated by the accumulated range data of the legs from the LRF.

III. PLATFORM

We use a telepresence robot developed in our lab as a testing platform, called Mcisbot [21], as shown in Fig. 1. The Mcisbot contains the Pioneer 3-AT as mobile robot base and a telepresence head. The robot head consists of a light LCD screen, a FF camera, a DF camera, and a speaker & microphone, and all together are mounted on a pan-tilt platform hold up by a vertical post. The post can be moved vertically to adjust the robot height ranging from $1200mm$-$1750mm$, covering school child and adult body heights. The robot has equipped with a FF camera primarily for video communication and a DF camera for navigation.

Different from other telepresence robots that are piloted through a traditional GUI with a key board and mouse or touchscreen GUI with a graphical buttons, the Mcisbot can be piloted through a touchable live video image based user interface (TIUI), as shown in Fig. 2. The TIUI only contains the live video(s) from the Mcisbot robot without or almost without graphical buttons, keys, and menus. A remote operator can use the TIUI to not only drive a telepresence robot but also operate local objects just by touching their live video images. In our scenario, a remote operator can directly choose the person to be followed simply by touching the live video image of the person with touchscreen gestures.

IV. PERSON-FOLLOWING FRAMEWORK

The framework mainly consists of the upper body tracking and feet tracking, based on correlation filters that have shown accurate and high speed performance on tracking benchmarks [22]. The fast discriminative scale space tracking (FDSST) algorithm [10] is used to track the upper body, and the general Kernelized correlation filter (KCF) [23] is used as the baseline tracker to track feet. We integrate color naming (CN) features into the histogram of oriented gradient (HOG) features to enhance the discriminative features of feet appearance. In particular, we also utilize spatial constraints between human body parts to make sure that feet trackers can track the correct target. Thus, we refer to our feet tracking method as the KCF-SC, as spatial constraints of human body parts are utilized.

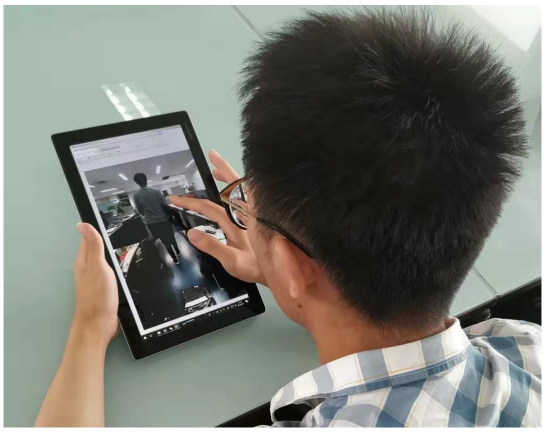

Fig. 2. Touching live video image user interface (TIUI).

A. Upper body tracking

The FDSST [10] is a high-performance tracker dealing with the scale change of the tracking target. A translation filter is first used to estimate the new target location and a scale filter is then applied to estimate the scale of the target.

1) Translation filter: We use the kernel correlation filter (KCF) as the translation filter to model the appearance of a target in the kernel space. The filter is trained on a $M \times N$ image patch p, which consists of each shifted sample $p_{m,n}$, $(m,n) \in \{0,1,...,M-1\} \times \{0,1,...,N-1\}$. The convolution response f is calculated by $f(p_{m,n}) = w \cdot \phi(p_{m,n})$, where ϕ denotes the mapping to the kernel space. Then the filer w can be trained in

$$w^* = \arg\min_w \sum_{m,n} (f(p_{m,n}) - g_{m,n})^2 + \lambda \|w\|^2, \quad (1)$$

where g is the desired correlation output that is usually a Gaussian function and λ denotes the regularization parameter. To optimizing the training process, a kernel $k(p,p') = \phi(p) \cdot \phi(p')$ is used to calculate the filter $w = \sum_{m,n} \alpha_{m,n} \phi(p_{m,n})$, where α is the dual variable of w. α is calculated in the Fourier domain by

$$\widehat{\alpha} = \frac{\widehat{g}}{\widehat{k^{pp}} + \lambda}, \quad (2)$$

where k^{pp} is a matrix whose element (m,n) is $k(p_{m,n},p)$ and $\widehat{g}$ denotes the DFT operation on g. When there comes a new frame, the patch z is extracted at the last target location and the response map is given by

$$f(z) = \mathcal{F}^{-1}(\widehat{k^{pz}} \odot \widehat{\alpha}), \quad (3)$$

where $\odot$ denotes the element-wise multiplication. The location of the target in the new frame can be estimated by searching the location of the max value of $f(z)$.

To further reduce the computations, two approuaches are used. One is the sub-grid interpolation of correlation scores with trigonometric polynomials and the other one is reducing the feature dimensionality using the Principal Component Analysis (PCA) [10].

2) Scale filter: The body scale will change while the person moves. The scale filter is applied to estimate the scale by computing correlation scores at the location obtained by the translation filter in the scale dimension [10]. The desired output correlation output g here is a one-dimensional Gaussian function as the lable of scales is one-dimensional. The learning and detection steps are similar to those of the translation filter, aiming to estimate the scale of the target based on the correlation results. The feature dimension of the scale filter can also be reduced by using the PCA.

B. Feet tracking

In our practice, unlike the upper body, the scale of the each foot changes little. Thus, we only use the KCF to perform feet tracking without scale estimation for decreasing the computational cost.

1) Incorporating CN features into HOG: The KCF tracker only takes HOG features to track targets. It has been confirmed that combining complex color features with luminance could show outstanding performance in object detection and tracking [24]. Eleven basic colors, black, blue, brown, gray, green, orange, ping, purple, red, white, and yellow, are used to represent pixels. We generate the CN feature map from the RGB color space and concatenate it with the HOG feature to obtain a new feature map of the tracking target.

2) Spatial constraint between upper body and feet: When a person is walking, his/her upper body is alway above the feet. Therefore the detection area of feet can be guided by the tracking result of the upper body, which improves the detection speed and reduce background interference. Fig. 3 shows the located area of feet. In our practice, the range of feet motion is approximate to a sector area of about $\theta = 80$ degrees. According to the location of the upper body, we estimate the deflection angle of the feet relative to the orientation of the robot. The horizontal coordinate can then be calculated according to the deflection angle.

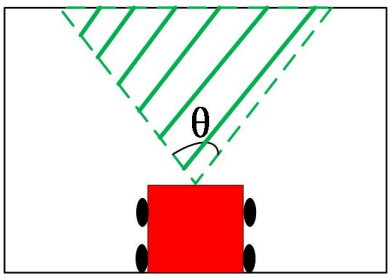

Fig. 3. The range of feet motion is a sector area of about 80 degrees based on the location of the upper body.

3) Spatial constraint between two feet: We assume that the displacement of the foot is not very large in successive frames. We multiply a binary mask with each value of the response map $f(z)$. The modified convolution conresponse value of each location (m, n) of the image patch z is $f^{'}(z_{m,n}) = \gamma_{m,n} f(z_{m,n})$. The distance between the location (m, n) and the center of the response map is defined as $d_{m,n}$. The value of each binary mask is defined as

$$\gamma_{m,n} = \begin{cases} c, & if\ d_{m,n} > T \\ 1 - c, & if\ d_{m,n} <= T. \end{cases} \tag{4}$$

Note that c is a constant close to zero. We set c to 0.1 and T to 2. Via the modified response map, it is effective to prevent any tracker from shifting to the other foot when the person walks straight. However, when the person turns, the foot of swing leg is occluded by the supporting leg. In this case the binary mask is invalid. To cope with this situation, the pixel distance d between the bounding boxes from the two feet trackers is used to control the value of a switch s, which equals to 0 when $d > 30$, and equals to 1 when $d <= 30$. When s equals 1, it means that one foot tracker has shifted to the other foot. In this case, if the location of the bounding box from the left/right foot transits, the corresponding foot tracker would detects an area at distance ℓ to the left/right side of the right/left foot.

C. Tracker recovery using Kalman filtering

A two-dimensional Kalman filtering method is used to estimate the location of a target to deal with the situation where the tracker fails. The state variable x_t at time t is a R^4 vector $(p_t^x, p_t^y, v_t^x, v_t^y)$. For each consecutive frame, the elapsed time Δt is used in state transition matrix F in

$$F = \begin{pmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{5}$$

Here Δt is fixed to $0.05s$. The initial value for the state is a zero vector since the coincides with the origin and it is stationary. When the tracking target is lost, the tracker will search around the location predicted by the Kalman filter until the max value of the response map is greater than a presetting threshold.

V. ROBOT CONTROL

A pre-specified distance D needs to be maintained between the robot and the followed person. The real distance between the robot and a person can be calculated directly based on the pixel distance between the robot and the center of two bounding boxes of the feet, as shown in Fig. 4 (a).

We define x^m as the horizontal coodinate of the image center from of the FF camera. x is the horizontal coordinate shown of the upper body, as shown in Fig. 4 (b). The distance between them is the parameter for controlling angular velocity of the robot. We use Proportional and

Fig. 4. (a) The distance d^p between the robot and the followed person can be calculated based on the feet tracking result. (b) The orientation information can be obtained based on the upper body tracking result.

Differential components of the PID controller to control the linear velocity ν and angular velocity ω at time t:

$$\begin{aligned} \nu &= K_p^l(d_t^r - D) + K_d^l(d_t^r - d_{t-\Delta t}^r), \\ \omega &= K_p^a(x_t - x^m) + K_d^a(x_t - x_{t-\Delta t}). \end{aligned} \tag{6}$$

K_p^l, K_d^l, K_p^a, K_d^a are PD constants, $(d_t^r - D)$, $(x_t - x^m)$ are error terms for linear velocity and angular velocity respectively, and dt is the time difference.

VI. EXPERIMENTS

A. Experiment on video sequences

We got four video sequences by controlling the telepresence robot manually to follow a person, among which two were captured in our laboratory and the other two were captured in ourdoor environment. All videos contain different feet tracking challenges. The duration of each video is $40s$, $25s$, $50s$, and $45s$, containing 993, 630, 1237 and 1144 frames, respectively. We evaluate our tracking methods based on tracking success rate. When the person turns around, the tracker will would to the other foot in some frames because of the self-occlusion. We still consider that the tracking is successful in this case as it is inevitable.

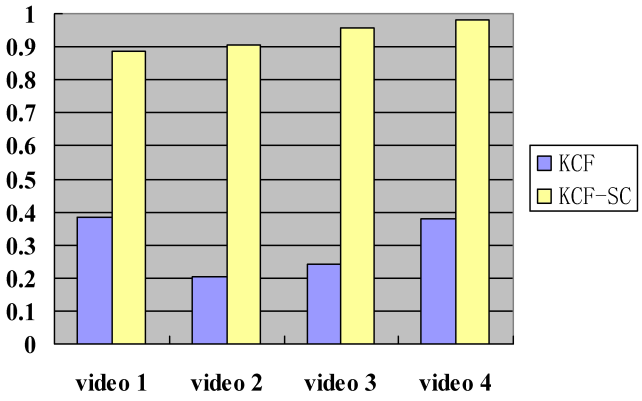

Fig. 5. Comparisons between the KCF-SC and the KCF.

The tracking results in the four videos are shown in Figures 6-9. The number under each col indicates the frame numbe. The first row is the results of upper body tracking, the second row and the third row show the cropped tracking results of the KCF trackers and the KCF-SC trackers, respectively.

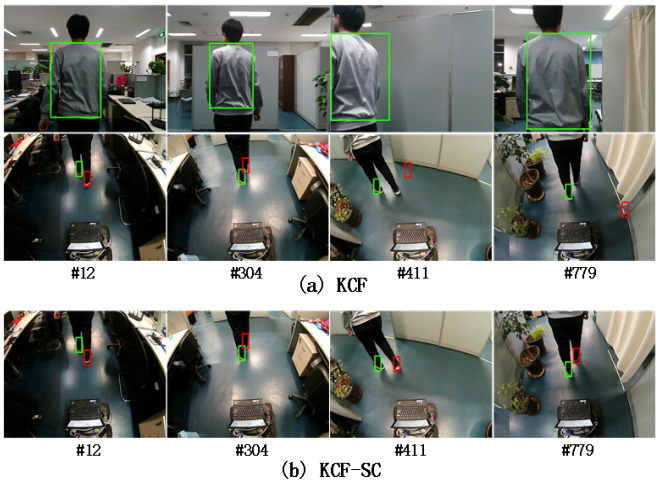

Fig. 6. Tracking results in the video 1. (a) The KCF trackers maintain successful in tracking in the first 304 frames, but after that, the right foot tracker fails due to the self-occlusion. (b) The KCF-SC trackers keep working well, even when the person turns around.

Fig. 6 shows the tracking results in the video 1. The video 1 was captured in the lab with background interference like chairs, flowerpots, etc. In the frame #304, the person turns left at a cornor. When the turning is over, the right foot tracker of the KCF-SC tracks the right foot again. However, the right foot tracker of the KCF shifts to the ground in the frame #411.

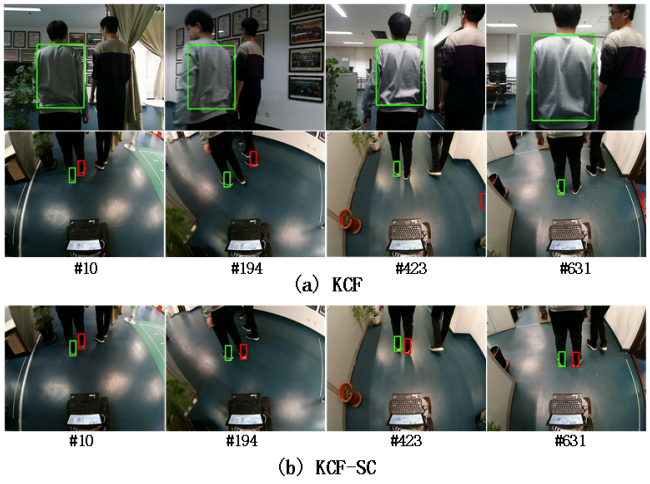

Fig. 7. Tracking results in the video 2. There are two persons walking side-by-side, and they wear the same color of pants and shoes. (a) The right foot tracker of the KCF shifts to the other person's left foot in the frame #194. In the frame #423, the tracker fails. (b) The KCF-SC trackers keep stable throughout the process.

Fig. 7 shows the tracking results in the video 2. The video 2 was also captured in the lab, and there are two persons wearing the same black pants and black shes, and walking

side by side. In the frame #194, they are all turning around. The KCF-SC trackers still work well after turning. However, the right foot tracker of the KCF shifts to the other person's left foot in the frame #194 and shifts to the ground later.

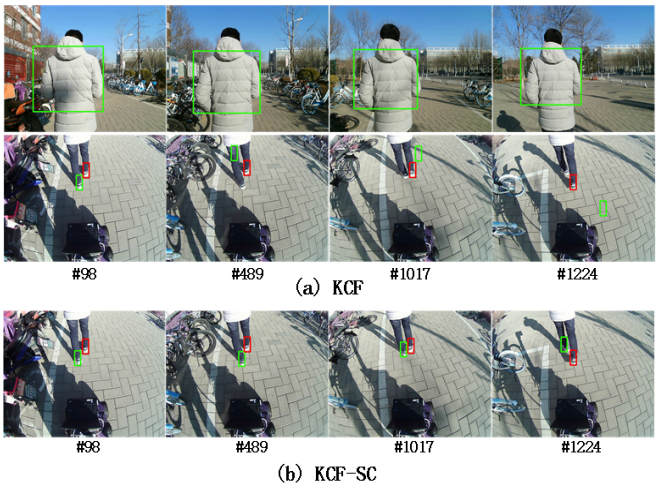

Fig. 8. Tracking results in video 3. The video was captured in a bike parking area where bikes might have an influence on the tracking. (a) The left foot tracker of the KCF shifts to the left leg in the frame #489 and then it shifts to the ground after a while. (b) Both feet trackers of the KCF-SC can track their targets accurately although tracking falilures occur in some frames.

Fig. 8 shows the tracking results in the video 3. The video 3 was captured in an outdoor bicycle parking lot. In this video, the KCF-SC trackers keep an accurate tracking of the targets thoughout the whole process. The left foot tracker fails and shifts to the left leg in the frame #489.

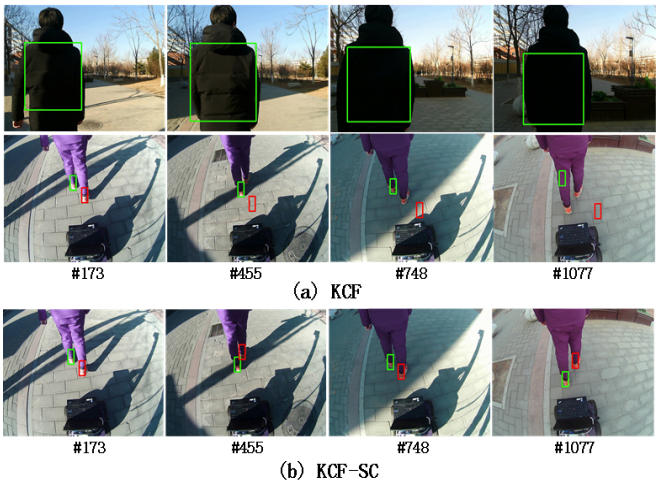

Fig. 9. Tracking results in video 4. The illumination change is the major challenge in this video. Due to the interference of infrared, the image taken by the DF camera has a color difference. (a) The right and left foot trackers of the KCF shift away in the frames #455 and #1077 respectively. (b) Both trackers of the KCF-SC maintain accurate tracking.

Fig. 9 shows the tracking results in the video 4. The video 4 was captured on outdoor pavement with illumination changes. In the frames #455 and #748 of the video, the brightness of the area where the feet are located suddenly dims. The KCF-SC trackers work well throughout the tracking process. The right foot tracker and left foot tracker of the KCF fail in the frames #455 and #1077 frames, respectively.

From the experiment results described above, we can see that, in contrast to the baseline KCF, the KCF-SC method has much more satisfactory tracking accuracy.

B. Experiment on person-following

We conducted an experiment on person following to verify the usage of our method in robotic telepresence systems.An remote operator first teleoperates the telepresnece robot to move to the front of the local person being followed, and activates the autonomous person-following function when the person turns around. After that, the robot can autonomously follow the local person walking around. In our experiment, the local person walks at a speed of about 0.5m/s. We recorded the distance between the robot and the followed person, and the horizontal difference between the center of the upper body and the the image center of the FF camera at 0.05s intervals. The testing data are shown in Figures 10 and 11. The duration of the following experiment is about 50s.

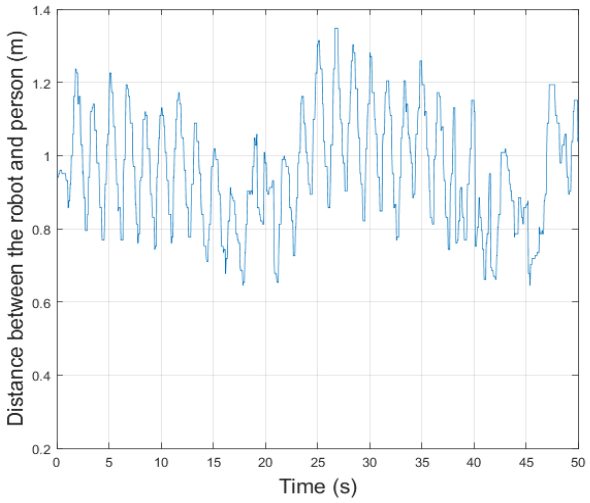

Fig. 10. The distance between the robot and the followed person and the robot for a run. It maintains between $0.6m$ and $1.4m$.

From the two figures, we can see that the distance between the robot and a person is ranging from $0.6m$ and $1.4m$, and the horizontal difference between upper body and FF camera's image center changes drastically when the turning happens at about $17s$ and $42s$. During the whole experiment, the telepresence robot could always follow the local person fluently. Although the trackers might fail in several frames, the tracking would be recovered soon so that it actually has no effect on the following.

VII. CONCLUSION

In this paper, we have presented an autonomous person-following framework for telepresence robots only using two web cameras. The framework combines the upper body

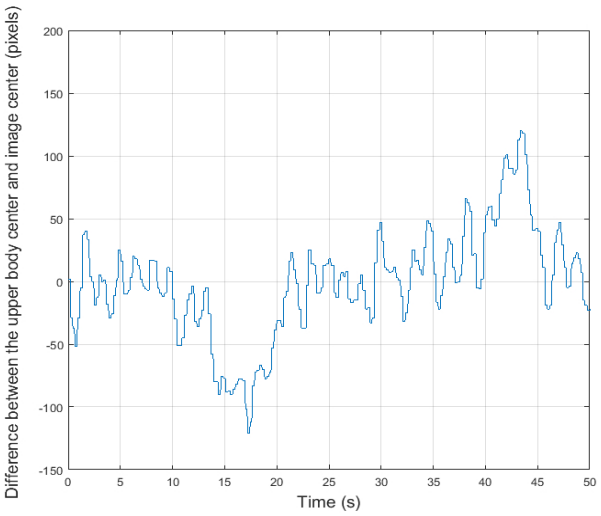

Fig. 11. The horizontal difference between the center of the upper body and the image center from the FF camera for a run. It changes drastically because of the turning.

tracking through an FF camera and the feet tracking through a DF camera to realize autonomous person-following robots without adding ranging sensors. The spatial constraint of upper body on feet can be used to prevent feet tracking from the disturbance of background, and the spatial constraint between two feet can ensure the robustness of trackers. The experimental results show the effectiveness of our method.

VIII. ACKNOWLEDGMENTS

This work was supported in part by Beijing Municipal Natural Science Foundation under Grant No. L172027, and the Natural Science Foundation of China (NSFC) under Grants No. 61702037 and No. 61773062, and Beijing Institute of Technology Research Fund Program for Young Scholars.

REFERENCES

[1] E. Paulos and J. Canny, "Prop: personal roving presence," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Press/Addison-Wesley Publishing Co., 1998, pp. 296–303.

[2] J. H. Lee, T. Tsubouchi, K. Yamamoto, and S. Egawa, "People tracking using a robot in motion with laser range finder," in *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*. Ieee, 2006, pp. 2936–2942.

[3] R. Gockley, J. Forlizzi, and R. Simmons, "Natural person-following behavior for social robots," in *Proceedings of the ACM/IEEE international conference on Human-robot interaction*. ACM, 2007, pp. 17–24.

[4] K. O. Arras, O. M. Mozos, and W. Burgard, "Using boosted features for the detection of people in 2d range data," in *Robotics and Automation, 2007 IEEE International Conference on*. IEEE, 2007, pp. 3402–3407.

[5] K. Zamani, G. Stavrinos, and S. Konstantopoulos, "Detecting and measuring human walking in laser scans," in *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*. ACM, 2018, p. 31.

[6] Z. Zhao, C. Fang, and Q. Ren, "People following system based on lrf," in *2018 11th International Workshop on Human Friendly Robotics (HFR)*. IEEE, 2018, pp. 78–83.

[7] B. X. Chen, R. Sahdev, and J. K. Tsotsos, "Person following robot using selected online ada-boosting with stereo camera," in *Computer and Robot Vision (CRV), 2017 14th Conference on*. IEEE, 2017, pp. 48–55.

[8] K. Shimura, Y. Ando, T. Yoshimi, and M. Mizukawa, "Research on person following system based on rgb-d features by autonomous robot with multi-kinect sensor," in *System Integration (SII), 2014 IEEE/SICE International Symposium on*. IEEE, 2014, pp. 304–309.

[9] A. Cosgun, D. A. Florencio, and H. I. Christensen, "Autonomous person following for telepresence robots," in *IEEE International Conference on Robotics and Automation*, 2013, pp. 4335–4342.

[10] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 8, pp. 1561–1575, 2017.

[11] Y. Li, S. Ding, Q. Zhai, Y. F. Zheng, and D. Xuan, "Human feet tracking guided by locomotion model," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2424–2429.

[12] 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.

[13] M. M. Loper, N. P. Koenig, S. H. Chernova, C. V. Jones, and O. C. Jenkins, "Mobile human-robot teaming with environmental tolerance," in *Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on*. IEEE, 2009, pp. 157–163.

[14] Y. Yoon, H. Yoon, and J. Kim, "Depth assisted person following robots," in *RO-MAN, 2013 IEEE*. IEEE, 2013, pp. 330–331.

[15] L. Pang, L. Zhang, Y. Yu, J. Yu, Z. Cao, and C. Zhou, "A human-following approach using binocular camera," in *Mechatronics and Automation (ICMA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1487–1492.

[16] T. Anezaki, K. Eimon, S. Tansuriyavong, and Y. Yagi, "Development of a human-tracking robot using qr code recognition," in *Frontiers of Computer Vision (FCV), 2011 17th Korea-Japan Joint Workshop on*. IEEE, 2011, pp. 1–6.

[17] A. Z. Shukor, N. A. Natrah, A. I. Tarmizi, A. A. Afiq, M. H. Jamaluddin, Z. A. Ghani, H. N. M. Shah, and M. Z. Ab Rashid, "Object tracking and following robot using color-based vision recognition for library environment," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 10, no. 2-7, pp. 79–83, 2018.

[18] C. Hu, X. Ma, and X. Dai, "A robust person tracking and following approach for mobile robot," in *Mechatronics and Automation, 2007. ICMA 2007. International Conference on*. IEEE, 2007, pp. 3571–3576.

[19] M. Tarokh and R. Shenoy, "Vision-based robotic person following in fast walking," 2014.

[20] 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.

[21] Y. Jia, B. Xu, J. Shen, M. Pei, Z. Dong, J. Hou, and M. Yang, "Telepresence interaction by touching live video images," *Computer Science*, 2016.

[22] J. Leng and Y. Liu, "Real-time rgb-d visual tracking with scale estimation and occlusion handling," *IEEE Access*, vol. 6, pp. 24 256–24 263, 2018.

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

[24] M. Danelljan, F. S. Khan, M. Felsberg, and J. V. D. Weijer, "Adaptive color attributes for real-time visual tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.