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

Target Tracking Control Based on Dual Model Fusion

Kang Li[1,2], Wenzhong Zha[1], Xiangrui Meng[1], Xiaoguang Zhao[2]

Abstract— **Recent years have witnessed rapid progress in target tracking. To track a moving target for mobile robots, however, both performance and speed of the algorithm are indispensable. This paper proposes a dual model fusion strategy to improve target tracking drift. Among them, the spatio-temporal context model in middle-level feature space (MFSTC) is utilized to ameliorate target tracking effect when illumination or appearance changes, the mean shift based on 3D back projection (MS3D) is fused to allow the algorithm to tackle occlusion and deformation. Tracking controller based on visual servo is also designed for mobile robots. We validate the efficiency of the proposed fusion model on the university of Birmingham RGB-D Tracking Benchmark (BTB) and show that our approach compares favorably with the state-of-the-arts, mobile robots using our approach can track targets robustly under various challenging scenes.**

I. INTRODUCTION

The objective of visual target tracking is to locate a target in consecutive video frames, for mobile robots, in the perspective of robotic vision. Over the past few years, there has been rapid progress in visual object tracking tasks, including augmented reality, automatic driving, intelligent video surveillance, robot navigation [1]–[4], etc. An amount of ongoing challenges still remain, such as handling object appearance changes, illumination variations, occlusions, shape deformations, and camera motion.

Fortunately, motivated by the fast development of affordable and increasingly reliable RGB-D sensors, and relatively large RGB-D datasets for visual tracking were compiled and released to the public [5], [6], numerous RGB-D trackers have been proposed [5]–[9]. Song et al. [5] developed an RGB-D tracker that performs quite well on the Princeton tracking benchmark. They adopt an exhaustive search paradigm coupled with an SVM classifier trained on color/depth HOG and point cloud features, which reduces the trackers runtime to only 0.26 FPS. Meshgi et al. [7] used an occlusion-aware particle filter framework to handle complex and persistent occlusions. However, this tracker uses a tightly constrained, predefined depth threshold. It therefore has difficulty tracking targets which exhibit large ranges of motion in the depth direction. Besides, it's speed is limited to 0.9 FPS, which is hard to use for mobile robots. Hannuna et al. [8] applied kernel correlation filters in color and depth maps, which demonstrates very promising performance with real-time speeds of up to 40 FPS. However, they might not make full use of the depth information and suffer the similar problem with [5]. Bibi et al. [9] proposed 3-D part-based tracker with automatic synchronization and registration. However, they judge the occlusion in terms of the thresholds computed in the first frame without updating, which fails in cases of significant and fast inwards/outwards movements.

Aforementioned methods are top four methods on the Princeton tracking benchmark, but [6] outlined four problems of this benchmark and publiced a new benchmark named the university of Birmingham RGB-D Tracking Benchmark (BTB), and proposed a RGB-D tracker using adaptive range-invariant depth models and spatio-temporal consistency constraints. Recently, a mean shift algorithm based on 3D back projection (MS3D) [10] with the best performance on BTB is proposed, which fused color and depth distribution efficiently in the mean-shift tracking scheme, however it is sensitive to appearance and illumination changes. In a sense, our previous work that the spatio-temporal context model in middle-level feature space (MFSTC) [11] can handle this problem, but it does't work well when the occlusion and deformation occurs.

The intuitive idea is to fuse the MS3D and MFSTC model to track a moving target. This paper proposed a target tracking control method based on dual model fusion for mobile robots. The dual model fusion mechanism is designed to give full play to advantages of MFSTC and MS3D model under various challenging scenes. Furthermore, we develop a tracking controller based on visual servo for mobile robots to track the moving target. The effectiveness and efficiency of our method is verified by the BTB dataset and real-life tracking experiments.

The rest of this paper is organized as follows. In Section II, we broadly introduce the overview of our approach and elaborate our RGB-D tracker. Section III provides details of our designed tracking controller. Experimental results and analysis are demonstrated in Section IV, and finally we give conclusions in Section V.

II. OUR APPROACH

A. Overview of our approach

Figure 1 shows the framework of the target tracking control method based on dual model fusion. The feedback of the entire closed-loop control system consists of our proposed RGB-D target tracking algorithm for real-time localization of the target's position in the robot's perspective. The deviation between the control expectation and the target position, as input of the tracking controller based on visual servo, can be calculated to adjust the tracking control amount to drive the

[1] Authors worked in Key Laboratory of Cognition and Intelligence Technology, China Electronics Technology Group Corporation Information Science Academy, CETC, Beijing, China lk530008489@163.com

[2] Author worked in the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China

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

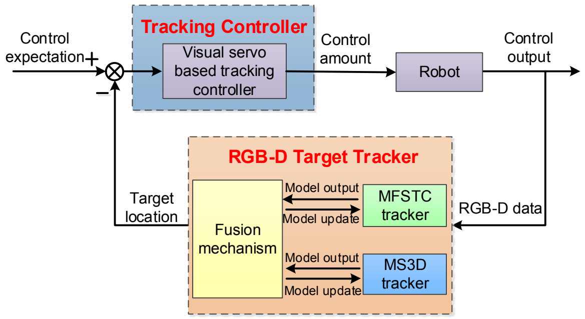


Fig. 1. Target tracking control algorithm framework based on dual model fusion.

robot to track the target robustly. The core is RGB-D tracker, which mainly includes three parts: MFSTC tracker, MS3D tracker and fusion mechanism.

B. Improved spatio-temporal context model

The traditional Spatio-Temporal Context (STC) model is fast for appearance modeling of targets, but it only models moving objects in RGB pixel space. When the light changes and the appearance changes, the target is easy to drift. In the KCF tracker, Henriques improved the generalization performance of the target appearance model by extracting the Histograms of Oriented Gradients (HOG). Inspired by it, our previous work proposed a spatio-temporal context model in middle-level feature space (MFSTC). The detail of the MFSTC tracker can be seen in [11].

C. Mean shift based on 3D back projection

Mean shift is a fast pattern matching algorithm based on kernel density estimation. Different from the spatio-temporal context model, it utilizes the feature similarity matching in each frame, which can not depend on the context of consecutive frames, the lost target can be re-detected. In our approach, a three-dimensional mean shift model (MS3D) based on RGB-D back projection [10] is fused, which integrates the depth information and designs the back projection image with color and depth probability density functions to improve the performance of target tracking.

D. Model fusion mechanism

Model fusion is an intuitive way to improve performance. The literature [6] proposed a two-layer structure tracking model: global processing layer and local processing layer. At the global processing layer, two KCF trackers are used to track the targets in the color image and the depth image, respectively, and then decide whether to open the cluster decision tree tracker using the local processing layer according to whether the tracking results of the two trackers are ambiguous. This two-layer tracking model architecture takes advantage of various trackers. Inspired by it, we design a fusion mechanism based on MFSTC model and MS3D model.

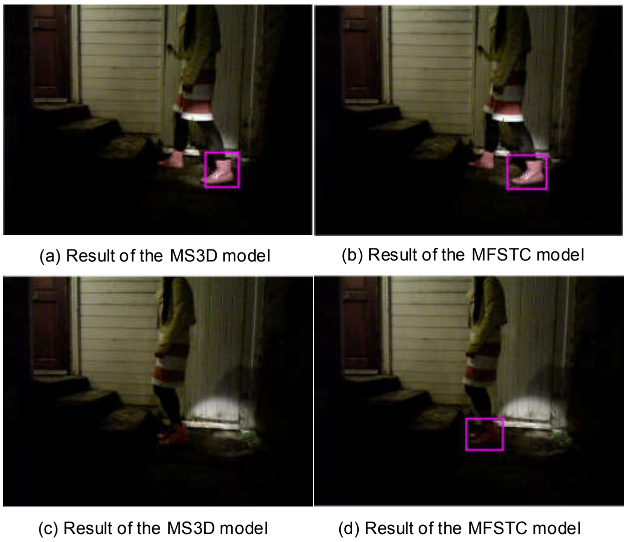

(a) Result of the MS3D model
(b) Result of the MFSTC model
(c) Result of the MS3D model
(d) Result of the MFSTC model

Fig. 2. Performance of different model under different illumination.

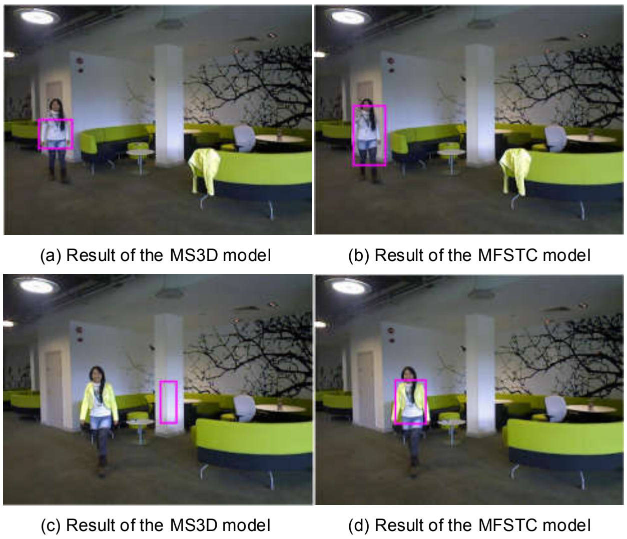
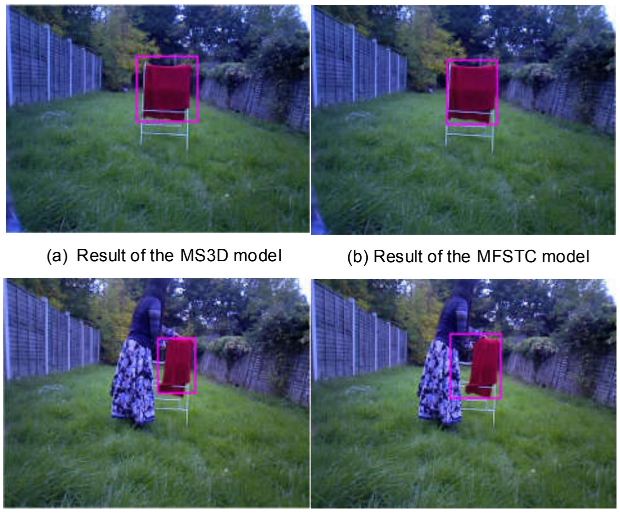

(a) Result of the MS3D model
(b) Result of the MFSTC model
(c) Result of the MS3D model
(d) Result of the MFSTC model

Fig. 3. Performance of different model when the appearance changes.

Since the color image analysis of the MS3D model still remains the traditional mean shift algorithm to model the target in the HSV color space, it is sensitive to illumination and appearance changes. While the MFSTC model extracts the spatio-temporal characteristics of the target, the feature expression is stronger. Figure 2 and Figure 3 show the processing of the MS3D model and MFSTC model under different lighting conditions and different appearance changes. It can be seen from the figure 2 that the back projection probability distribution of target on MS3D model has undergone a serious change, causing the shoes to be lost, while the performance of the MFSTC model is more robust when the illumination changes. From the figure 3, we can see that after the target has changed clothes of different colors, the MS3D model drifts, and the MFSTC can still correctly locate the target. Therefore, the proposed tracker first uses the MFSTC model to track the target frame by frame, so as to take full advantages of the robust target tracking of



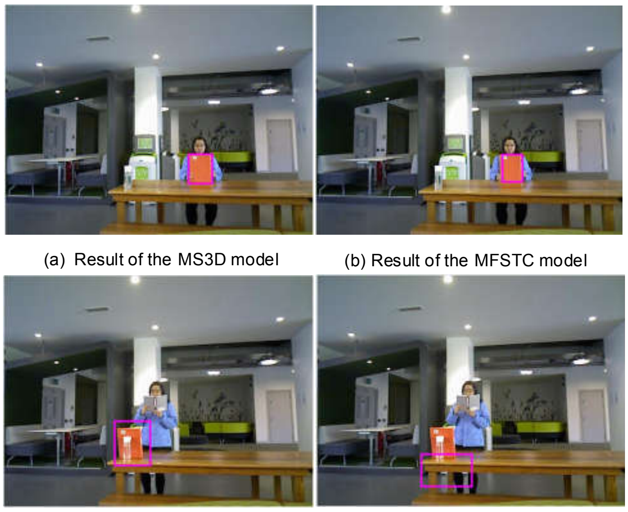
(a) Result of the MS3D model

(b) Result of the MFSTC model

(c) Result of the MS3D model

(d) Result of the MFSTC model

Fig. 4. Performance of different model when the deformation occurs.



(a) Result of the MS3D model

(b) Result of the MFSTC model

(c) Result of the MS3D model

(d) Result of the MFSTC model

Fig. 5. Performance of different model when the occlusion occurs.

MFSTC model for illumination and appearance changes.

However, when a challenge such as deformation or occlusion occurs during the target tracking process, the MFSTC model inevitably drift target due to lacking of various challenge coping strategies designing. Figure 4 and Figure 5 show the performance of the MS3D model and the MFSTC model when deformation and partial occlusion occur, respectively. It can be seen that the MS3D model performs better for the above challenges, so the advantages of the MS3D model should be exploited at this time.

When the deformation and occlusion conditions are encountered during the execution of the MFSTC model, it is usually accompanied by a case where the maximum response value becomes smaller. When the maximum response value c_{max} of the model is less than a certain empirical parameter η_c, it reflects the difference between the current candidate target and the tracking target to some extent, and tracking drift is likely to occur. When the candidate target is occluded, it is usually accompanied by a small depth region having a larger number of pixels in the ROI depth distribution. Assume that the depth value of the target ROI region is subject to a Gaussian distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \qquad (1)$$

where x is the depth value of the target ROI area, μ and σ are the mean and standard deviation of the depth value of the target ROI area, respectively. According to the concentration of the Gaussian distribution (3σ principles), pixels in the range of $[\mu-3\sigma, \mu+3\sigma]$ are considered as targets, and pixels smaller than $\mu-3\sigma$ are likely to be occluders, then calculate:

$$O = \sum\nolimits_{d=0}^{\mu-3\sigma} h(d) \Big/ \sum\nolimits_{d} h(d) \qquad (2)$$

where $h(d)$ represents the number of pixels in the dth cube of the target depth histogram, and $d=0$ is the depth at which the camera is located. When O is greater than an empirical parameter η_o, the target is likely to track failure. When the above two conditions are met in the tracking process, the confidence of the MFSTC model tracking failure is very high. At this time, the MS3D model real-time positioning target will be turned on, and the MFSTC model is updated according to the tracking result of MS3D until the model response value is renewed. If it is greater than the empirical parameter and the depth distribution returns to normal, re-switch to the MFSTC model tracking target. Through the above fusion mechanism, the advantages of the two models can be fully utilized, thereby improving the robustness of target tracking.

III. TARGET TRACKING CONTROL BASED ON VISUAL SERVO

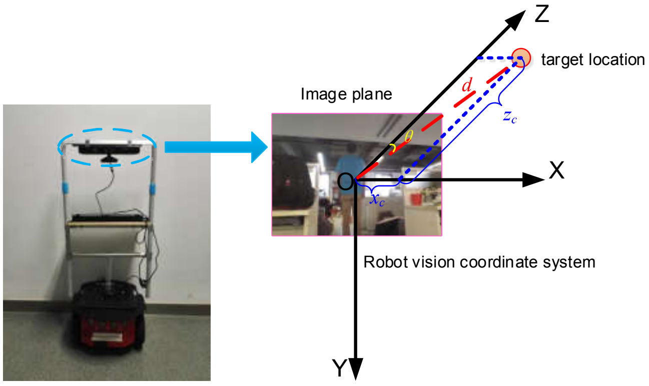


Fig. 6. Robot vision coordinate system

As shown in Figure 6, in order to make the robot safely and stably track the target, the distance d between the robot and the target can be controlled to be the safe distance d_s, and the angle θ is always 0°. According to the coordinates(x_c, y_c, z_c) of the target center point in the

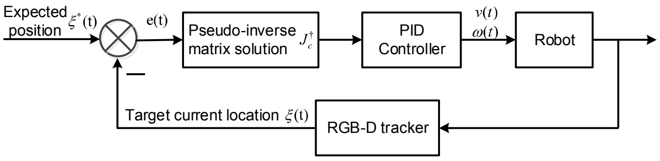

Expected position ξ^*(t)
e(t)
Pseudo-inverse matrix solution $J_c^\dagger$
PID Controller
$v(t)$
$\omega(t)$
Robot
—
Target current location ξ(t)
RGB-D tracker

Fig. 7. Target tracking closed-loop control diagram

robot vision coordinate system, the distance and angle of the current robot from the tracking target can be calculated by:

$$d = \sqrt{{x_c}^2 + {z_c}^2} \tag{3}$$

$$\theta = \arctan(x_c / z_c) \tag{4}$$

Thereby, the amount of change in the control deviation of the target center point is calculated. As shown in Figure 7, defining the desired position of the target center point in the robot's visual coordinate system as $\xi^*(t)$, and the current position is $\xi(t)$, then system deviation $e(t) = \xi^*(t) - \xi(t)$. Designing the PID controller to adjust the robot's linear speed $v(t)$ and angular velocity $\omega(t)$:

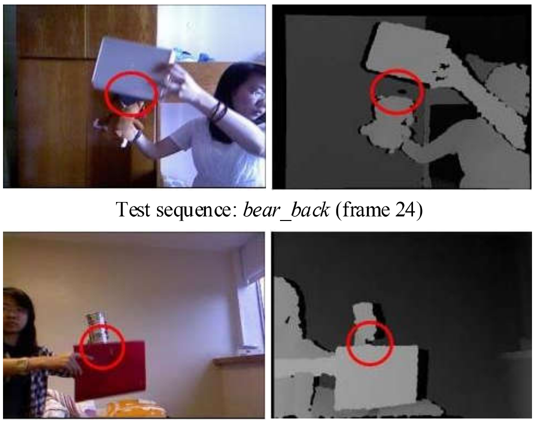
$$\begin{bmatrix} v(t) & \omega(t) \end{bmatrix}^T = -\lambda J_c^\dagger [k_p e(t) + k_i \textstyle\sum_{t=0}^{T} e(t) + k_d \left(e(t) - e(t-1)\right))] \tag{5}$$

where k_p, k_i, and k_d are proportional, integral, and differential coefficients, respectively.

IV. EXPERIMENTS AND ANALYSIS

A. Experiments with database

As mentioned above, we evaluated the proposed tracker on a new public RGB-D object tracking dataset built by [6] rather than the famous PTB benchmark [5]. In [6], the author outlined 4 problems of PTB, where the most serious problem is that it contains some sequences where the color and depth images pairs are unsynchronized (see Figure.8), and the majority of the benchmark videos are captured by a stationary camera. The BTB dataset comprises 36 video sequences in different challenge scenarios. The time-stamped color image and depth image are time-aligned, and two sets of sequences for each scene are recorded by still camera and moving camera respectively. Ten different attributes are counted by manual labeling and computer labeling to distinguish video. The ten attributes are: Illumination variation (IV), Depth variation (DV), Scale variation (SV), Color distribution variation (CDV), Depth distribution variation (DDV), Surrounding depth clutter (SDC), Surrounding color clutter (SCC), Background color confusion (BCC), Background shape camouflages (BSC), Partial occlusion (PO). A state-of-the-art methodologies named OTB [12] is selected as evaluation index. We present the results of comparing the proposed tracker with the top 4 RGB-D trackers on the PTB (Princeton tracker (PT) [5], depth-scaled correlation filter model (DS-KCF) [13], shape-adaptive DS-KCF* model (DS-KCF*) [8], occlusion-aware particle filter model (OAPF) [7], baseline algorithm (STC) [6] and the 3D mean shift algorithm (MS3D) with the best performance on BTB [10].



Test sequence: *bear_back* (frame 24)

Test sequence: *cup_book* (frame 11)

Fig. 8. An example of the color and depth images pairs are unsynchronized in the Princeton tracking benchmark

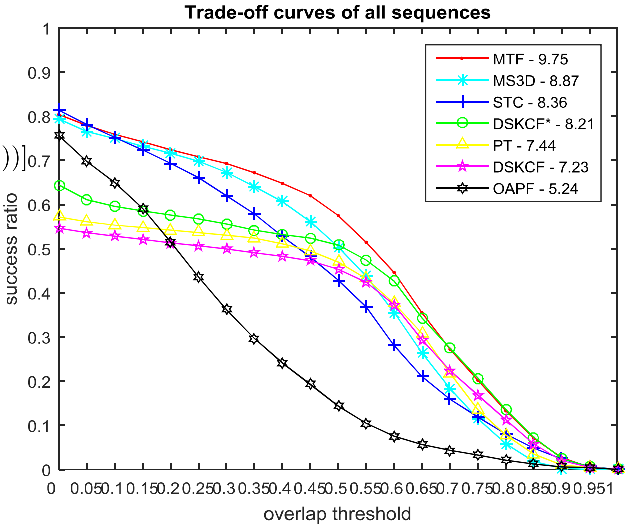


Fig. 9. Trade-off curves of all sequences

Qualitative Evaluation As shown in Figure 10, in the sequence *Backpack*, the red trash in the background of the 287th frame interferes with the judgment of MS3D model, and it can be seen from the video key frame that when other models fail, our Mfstc-ms3d Tracker Fusion (MTF) model can still locate the target; in the sequence *Face*, the book blocks part of the face at the 42th frame, only the MS3D, STC and MTF model can work. Until the 67th frame, only MTF model can follow the face and other models drift. In the sequence *Book*, only the MTF model can track the orange book at the 142th frame. At the 302th frame, the book is opened, the appearance and shape change drastically, other models are completely ineffective, and the MTF can still track part of the target; in the sequence *Toytank*, as the light changes and the shape changes, MS3D Model tracking fails, and the MTF model can still track toy tanks. Experimental results demonstrate our MTF model achieves competitive performance in a variety of challenging situations.

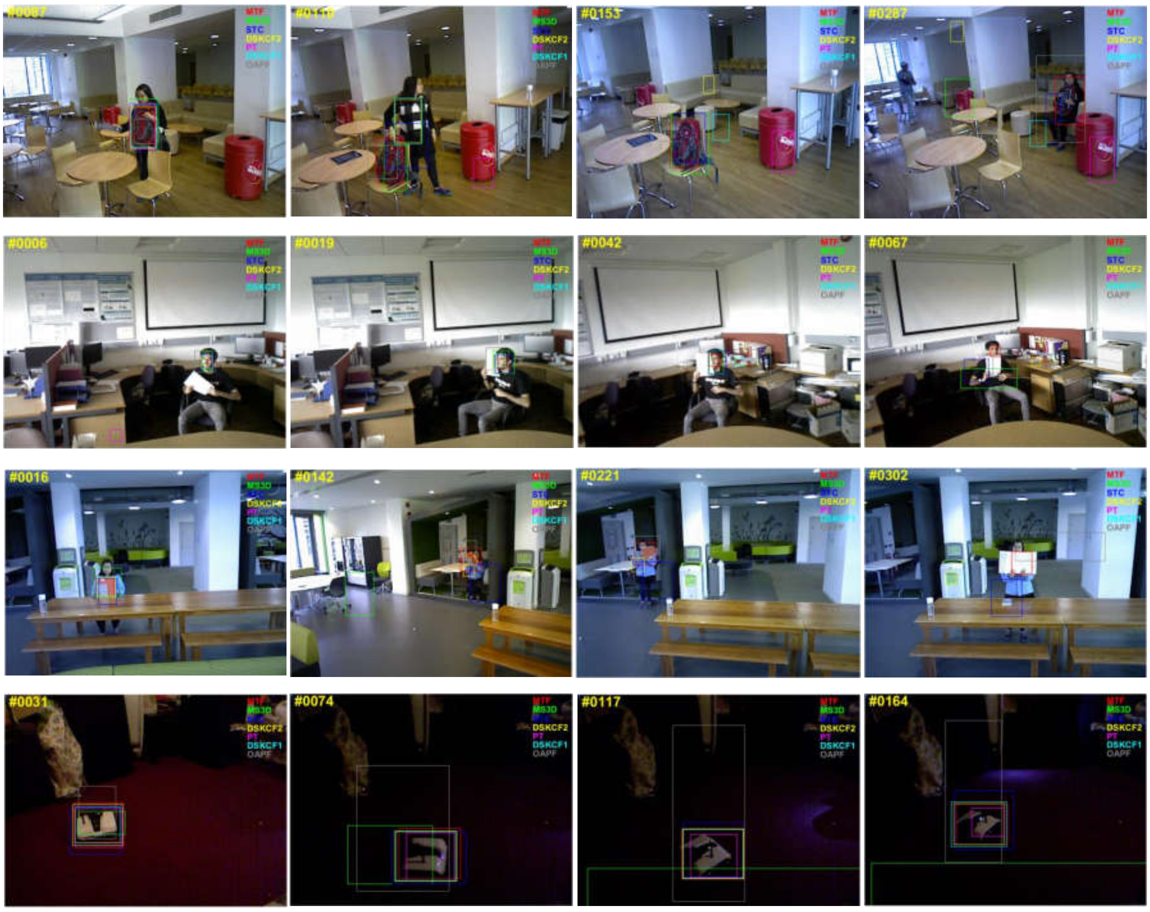


Fig. 10. A visual review of the trackers' performance in sequences with various of challenges. The sequence from top to bottom is: *Backpack*, *Face*, *Book*, *Toytank*. It contains complex challenges such as background confusion, depth changes, partial occlusion, appearance and lighting changes, etc.

TABLE I

AUC OF BOUNDING BOX OVERLAP, RED DENOTES BEST PERFORMING TRACKER AND BLUE DENOTES THE SECOND

Algorithm	Overall	Stationary	Moving	IV	DV	SV	CDV	DDV	SDC	SCC	BCC	BSC	PO	FPS
MTF (Ours)	9.75	10.42	9.12	7.34	8.90	6.47	5.43	8.72	8.87	10.07	8.81	8.69	9.21	22~37
MS3D	8.87	10.2	7.63	4.80	7.63	5.89	3.13	8.82	8.48	9.57	8.17	8.24	8.08	30~50
STC	8.36	9.57	7.18	5.78	7.56	5.07	5.12	7.66	8.01	9.53	6.67	7.17	7.73	14.73
DS-KCF*	8.21	9.36	7.13	6.10	7.88	4.39	0.94	5.26	7.93	9.81	5.66	6.50	7.76	17.89
PT	7.44	8.54	6.43	4.15	6.65	2.80	0.43	3.48	6.79	8.23	5.74	5.76	6.20	0.14
DS-KCF	7.23	7.52	6.85	5.50	7.06	3.36	1.43	4.16	7.89	8.25	4.82	5.16	6.02	20.95
OAPF	5.24	6.0	4.54	3.19	4.45	3.07	3.22	3.71	5.00	6.13	3.79	4.82	5.82	1.30

Quantitative Evaluation Area under the curve (AUC) is exploited to evaluate the performance of state-of-the-art RGB-D trackers. The trade-off curves of all sequences on the BTB is shown in Figure 9. The overall performance has been improved by 0.88 by comparing MTF model with MS3D. Table I shows the AUC performance of RGB-D trackers in different attributes. On the one hand, Our tracker outperforms others in both IV, CDV and SCC, which reflects that MFSTC model can mitigate tracking drift since the illumination or appearance changes. On the other hand, The performance on DV, SV, BSC and PO also demonstrates that the MTF model takes advantage of the MS3D model in occlusion and depth changes. In addition, the average frame rate has been arrived at 22∼37 FPS, which satisfied the real-time requirements of robot tracking.

B. Experiments with mobile robot

In order to further verify the effectiveness of the proposed MTF model, we executed the robot tracking experiment in real scene. The safe tracking distance d_s is set to 1.8 m,

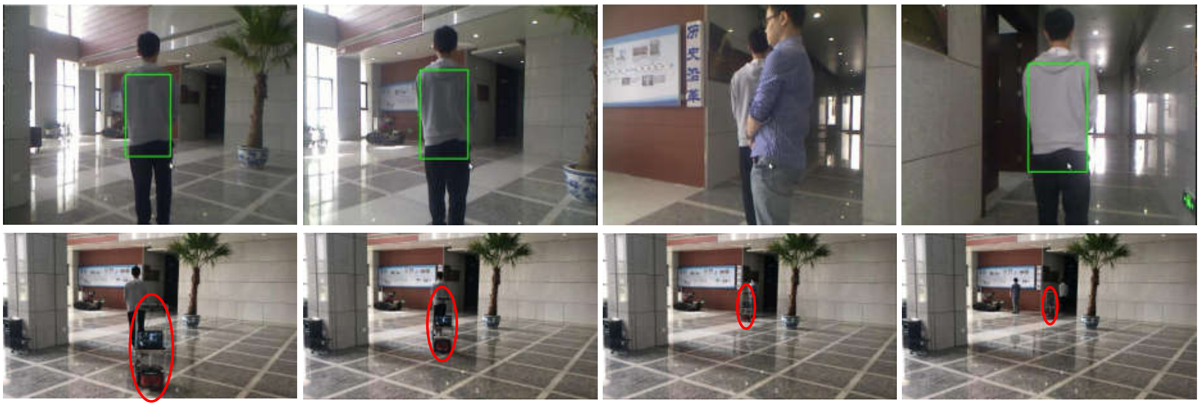


Fig. 11. Experimental results of the MTF tracking model under partial occlusion. The first row represents the tracking result of the robot perspective, and the second row represents the tracking result of the third angle.

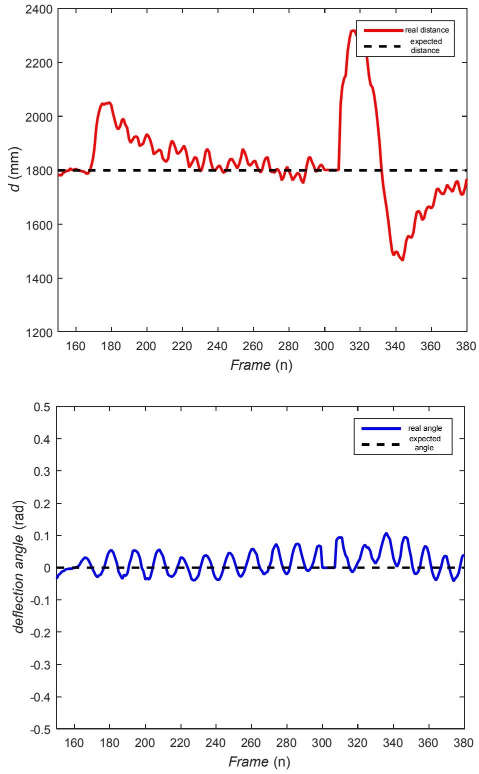


Fig. 12. Distance and angle change between the robot and moving target

and the ideal angle between the robot and the target is 0°. Figure 11 and Figure 12 shows the performance of our tracker when the partial occlusion is occurred. we can see that the target is not lost, and the distance d and declination fluctuates around the safe distance d_s and 0° respectively.

V. CONCLUSIONS

In this work, we proposed a novel and practical RGB-D target tracking control method for robot tracking task. Both MFSTC tracker and MS3D tracker are integrated to improve the tracking effect. Tracking controller based on visual servo is also designed. At last, we evaluated our tracker on a new RGB-D tracking dataset named BTB, which runs in real time and ranks in the first position. Furthermore, we choose a challenging occlusion test for robot tracking, our robot can track the moving target robustly.

REFERENCES

[1] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, “Visual tracking: An experimental survey,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1442–1468, 2013.

[2] X. Li, W. Hu, C. Shen, Z. Zhang, A. Dick, and A. V. D. Hengel, “A survey of appearance models in visual object tracking,” *ACM transactions on Intelligent Systems and Technology (TIST)*, vol. 4, no. 4, p. 58, 2013.

[3] A. Yilmaz, O. Javed, and M. Shah, “Object tracking: A survey,” *Acm computing surveys (CSUR)*, vol. 38, no. 4, p. 13, 2006.

[4] P. Li, D. Wang, L. Wang, and H. Lu, “Deep visual tracking: Review and experimental comparison,” *Pattern Recognition*, vol. 76, pp. 323–338, 2018.

[5] S. Song and J. Xiao, “Tracking revisited using rgbd camera: Unified benchmark and baselines,” in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 233–240.

[6] J. Xiao, R. Stolkin, Y. Gao, and A. Leonardis, “Robust fusion of color and depth data for rgb-d target tracking using adaptive range-invariant depth models and spatio-temporal consistency constraints,” *IEEE transactions on cybernetics*, vol. 48, no. 8, pp. 2485–2499, 2018.

[7] K. Meshgi, S.-i. Maeda, S. Oba, H. Skibbe, Y.-z. Li, and S. Ishii, “An occlusion-aware particle filter tracker to handle complex and persistent occlusions,” *Computer Vision and Image Understanding*, vol. 150, pp. 81–94, 2016.

[8] S. Hannuna, M. Camplani, J. Hall, M. Mirmehdi, D. Damen, T. Burghardt, A. Paiement, and L. Tao, “Ds-kcf: a real-time tracker for rgb-d data,” *Journal of Real-Time Image Processing*, pp. 1–20, 2016.

[9] A. Bibi, T. Zhang, and B. Ghanem, “3d part-based sparse tracker with automatic synchronization and registration,” in *Computer Vision & Pattern Recognition*, 2016.

[10] Y. Zhao and E. Menegatti, “Ms3d: Mean-shift object tracking boosted by joint back projection of color and depth,” in *International Conference on Intelligent Autonomous Systems*. Springer, 2018, pp. 222–236.

[11] K. Li, X. Zhao, S. Sun, and M. Tan, “Robust target tracking and following for a mobile robot,” *International Journal of Robotics and Automation*, vol. 33, no. 4, 2018.

[12] Y. Wu, J. Lim, and M.-H. Yang, “Object tracking benchmark,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, 2015.

[13] M. Camplani, S. L. Hannuna, M. Mirmehdi, D. Damen, A. Paiement, L. Tao, and T. Burghardt, “Real-time rgb-d tracking with depth scaling kernelised correlation filters and occlusion handling.” in *BMVC*, 2015, pp. 145–1.