TSDM: Tracking by SiamRPN++ with a Depth-refiner and a Mask-generator

Pengyao Zhao, Quanli Liu*, Wei Wang, Qiang Guo
Key Laboratory of Intelligent Control and Optimization for Industrial Equipment, Dalian University of Technology
yaoyaoasdzxc@163.com

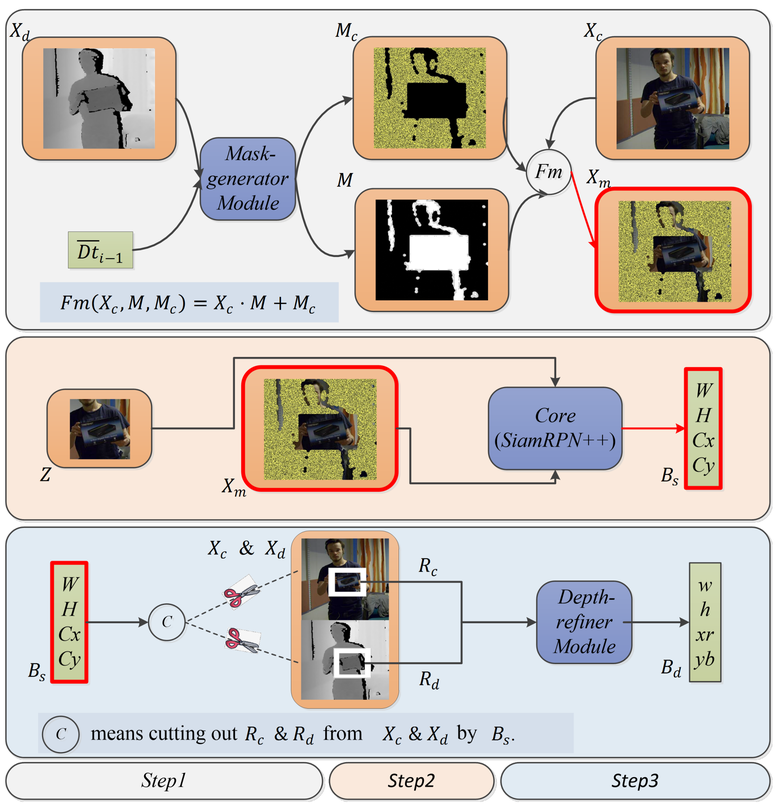
1. Introduction

Depth images provide informative cues for a generic object tracking. However few trackers have used depth information due to the lack of a suitable model. In this paper, a RGB-D tracker named TSDM is proposed, which is composed of a Mask-generator (M-g), SiamRPN++[1] and a Depth-refiner (D-r). The M-g generates the background masks, and updates them as the target 3D position changes. The D-r optimizes the target bounding box based on the depth difference between the target and the surrounding background. Extensive evaluation on the Princeton Tracking Benchmark and the Visual Object Tracking challenge shows that our tracker outperforms the state-of-the-art while achieving 31 FPS. Code and models of TSDM are available at https://github.com/lql-team/TSDM.

2. Contribution

1. Propose TSDM, a new RGB-D tracker, which can ignore background distractors and output an accurate target state.
2. Propose two novel depth modules which can overcome the obstacle above and make use of depth information effectively.

3. Method pipeline



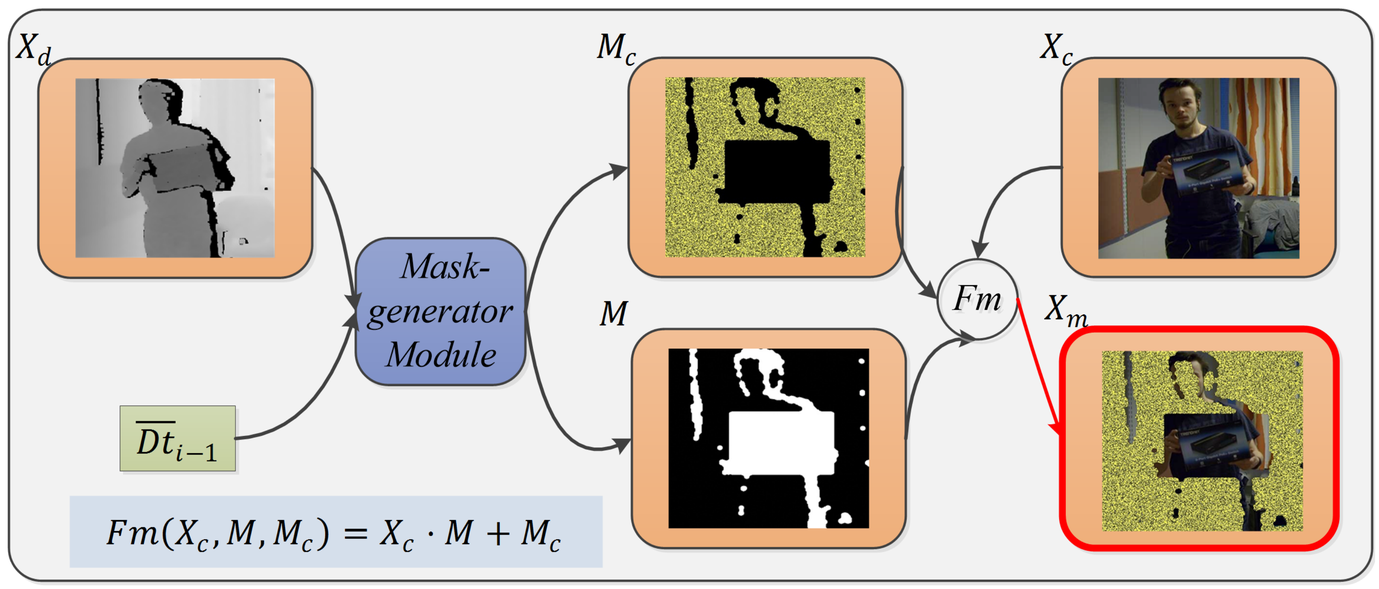
1. Input X_d and $\overline{Dt}_{i-1}$ into M-g to get M and M_c. Then use $F_m(\cdot)$ to get X_m.
2. Input Z and X_m into the core. Then the core outputs the target bounding box B_s.
3. Cut out R_c and R_d from X_c and X_d by B_s respectively. Then input R_c and R_d into D-r to get the refined target bounding box B_d.

4. Mask-generator

For reducing the interference of background distractors and clearing out irrelevant image information to the target, M-g generates two background mask images M and M_c by Eq. 1 and Eq. 2 respectively. Where M makes fore-background separation, and M_c colors the background.

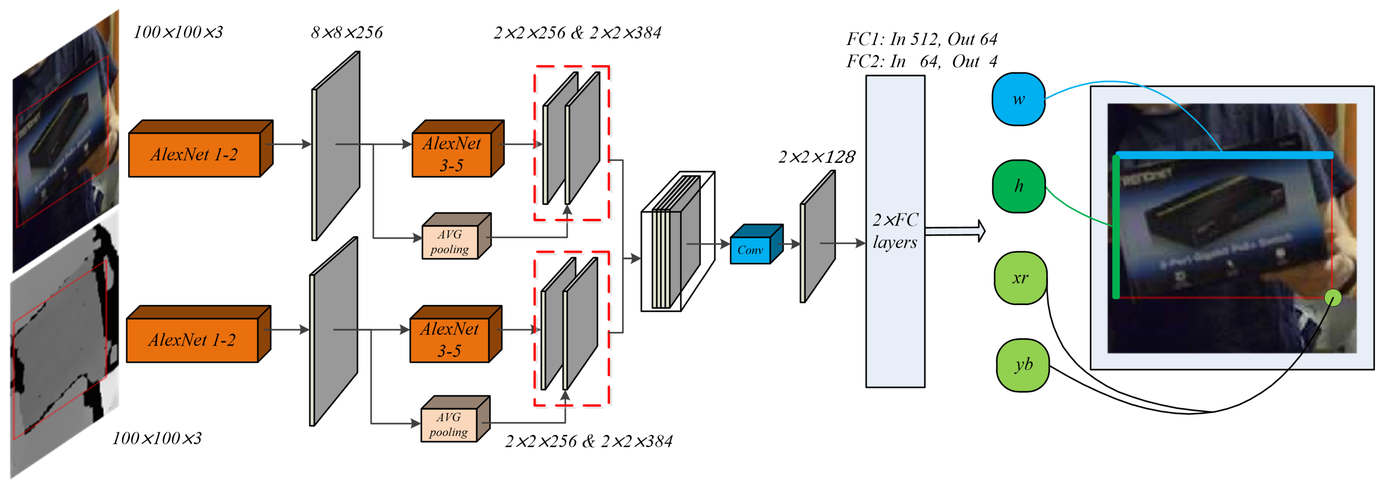
$$M(x,y) = \begin{cases} 1, & out\,of\,depth\,range \\ 0, & otherwise \end{cases} \tag{1}$$

$$M_c(x,y,c) = \rho(x,y,c) \times [1 - M(x,y)], \quad c \in \{1,2,3\} \tag{2}$$



5. Depth-refiner

In essentially, D-r can be treated as an information fusion network. It uses depth information to optimize the target state, and meanwhile, color as the correction information overcomes the slight color-depth mismatch. By integrating the two information, the network can output a more accurate bounding box. In the network, we adopt double Alexnet[2] rather than a single new 4-channel-input net as the backbone for a quick transfer learning; the 1×1 convolution layer fuses cross-channel information and reduces feature map dimension; the two fully connected layers are used to output the refined bounding box.



6. Result on benchmark

On the PTB[3], TSDM is compared with 7 trackers in the baseline. All these trackers use depth information but are not based end-to-end CNN. Table II shows the average IOU overlap of each category and the overall average IOU overlap. Our tracker performs the best for Overall.

	Average IOU overlap											
Mehtod	Overall	*Human*	*Animal*	*Rigid*	*Large*	*Small*	*Slow*	*Fast*	*Occ.*	*No-Occ.*	*Passive*	*Active*
TSDM	**0.792**	0.71(6)	**0.85(1)**	**0.86(1)**	0.77(2)	**0.81(1)**	**0.87(1)**	**0.76(1)**	0.69(5)	**0.94(1)**	0.84(3)	**0.78(1)**
OTR [12]	0.769	0.77(2)	0.68(3)	0.81(3)	0.76(4)	0.77(2)	0.81(2)	0.75(2)	0.71(2)	0.85(4)	**0.85(1)**	0.74(2)
ECO-TA [18]	0.754	0.77(3)	0.65(5)	0.80(4)	0.77(3)	0.74(4)	0.79(5)	0.41(8)	0.68(6)	0.85(3)	0.84(2)	0.72(4)
3D-T [2]	0.750	**0.81(1)**	0.64(6)	0.73(8)	**0.80(1)**	0.71(7)	0.75(8)	0.75(3)	**0.73(1)**	0.78(6)	0.79(7)	0.74(3)
CSR-rgbd++[11]	0.740	0.77(4)	0.65(4)	0.76(7)	0.75(5)	0.73(5)	0.80(4)	0.72(4)	0.70(3)	0.79(5)	0.79(6)	0.72(5)
Ca3dms [16]	0.737	0.66(8)	0.74(2)	0.82(2)	0.73(6)	0.74(3)	0.80(3)	0.71(6)	0.63(8)	0.88(2)	0.83(4)	0.70(6)
DM-DCF [10]	0.726	0.76(5)	0.58(8)	0.77(5)	0.72(7)	0.73(6)	0.75(7)	0.72(5)	0.69(4)	0.78(8)	0.83(5)	0.69(7)
DS-KCF [6]	0.693	0.67(7)	0.61(7)	0.76(6)	0.69(8)	0.70(8)	0.75(6)	0.67(7)	0.63(7)	0.78(7)	0.79(8)	0.66(8)

7. References

References

[1] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing and J. Yan. SiamRPN++: Evolution of siamese visual tracking with very deep networks. In CVPR, 2019.

[2] A. Krizhevsky, I. Sutskever and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012.

[3] S. Song and J. Xiao. Tracking revisited using RGBD camera: Unified benchmark and baselines. In ICCV, 2013.