FC-TrackNet: Fast Convergence Net for 6D Pose Tracking in Synthetic Domains

Di Jia[1*], Qian Wang[1*†], Jun Cao[2], Peng Cai[1], Zhiyang Jin[1]
[1] Liaoning Technical University
[2] Intel Corporation
jiadi@lntu.edu.cn, lntu_wangqian@163.com, jun.cao@intel.com, lntu_caipeng@163.com, lntu_jzy@163.com

Abstract

In this work, we propose a fast convergence track net, or FC-TrackNet, based on a synthetic data-driven approach to maintaining long-term 6D pose tracking. Comparison experiments are performed on two different datasets, The results demonstrate that our approach can achieve a consistent tracking frequency of 90.9 Hz as well as higher accuracy than the state-of-the art approaches.

Introduction

Estimating the six-dimensional (6D) pose of an object and accurately tracking it from an image sequence is an essential task for virtual reality, augmented reality, and robotic manipulations. The single-image estimation approach requires initializing the pose of the object in each frame in the video sequence, without considering the spatial and temporal information features with a similar pose of the object in frames, which requires intensive calculation and generates a large amount of redundant data when tracking the video sequence (Deng et. al. 2020; Deng et. al. 2021; He et. al. 2020; Issac et. al. 2016; Li, Wang, and Ji 2019; Mitash et. al. 2020; Sundermeyer et. al. 2018; Wang et. al. 2019; Xiang et. al. 2018). At present, deep learning–based approaches have become mainstream in 6D pose estimation for objects, which significantly improves the accuracy and robustness of pose estimation (Pavlakos et. al. 2017; Kehl et. al. 2017; Peng et. al. 2020; Zeng et. al. 2017). Tracking the 6D pose of the objects in video sequences requires extensive, hand-annotated, real data for training, which are costly to acquire and label (Li et. al. 2020). The availability of synthetic data enables easy access to training data, which can provide sufficient simulation variability for the network during training, and the model can be generalized for real usage during testing (Tobin et. al. 2017). We propose a fast convergent track net, or FC-TrackNet, which is based on a synthetic data-driven approach to allow long-term, stable, and high-precision pose tracking of objects. It has strong robustness against severe occlusion of the object, and it can be widely used in real scenarios, as shown in figure 1. The experiment results show that the pose tracking of our approach outperforms that of other state-of-the-art approaches under both small and large datasets, so it has the potential for wide application in reality.

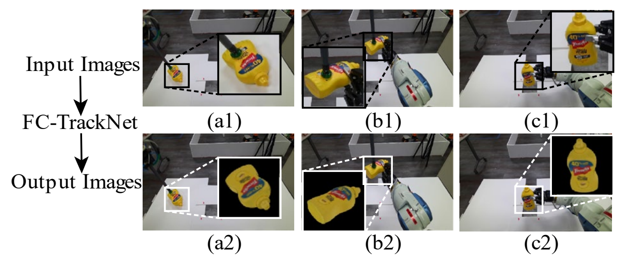

Figure 1: (a1–c1) are the three input images, and (a2–c2) are the predicted pose images through FC-TrackNet.

Approach

Our proposed network is shown in figure 2. The input data are two groups of images containing RGBD information. At the training stage, we use synthetic data, and at the test stage, the input images are the real data. O_t is the current observed image tensor, and R_{t-1} is the rendered image tensor of the network as the output of the previous frame. The relative pose transition (ΔP_t) and the characteristic discrepancy (Φ) of the predicted pose after network processing are obtained, where ΔP_t is the change in pose from P_{t-1} to O_t, and the characteristic discrepancy Φ is

$$\Phi = \delta(\psi_{I1}(\overline{P}) - \psi_{I2}(P)),$$

where δ denotes the predefined robust loss function, $\psi()$ denotes the direct pixel intensity values, P and $\overline{P}$ are the object pose, and the cost function is determined by P and $\overline{P}$ to measure the characteristic discrepancy of pose Φ. The current predicted pose (P_t) is the forward propagation result

[*]Corresponding author. [†]Equal contributions.

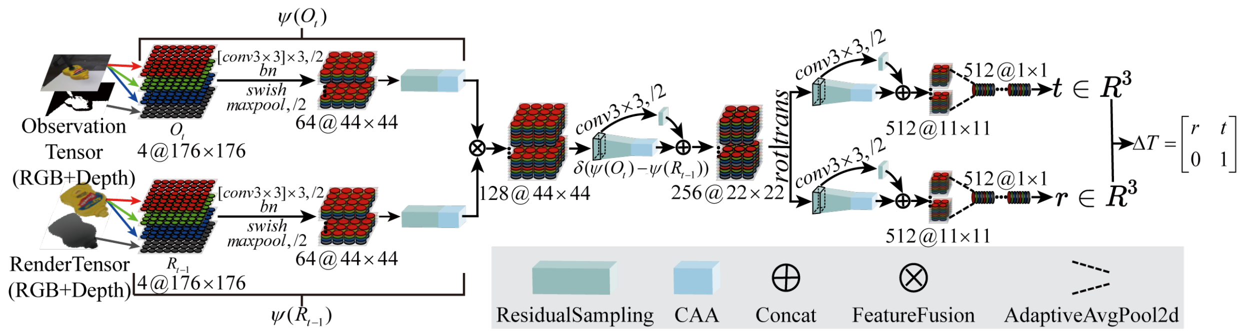

Figure 2: The RGB-D observation image (O_t) at frame t of the video sequence and the RGB-D rendered image (R_{t-1}) at frame $t-1$ are used as inputs. The results is used to predict the translation matrix (t) and rotation matrix (r) of the object.

by the relative pose transition (ΔP_t). The optimal relative transition solution (ΔP_t^*) is

$$\Delta P_t^* = \operatorname{argmin}\left\{\delta(\psi_{O_t}(P_t) - \psi_R(P_{t-1}) - J(P_{t-1})\Delta P_t\right\},$$

where J is the Jacobian matrix of pixel intensity values (ψ_R) with respect to the object pose P. The loss function of rotation and shift is calculated by the mean squared error. For the network structure of this paper, t is the object shift difference, and r is the object rotation difference. The MSE is rewritten as follows to obtain L :

$$L = (t - \bar{t})^2 + (r - \bar{r})^2 ,$$

We decompose the extracted features into one-dimensional feature encoding with two differents directional components aggregated along the horizontal and vertical coordinates. The final output $y(i,j)$ is obtained as follows:

$$y(i,j) = x_c(i,j) \times T_c^h(i) \times T_c^w(j) ,$$

where $x_c(i,j)$ is the original feature map, horizontal component (t^w) of size $1 \times w$ and a vertical component (t^h) of size $h \times 1$ after systematic transformation, which is increased channels convolutionally transformed by 1×1 F_w , F_h , and the sigmoid function $\sigma(x)$ to generate the horizontal component $T^w = \sigma(F_w(t^w))$ and the vertical component $T^h = \sigma(F_h(t^h))$.

Experiments

Two benchmark datasets, YCB-Video (Xiang et. al. 2018) and YCBInEoAT (Wen et. al. 2020), are selected for training and validation. Object pose reinitialization is not allowed in the evaluation owing to the large cost and other errors. Moreover, area under the curve (AUC) and average distance of model points (ADD) are used to evaluate the results in the video sequence(Xiang et. al. 2018). All the results(Ge and Loianno 2021; Issac et. al. 2016; Li et. al. 2020; Jonathan et. al. 2018; Wang et. al. 2019; Wen et. al. 2020; Wüthrich et. al. 2013) are shown in Figure 3. Our approach uses the same PPDR for synthetic data generation as the state-of-the-art se(3)-TrackNet (Wen et. al. 2020), and the small datasets 6k, 8k and 10k are selected for evaluation. Our method is significantly better than the comparison group on different numbers of small datasets, and is able to reach 90.11% and 94.99% for ADD and ADD-S metrics at 10k. Our approach still provides tracking closer to the real pose of the object, demonstrating that the proposed network has a high convergence speed.

Conclusion

In this work, we have proposed a new network structure, FC-TrackNet to provide long-term effective object 6D pose tracking with only one initialization. The network can quickly reach a state of network convergence by using a small amount of synthetic data, achieve ideal tracking performance in both severe occlusion and drastic motion tests.

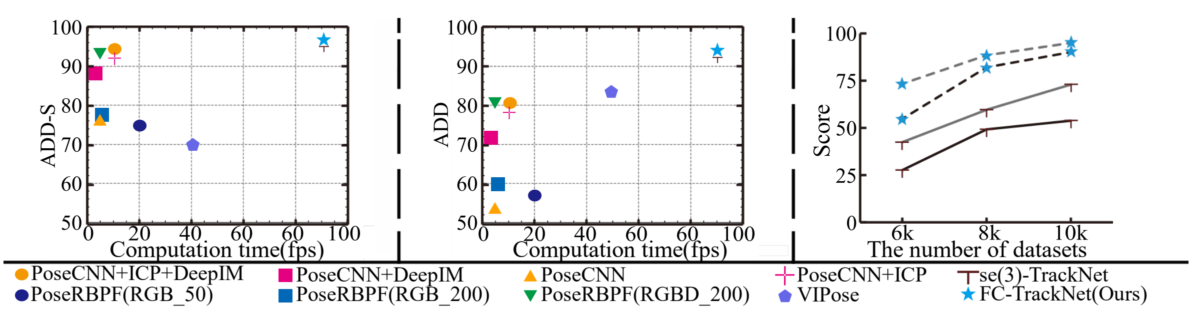

Figure 3. Comparison of methods in the complete YCB-Video(left and middle) and in the small YCB-InEOAT(right)

References

Deng, X.; Xiang, Y.; Mousavian, A.; Eppner, C.; Bretl, T.; and Fox, D. 2020. Self-supervised 6D Object Pose Estimation for Robot Manipulation. In proceedings of the IEEE International Conference on Robotics and Automation.Paris: Institute of Electrical and Electronics Engineers. doi.org/10.1109/ICRA40945. 2020. 9196714.

Deng, X.; Mousavian, A.; Xiang, Y.; Xia, F.; Bretl, T.; and Fox, D. 2021. PoseRBPF: A Rao–Blackwellized Particle Filter for 6-D Object Pose Tracking. IEEE Transactions on Robotics 37(5): 1328–1342. doi.org/10.1109/TRO. 2021. 3056043.

Ge, R.; and Loianno, G. 2021. VIPose: Real-time Visual-Inertial 6D Object Pose Tracking. In proceedings of the Intelligent Robots and Systems.Prague: Institute of Electrical and Electronics Engineers. doi.org/10.1109/IROS51168. 2021. 9636283.

He, Y.; Sun, W.; Huang, H.; Liu, J.; Fan, H.; and Sun, J. 2020. PVN3D: A Deep Point-Wise 3D Keypoints Voting Network for 6DoF Pose Estimation. In proceedings of the Computer Vision and Pattern Recognition.Washington: Computer Vision and Pattern Recognition. doi.org/10.1109/CVPR42600. 2020. 01165.

Issac, J.; Wüthrich, M.; Cifuentes, C. G.; Bohg, J.; Trimpe, S.; and Schaal, S. 2016. Depth-based object tracking using a Robust Gaussian Filter. In proceedings of the IEEE International Conference on Robotics and Automation.Stockholm: Institute of Electrical and Electronics Engineers. doi.org/10.1109/ICRA. 2016. 7487184.

Jonathan, T.; Thang, T.; Balakumar, S.; Yu, X.; Dieter, F.; and Stan, B. 2018. Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects. Paper presented at the 2nd Conference on Robot Learning. Zurich, CH, October 29-31.

Kehl, W.; Manhardt, F.; Tombari, F.; Ilic, S.; and Navab, N. 2017. SSD-6D: Making RGB-Based 3D Detection and 6D Pose Estimation Great Again. In proceedings of the International Conference on Computer Vision.Venice: Institute of Electrical and Electronics Engineers. doi.org/10.1109/ICCV. 2019. 00777.

Li, Y.; Wang, G.; Ji, X.; and Fox, D. 2020. DeepIM: Deep Iterative Matching for 6D Pose Estimation. Int J Comput Vis 128: 657–678. doi.org/10.1007/s11263-019-01250-9.

Li, Z.; Wang, G.; and Ji, X. 2019. CDPN: Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6-DoF Object Pose Estimation. In proceedings of the International Conference on Computer Vision.Seoul: Institute of Electrical and Electronics Engineers. doi.org/10.1109/ICCV. 2019. 00777.

Mitash, C.; Bowen, W.; Kostas, B.; and Abdeslam, B. 2020. Scene-level Pose Estimation for Multiple Instances of Densely Packed Objects. Paper presented at the Conference on Robot Learning. Boston, MA, October 30-November 1.

Pavlakos, G.; Zhou, X.; Chan, A.; Derpanis, K. G.; and Daniilidis, K. 2017. 6-DoF object pose from semantic keypoints. In proceedings of the IEEE International Conference on Robotics and Automation.Singapore: Institute of Electrical and Electronics Engineers. doi.org/10.1109/ICRA. 2017. 7989233.

Peng, S.; Zhou, X.; Liu, Y.; Lin, H.; Huang, Q.; and Bao, H. 2020. PVNet: Pixel-Wise Voting Network for 6DoF Object Pose Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(6): 3212–3223. doi.org/10.1109/TPAMI. 2020. 3047388.

Sundermeyer, M.; Marton, Z.-C.; Durner, M.; Brucker, M.; and Triebel, R. 2018. Implicit 3D Orientation Learning for 6D Object Detection from RGB Images. Paper presented at the European Conference on Computer Vision. Munich, DE, September 8-14.

Tobin, J.; Fong, R.; Ray, A.; Schneider, J.; Zaremba, W.; and Abbeel, P. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In proceedings of the Intelligent Robots and Systems.Venice: Institute of Electrical and Electronics Engineers. doi.org/10.1109/IROS. 2017. 8202133.

Wang, C.; Xu, D.; Zhu, Y.; Martín-Martín, R.; Lu, C.; Fei-Fei, L.; and Savarese, S. 2019. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion. In proceedings of the Computer Vision and Pattern Recognition.California: Computer Vision and Pattern Recognition. doi.org/10.1109/CVPR. 2019. 00346.

Wen, B.; Mitash, C.; Ren, B.; and Bekris, K. E. 2020. se(3)-TrackNet: Data-driven 6D Pose Tracking by Calibrating Image Residuals in Synthetic Domains. In proceedings of the Intelligent Robots and Systems.Nevada: Institute of Electrical and Electronics Engineers. doi.org/10.1109/IROS45743. 2020. 9341314.

Wüthrich, M.; Pastor, P.; Kalakrishnan, M.; Bohg, J.; and Schaal, S. 2013. Probabilistic object tracking using a range camera. In proceedings of the Intelligent Robots and Systems.Tokyo: Institute of Electrical and Electronics Engineers. doi.org/10.1109/IROS. 2013. 6696810.

Xiang, Y.; Schmidt, T.; Narayanan, V.; and Fox, D. 2018. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In Proceedings of the Robotics: Science and Systems XIV.Pennsylvania: Robotics: Science and Systems. doi.org/10.15607/RSS. 2018. XIV. 019.

Zeng, A.; Yu, K. -T.; Song, S.; Suo, D.; Walker, E.; Rodriguez, A.; and Xiao, J. 2017. Multi-view self-supervised deep learning for 6D pose estimation in the Amazon Picking Challenge. In proceedings of the IEEE International Conference on Robotics and Automation.Singapore: Institute of Electrical and Electronics Engineers. doi.org/10.1109/ICRA. 2017. 7989165.