CSC478 Project Proposal

Riddhesh Shah
University of Toronto
riddhesh.shah@mail.utoronto.ca

Abstract

Multi-Object tracking (MOT) is a growing field and important for advancements in popular tech such as autonomous driving systems. Previously popular methods either worked in 2D image space or in 3D point cloud representation spaces. New research indicates that additional performance can be extracted by fusing 2D and 3D methods. We propose changes to the current state-of-the-art fusion MOT method EagerMOT [1] and hope to have comparable results on the KITTIMOT[2] and NuScenes[5] datasets.

1 Introduction

Multi-object tracking (MOT) enables mobile robots to perform well-informed motion planning and navigation by localizing surrounding objects in 3D space and time. Majority of existing research in this field takes place on data from depth sensors (eg. LiDAR) while current state-of-the-art results like those achieved by EagerMOT (Kim et al. [1]) which uses a fusion of LiDAR and camera data. This is done in order to achieve a larger sensing range by leveraging detections in image space while also leveraging LiDAR detections for a higher recall. Using images allows identification of distant incoming objects, while depth estimates allow for precise trajectory localization as soon as objects are within the depth-sensing range. EagerMOT [1] can be divided into the following rough steps:

- Object detection in 2D and 3D spaces followed by the fusion of these detections based on 2D overlap.
- Fused instances then go through a two staged matching process to update their tracks using 3D and/or 2D information.
- An object's track lifecycle is determined by a set of simple rules to determine whether to keep or discard an object's track.

There are many areas of improvement within EagerMOT as it simply outlines a working proof-of-concept template that can be extended for better performance. Areas of research that have been left unexplored are improvements in different stages of this process such as the object detection, object track filtering, sensor fusion methods and data association methods.

2 Related Works

Related works can be separated into 2D MOT, 3D MOT and fusion MOT methods. While a lot of research work has been in 2D MOT and 3D MOT, Fusion MOT remains a relatively underexplored area of research.

- 2D MOT: Early advancements in the field of Multi-Object Tracking was fueled by advancements in deep learning based object detection methods. TrackR-CNN [8] is a popular MOT method that extends Mask R-CNN [9] with 3D convolutional networks to improve temporal consistency of the detector and uses object re-identification as a cue for the association.

Tracktor [10] is another popular 2D MOT method that repurposes the regression head of Faster-RCNN [11] to follow targets. Research in 2D MOT is currently leaning towards end-to-end learning and using graph neural networks for data association.

- 3D MOT: Due to recent advancements in point cloud representation and 3D object detection methods, LiDAR based tracking-by-detection has been gaining popularity. Weng et al.[3] proposes a well performing 3D MOT method that EagerMOT[1] and we will improve upon. Due to strong dependence on 3D detections, 3D MOT methods, included that proposed by Weng et al.[3] is susceptible to false positives and struggles with bridging longer occlusion gaps.
- Fusion MOT: The state-of-the-art method in fusion MOT is EagerMOT. However it is not the first. Previous research in this field includes methods like MOTSFusion[12] that fuses optical flow, scene flow, stereo-depth, and 2D object detections to track objects in 3D space. We will hope to have comparable results to both EagerMOT[1] and MOTSFusion[12] by the end of our project.

3 Proposed Approach

We hope to experiment with changes in the following fields and study their effects on the framework proposed by EagerMOT.

- Replacing the 2D object detection model (Currently RRC[13] for cars and TrackR-CNN[8] for pedestrians) with single stage detectors such as YOLO[7] as well as DETR[6]. We'd like to see if single stage detectors positively impact the detection speed and if DETR improves detection performance.
- Experimenting with more complex fusion methods as opposed to the current overlap method used by EagerMOT. We hope to be able to implement methods proposed by Robert Langaniere[4] such as Late and Mid-level fusions
- EagerMOT uses a Kalman Filter with a linear dynamics model (constant velocity model) for track parameterization. As we know, almost nothing in the real world follows linear dynamics. We hope to achieve better track parameterization by using an Extended Kalman filter to deal with the non-linearity observed in the dataset.

4 References

[1] Kim, Aleksandr, Ošep, Aljoša and Leal-Taix’e, Laura. EagerMOT: 3D Multi-Object Tracking via Sensor Fusion. IEEE International Conference on Robotics and Automation (ICRA), 2021.

[2] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In CVPR, 2012.

[3] X. Weng, J. Wang, D. Held, and K. Kitani. 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. IROS, 2020.

[4] Robert Langaniere, Faculty of Engineering, uOttawa on Synopsis, Youtube. Sensor Fusion for Autonomous Vehicles: Strategies, Methods, and Tradeoffs | Synopsys - YouTube

[5] Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. arXiv preprint arXiv:1903.11027, 2019.

[6] Carion, Massa, Synnaeve, Usunier, Kirillov and Zagoruyko: End-to-End Object Detection with Transformers. ECCV, 2020.

[7] Joseph, Divvala, Girshick and Farhadi: You Only Look Once: Unified, Real-Time Object Detection. CVPR, 2016.

[8] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe. MOTS: Multi-object tracking and segmentation. In CVPR, 2019.

[9] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask R-CNN. In ICCV, 2017.

[10] P. Bergmann, T. Meinhardt, and L. Leal-Taixe. Tracking without bells and whistles. In ICCV, 2019.

[11] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.

[12] J. Luiten, T. Fischer, and B. Leibe. Track to reconstruct and reconstruct to track. IEEE RAL, 5(2):1803–1810, 2020

[13] J. Ren, X. Chen, J. Liu, W. Sun, J. Pang, Q. Yan, Y.-W. Tai, and L. Xu. Accurate single stage detector using recurrent rolling convolution. In CVPR, 2017.