CSC478 Project Report

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. I attempt to implement a robust MOT method [2] on the KITTIMOT test set [1] and try to study the impacts of the second data association step outlined by EagerMOT.

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 [2] 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. [2] 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 life cycle is determined by a set of simple rules to determine whether to keep or discard an object's track.

In my project proposal, I proposed improvements in detections, implementing better filtering and better fusion methods. While I was able to implement a high performing version of EagerMOT by using FusionMOT RRC [3], TrackRCNN [5] for the 2D detections and PointGNN [4] for the 3D detections. I attempted to integrate an extended Kalmann filter while also including imu data provided by KITTI [1], however I ran into some roadblocks that I will describe here.

2 Attempted Methods

2.1 Detection and Tracking

For the KITTI dataset, we first use ego motion estimates in order to convert the coordinates to world coordinates. The conversion files are provided by the authors of EagerMOT [2]. I perform all evaluations on the KITTIMOT[1] testing set as training/validation segmentations are unavailable for trackRCNN [3] and requires large computing resources in order to generate. This however doesn't negatively impact the research as the testing set has significantly more data points than the training set and is used by other comparable methods to set benchmark metrics.

The 2D detections from trackRCNN[3]+RRC[5] and the 3D detections from PointGNN [4]

are then fused by tracing around the intersection of their detections. A 3D detections corresponds to a 2D detection if their IoU is at least (0.1, 0.1) when they are both projected in 2D space.

Once new detections are made, respective tracks are initialized using a Kalman filter(with a constant velocity model). The "track" is simply a prediction of where the detection should appear next. The detections then go through two association stages where each detection is associated to a track. The first association is done in 3D space using greedy matching. The score used to match a detection to its track in 3d space is the one described in EagerMOT[2] called "scaled distance" which is defined as Euclidian distance * cosinedistance(orientations).

A second data association step takes place in 2D by taking unmatched tracks from the first association step and trying to assign detections to them in 2D space. This is done to improve performance of far-off objects as they are more detectable in 2D space than 3D, as well as to account for any occlusions or disturbances in the 3D detections. These associations are done using greedy matching on 2D IoU. The following observation depicts the improvements achieved by using a second stage on the KITTIMOTS [1] testing dataset.

Table I

Matched Tracks after 1st Association	1949
Unmatched Tracks after 1st Association	934
Matched Tracks after 2nd Association	196
Unmatched Tracks after 2nd Association	738

As we can see, the first association step left 1949 tracks unmatched, out of which 196 were recovered by the second association step, clearly showing merit in the two stage approach.

2.2 Filtering (Incomplete)

As proposed in my Project Proposal, I explored methods to improve the filtering done for track initialization and parametrization. The original paper [2] uses a kalman filter with linear dynamics that takes into consideration only the visually observed position and orientation changes. Since KITTI is a dataset with moving vehicles and people, it makes sense to upgrade the filtering method to take into account non linear dynamic modelling. I attempted to do this by integrating an extended Kalman filter that includes information from the IMU of the system collecting data for the KITTIMOTs dataset [1]. The IMU provides latitude+longitude data which is used to provide more accurate x-y world coordinates. We also use acceleration and angular rate from the IMU in our state estimation.

I however ran into a roadblock in the integration step of the extended kalman filter during updating the state vector for the tracking system. The previous scoring function expects a 7D state vector to assign "scaled distance" based on which the first data association is done. The introduction of additional variables from the IMU in the state vector requires the design of a new scoring function that suits our purposes as well as modifications to other structures that need to be adjusted to the new state representation. I ran out of time and so, while the extended kalman filter class using the IMU data is set up, it isn't integrated into my implementation of EagerMOT.

3 Conclusion

I was able to recreate the best results claimed to be achieved by EagerMOT [2] on the KITTI dataset. However the results still struggle with pedestrians and crowded scenes. My inference is that this is due to weak association metrics such as using greedy matching with IoU that causes problems when targets are very close to each other causing "overlapping" observations that break the greedy matching over IoU. The use of linear dynamics in the Kalman filter also seems like a error prone method for predicting future object tracks. An extended Kalman filter or transformer based network to make the track predictions could lead to better results.

The simple yet effective template laid out by EagerMOT leaves room for much improvement and possible further research on real datasets such as KITTI [1]

References

[1] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[2] Ošep Aljoša Kim, Aleksandr and Laura Leal-Taix'e. Eagermot: 3d multi-object tracking via sensor fusion. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[3] Jonathon Luiten, Tobias Fischer, and Bastian Leibe. Track to reconstruct and reconstruct to track. *arXiv:1910.00130*, 2019.

[4] Weijing Shi and Ragunathan (Raj) Rajkumar. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[5] Paul Voigtlaender, Michael Krause, Aljoša Ošep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. MOTS: Multi-object tracking and segmentation. In *CVPR*, 2019.