Graph Neural Networks for 3D Multi-Object Tracking

Xinshuo Weng, Yongxin Wang, Yunze Man, and Kris Kitani

Robotics Institute, Carnegie Mellon University
{xinshuow, yongxinw, yman, kkitani}@cs.cmu.edu

Abstract. 3D Multi-object tracking (MOT) is crucial to autonomous systems. Recent work often uses a tracking-by-detection pipeline, where the feature of each object is extracted independently to compute an affinity matrix. Then, the affinity matrix is passed to the Hungarian algorithm for data association. A key process of this pipeline is to learn discriminative features for different objects in order to reduce confusion during data association. To that end, we propose two innovative techniques: (1) instead of obtaining the features for each object independently, we propose a novel feature interaction mechanism by introducing Graph Neural Networks; (2) instead of obtaining the features from either 2D or 3D space as in prior work, we propose a novel joint feature extractor to learn appearance and motion features from 2D and 3D space. Through experiments on the KITTI dataset, our proposed method achieves state-of-the-art 3D MOT performance. Our project website is at `http://www.xinshuoweng.com/projects/GNN3DMOT`.

Keywords: multi-object tracking, graph neural networks

1 Introduction

Multi-object tracking (MOT) is an indispensable component of applications such as autonomous driving [8] and assistive robots [4]. Recent work approaches MOT in an online manner with a tracking-by-detection pipeline, where a detector [6,9] is applied to all frames and features are extracted *independently* from each object. Then, the pairwise feature similarity is computed between objects and used to solve MOT with a Hungarian algorithm [7]. The key process of this pipeline is to learn discriminative features for objects with different identities.

Our observation is that the feature extraction in prior work is independent for each object as shown in Figure 1 (Top) and there is no interaction. For example, an object's 2D appearance feature is computed only from its own image patch. We found that *independent feature extraction leads to inferior discriminative feature learning*, and object dependency is the key to obtaining discriminative features. Intuitively, the features of the same object over two frames should be as similar as possible and the features between two different objects should be as different as possible to avoid confusion during data association. This can only be achieved if object features can be obtained as a dependent or context-aware process, i.e., modeling interactions between objects.

Based on the observation, we propose a novel *feature interaction mechanism* for MOT as shown in Figure 1 (Bottom). We achieve this by introducing the Graph Neural Networks (GNNs). To the best of our knowledge, our work is the

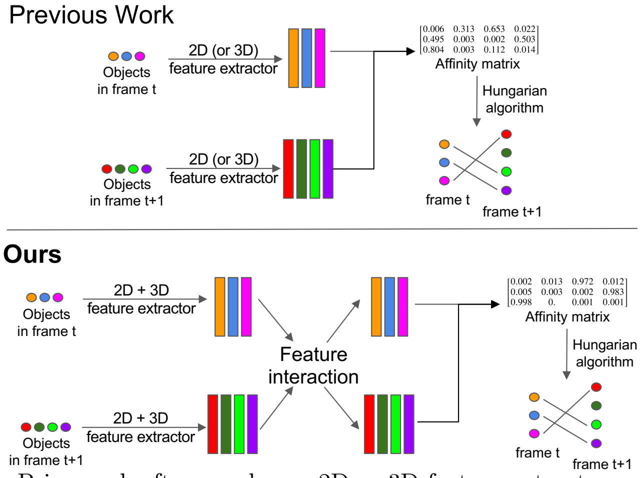


Fig. 1. (**Top**): Prior work often employs a 2D or 3D feature extractor and obtain the feature independently from each object. (**Bottom**): Our work proposes a joint 2D and 3D feature extractor and a feature interaction mechanism to improve the discriminative feature learning for data association in MOT.

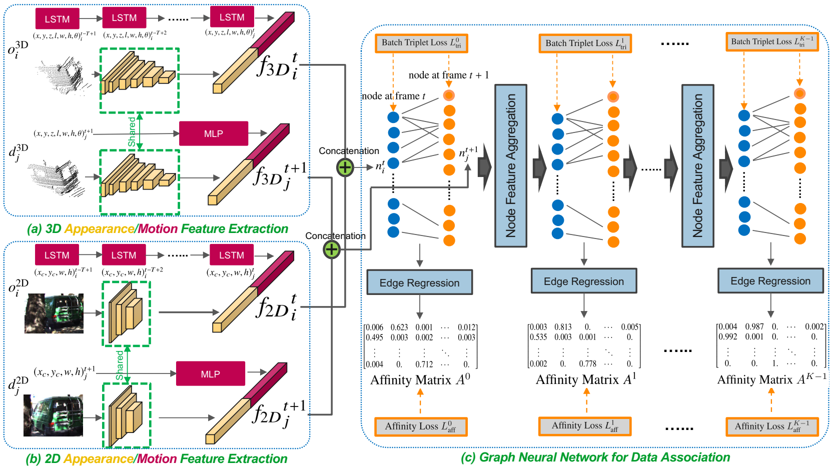
first applying the GNNs to MOT. Specifically, we construct a graph with each node being the object feature. Then, at every layer of the GNNs, each node can update its feature by aggregating features from other nodes. This node feature aggregation process is useful because each object feature is not isolated and can be adapted with respect to other objects. We observe that, after a few GNN layers, the computed affinity matrix becomes more discriminative.

In addition to the feature interaction, another critical aspect to discriminative feature learning in MOT is feature selection. Among different features, motion and appearance are proved to be the most useful features. Although prior works [1] have explored using appearance and motion features, they only focus on 2D or 3D space as shown in Figure 1 (top). That means, prior work only uses the 2D feature when approaching 2D MOT or only uses the 3D feature for 3D MOT. However, this is not optimal as 2D and 3D information are complementary. For example, two objects can be very close in the image but actually at a distance in 3D space because of depth discrepancy. As a result, the 3D information is more discriminative in this case. On the other hand, 3D detection might not be very accurate for faraway objects which will result in noisy 3D motion. In this case, the 2D information might be more accurate for data association.

To this end, we also propose a novel feature extractor that jointly learns motion and appearance features from both 2D and 3D space as shown in Figure 1 (bottom). Specifically, the joint feature extractor has four branches with each branch being responsible for 2D appearance, 2D motion, 3D appearance and 3D motion feature, respectively. Features from all four branches are fused before feeding into the GNNs for feature interaction.

2 Approach

The goal of online MOT is to associate existing tracked objects from previous frame with detected objects in the current frame. Given M tracked objects



(a) 3D Appearance/Motion Feature Extraction

(b) 2D Appearance/Motion Feature Extraction

(c) Graph Neural Network for Data Association

Fig. 2. Proposed Network.

$o_i \in O$ at frame t where $i \in \{1, 2, \cdots, M\}$ and also N detected objects $d_j \in D$ in frame t+1 where $j \in \{1, 2, \cdots, N\}$, we aim to learn discriminative features from O and D and find the matching based on the pairwise similarity.

In Figure 2, our entire network consists of: (a) a 3D appearance and motion feature extractor; (b) a 2D appearance and motion feature extractor. Both 2D and 3D feature extractors are applied to all objects in O and D and then the extracted features are fused together, (c) a graph neural network that takes the fused object feature as inputs and constructs a graph with node being the object feature in the frame t and t+1. Then, the GNNs iteratively aggregates the node feature from the neighborhood and computes the affinity matrix for matching using edge regression. To train the network, we employ the batch triplet loss on the node features and the affinity loss on the predicted affinity matrix.

3 Experiments

Dataset. To evaluate our 3D MOT method that requires both 2D and 3D data as inputs, we use the KITTI [3] dataset, which provides both the 2D (images and 2D boxes) and 3D data (LiDAR point cloud and 3D boxes). Same as most prior works, we primarily report results on the car subset for comparison.

Evaluation Metrics. We use standard CLEAR metrics [2] (MOTA, MOTP, IDS, FRAG) and also the new sAMOTA, AMOTA and AMOTP metrics proposed in [10] for MOT evaluation. Since we are evaluating 3D MOT methods, all above metrics need to be defined in 3D space using the criteria of the 3D IoU or 3D distance. However, the KITTI dataset only supports 2D MOT evaluation, i.e., metrics defined in 2D space for evaluating image-based MOT methods. Therefore, instead of using KITTI 2D MOT evaluation server, we use 3D MOT evaluation code released by [10] for evaluation. Accordingly, evaluation must be done on the validation set as we do not have access to the ground truth of the test set for 3D MOT evaluation. As KITTI does not have an official train / validation split, we use the one proposed by [5].

Table 1. Quantitative comparison of 3D MOT performance on the KITTI dataset.

Method	**sAMOTA**↑	AMOTA↑	AMOTP↑	MOTA↑	MOTP↑	IDS↓	FRAG↓
mmMOT [11]	70.61	33.08	72.45	74.07	78.16	10	125
FANTrack [1]	82.97	40.03	75.01	74.30	75.24	35	202
AB3DMOT[10]	93.28	45.43	77.41	**86.24**	78.43	**0**	15
Ours	**93.92**	**45.83**	**78.10**	86.03	**79.03**	**0**	**10**

Baselines. For 3D MOT, we compare with recent open-source 3D MOT systems such as FANTrack [1], mmMOT [11] and AB3DMOT [10]. To achieve a fair comparison, we use the same 3D detections obtained by PointRCNN [6] for our proposed method and all baselines [1,11,10] that require 3D detections as inputs. For baselines [1,11] that also require 2D detections as inputs, we use the 2D projection of the 3D detections.

Results. We summarize the results in Table 1, where our method consistently outperforms other modern 3D MOT systems in most metrics.

4 Conclusion

To improve discriminative feature learning, we proposed a new 3D MOT method with a novel joint feature extractor and a novel feature interaction mechanism achieved by GNNs. Through experiments on the KITTI dataset, we showed effectiveness of our method and established new S.O.T.A. 3D MOT performance.

References

1. Baser, E., Balasubramanian, V., Bhattacharyya, P., Czarnecki, K.: FANTrack: 3D Multi-Object Tracking with Feature Association Network. IV (2020)
2. Bernardin, K., Stiefelhagen, R.: Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. Journal on Image and Video Processing (2008)
3. Geiger, A., Lenz, P., Urtasun, R.: Are We Ready for Autonomous Driving? the KITTI Vision Benchmark Suite. CVPR (2012)
4. Manglik, A., Weng, X., Ohn-bar, E., Kitani, K.M.: Forecasting Time-to-Collision from Monocular Video: Feasibility, Dataset, and Challenges. IROS (2019)
5. Scheidegger, S., Benjaminsson, J., Rosenberg, E., Krishnan, A., Granstr, K.: Mono-Camera 3D Multi-Object Tracking Using Deep Learning Detections and PMBM Filtering. IV (2018)
6. Shi, S., Wang, X., Li, H.: PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud. CVPR (2019)
7. W Kuhn, H.: The Hungarian Method for the Assignment Problem. Naval Research Logistics Quarterly (1955)
8. Wang, S., Jia, D., Weng, X.: Deep Reinforcement Learning for Autonomous Driving. arXiv:1811.11329 (2018)
9. Weng, X., Kitani, K.: Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud. ICCVW (2019)
10. Weng, X., Wang, J., Held, D., Kitani, K.: 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. IROS (2020)
11. Zhang, W., Zhou, H., Sun, S., Wang, Z., Shi, J., Loy, C.C.: Robust Multi-Modality Multi-Object Tracking. ICCV (2019)