Robust 3D Object Tracking in Autonomous Vehicles

Eric Chan (erchan@stanford.edu), Anthony Galczak (agalczak@stanford.edu), Anthony Li (antli@stanford.edu)

Department of Computer Science, Stanford University

Abstract

We present a stereo-camera based 3D multiple-vehicle-tracking system that utilizes Kalman filtering to improve robustness. The objective of our system is to accurately predict locations and orientations of vehicles from stereo camera data. It consists of three modules: a 2D object detection network, 3D position extraction, and 3D object correlation / smoothing. The system approaches the 3D localization performance of LIDAR and significantly outperforms the state-of-the-art monocular vehicle tracking systems.

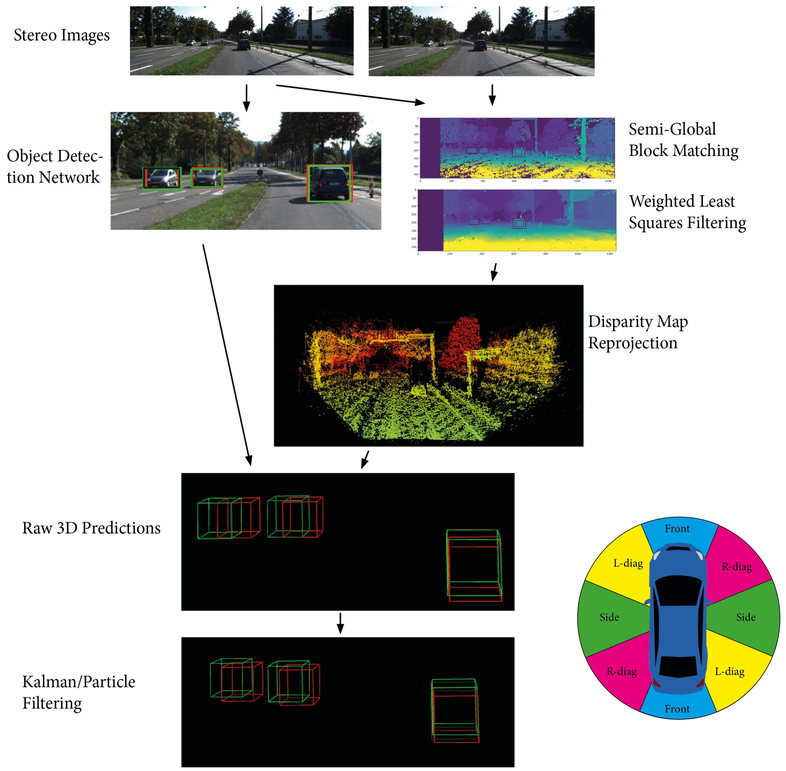
Data

2D object detector training data set

- KITTI Object Detection 2012[1]
  - 2D bounding boxes, observation angle.
  - 7481/7518 train/test split[2]

Tracking evaluation data set

- KITTI Object Tracking 2012[1]
  - Sequential stereo camera images
  - 21 labeled sequences, 200+ frames ea.

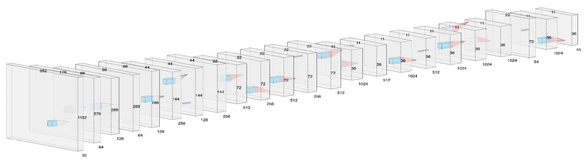


Features

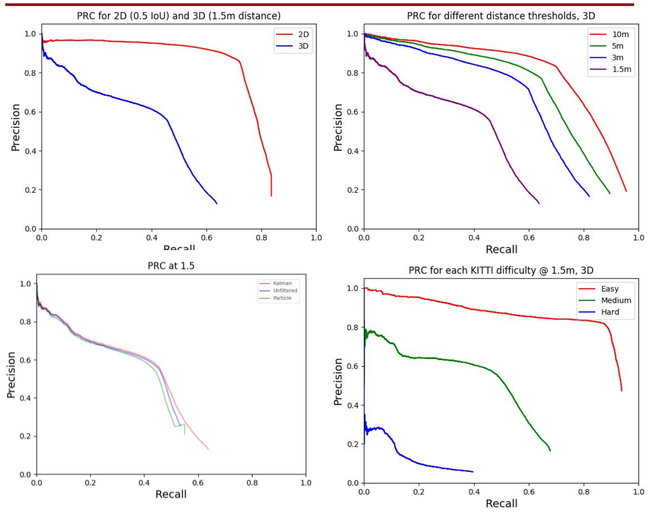
The features for our model are the left image of the object tracking data set. From here, we calculate a 2D bounding box from our YOLO model and combine it with the corresponding right image to produce a depth prediction. Finally, we output a 3D position of the object which we reconstruct into a 3D bounding box.

Models

We use transfer learning on top of YOLOv2[5] to extract image-space bounding boxes and observation angles from our imagery. This is necessary as YOLOv2 performance for detecting vehicles on the KITTI data set has been shown to be very poor out of the box[6].



Results



	Unfiltered	**Kalman**	Particle	Mono[3]	LIDAR[4]
Easy	0.799	0.8079	0.7251	.1805	0.8661
Med	0.3737	0.4027	0.3671	.1498	0.7763
Hard	0.1198	0.1269	0.1192	.1342	0.7606

Discussion

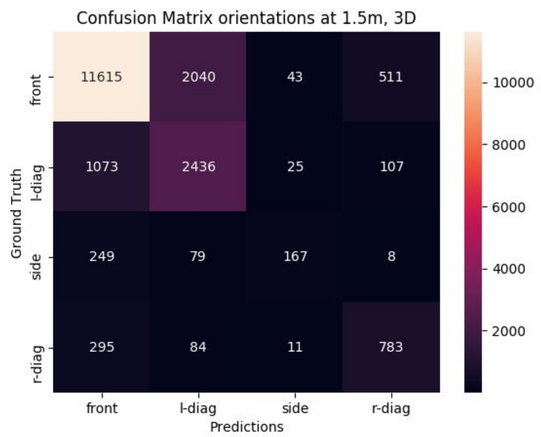
Although our system lacks the fine-grained precision of LIDAR, it can still adequately track most vehicles.

Our system achieves comparable precision to LIDAR for vehicles that are “Easy” to detect, but performs significantly worse for more difficult vehicles.

Our system significantly outperforms state-of-the-art monocular detectors, achieving a 450% MAP improvement for “Easy” detections and 270% improvement for “Moderate” difficulty vehicles.

Our 2D object detection performance is significantly better than our 3D performance, implying our results could be better with improved ranging accuracy.

Kalman filtering outperformed both our unfiltand particle filtering for 3D localization performance.



Future

We have identified two areas that would result in significant performance improvements.

To provide vehicle *tracking*, our system correlates each 3D detection with the most likely tracked vehicle. Our current algorithm uses a distance-based heuristic, which is sometimes confused by closely clustered vehicles. A possible extension is an appearance-based matching algorithm based on image embeddings.

Our ranging system performs poorly on vehicles that are significantly occluded because it ranges the occluding object rather than the occluded vehicle. We hypothesize that a ranging algorithm that includes semantic segmentation could produce significant performance improvements for occluded vehicles.

References

[1] Andreas Geiger, Philip Lenz, Christoph Stiller, Raquel Urtasun. Vision Meets Robotics: The KITTI Dataset. International Journal of Robotics Research. 2013.

[2] Andreas Geiger, Philip Lenz, Raquel Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. Conference on Computer Vision and Pattern Recognition (CVPR). 2012.

[3] Simonelli, Andrea et al. “Disentangling Monocular 3D Object Detection.” ArXiv abs/1905.12365 (2019).

[4] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, Jiaya Jia; STD: Sparse-to-Dense 3D Object Detector for Point Cloud. The IEEE International Conference on Computer Vision (ICCV), 2019, pp. 1951-1960

[5] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 6517-6525. doi: 10.1109/CVPR.2017.690

[6] Asvadi, Alireza. “YOLOv2 416x416 Detection Framework Experiment on KITTI.” The KITTI Vision Benchmark Suite,

[7] W. Choi, C. Pantofaru and S. Savarese, "A General Framework for Tracking Multiple People from a Moving Camera," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 7, pp. 1577-1591, July 2013