Joint SFM and Detection Cues for Monocular 3D Localization in Road Scenes

Shiyu Song, Manmohan Chandraker
NEC Labs America, Cupertino, CA

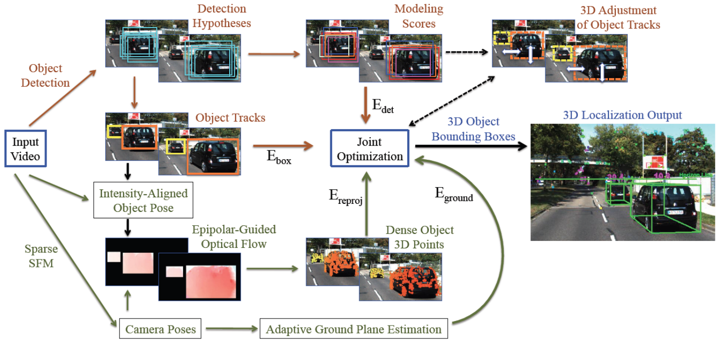

Figure 1: Overview of our system for 3D object localization by combining SFM cues (green) with object detection cues (brown). Given monocular video input, camera poses and ground plane are estimated by SFM, while a dense tracking framework yields 3D points on objects. These are combined with cues from object detection hypotheses and object tracks in a joint optimization framework that allows for soft adjustment of track positions to maximize consistency with 3D cues, bounding boxes and detection scores.

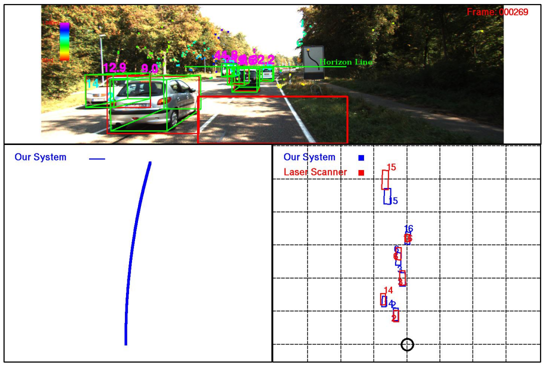

Figure 2: Output of our localization system. The bottom left panel shows the monocular SFM camera trajectory. The top panel shows input 2D bounding boxes in red, horizon from estimated ground plane and the estimated 3D bounding boxes in green with distances in magenta. The bottom right panel shows the top view of the ground truth object localization from laser scanner in red, compared to our 3D object localization in blue.

We present a framework for fast and highly accurate 3D localization of objects such as cars in autonomous driving applications, using a single camera. Our localization framework jointly uses information from complementary modalities such as structure from motion (SFM) and object detection to achieve high localization accuracy in both near and far fields. This is in contrast to prior works that rely purely on detector outputs, or motion segmentation based on sparse feature tracks. Rather than completely commit to tracklets generated by a 2D tracker, we make novel use of raw detection scores to allow our 3D bounding boxes to adapt to better quality 3D cues. To extract SFM cues, we demonstrate the advantages of dense tracking over sparse mechanisms in autonomous driving scenarios. In contrast to complex scene understanding, our formulation for 3D localization is efficient and can be regarded as an extension of sparse bundle adjustment to incorporate object detection cues. Figure 1 illustrates an overview of the system.

Given monocular video input, we estimate the pose and ground plane corresponding to the camera using the system of [3]. Intuitively, SFM can estimate accurate 3D points on nearby objects, but suffers due to the low resolution of those far away. On the other hand, bounding boxes from object detection are obtainable for distant objects, but are often inconsistent with the 3D scene in the near field. Thus, we seek 3D bounding boxes that are most consistent with 2D tracked ones, while also maximizing the alignment of estimated object pose with tracked 3D points. We define a combined energy function to be minimized over the set of object poses $\{\mathbf{\Omega}^i(t)\}$, 3D bounding box dimensions $\mathbf{B}^i$ and the set of tracked 3D points on each object $\mathbf{X}_o^i$, for objects $i = 1, \cdots, N$, as:

$$\mathcal{E}\Big(\{\mathbf{\Omega}^i(t)\}, \{\mathbf{B}^i\}, \{\mathbf{X}_o^i\}\Big) = \mathcal{E}_{sfm} + \lambda_o \mathcal{E}_{obj} + \lambda_p \mathcal{E}_{prior}, \qquad (1)$$

where $\mathcal{E}_{sfm}$, $\mathcal{E}_{obj}$ and $\mathcal{E}_{prior}$ are the SFM cost, object cost and the prior cost, respectively. We formulate the objective function in (1) as an extension of traditional bundle adjustment to incorporate object cues, since it is defined over a set of variables $\{\mathbf{\Omega}^i(t)\}$ that constitutes "poses" and another set given by $\{\mathbf{B}^i, \mathbf{X}_o^i\}$ that constitutes "3D points".

To obtain 3D points on objects, the PnP-based pose computation of background SFM does not suffice since it requires prior knowledge of feature tracks, which are not plentiful on objects like cars. Instead, given the object pose at time t and an existing set of 3D points, the pose at time $t+1$ is computed by minimizing intensity differences. Using object pose from intensity alignment rather than feature tracks allows epipolar guidance for a TV-L1 optical flow process that generates dense feature tracks. The flow computation reduces to a 1D search and is performed only within the object bounding boxes. Note that the intensity-aligned pose and triangulated 3D points from dense tracking are refined by the joint optimization framework that also incorporates other cues such as objects and ground plane. The total SFM cost $\mathcal{E}_{sfm}$ favors object poses that minimize reprojection error for object 3D points and best align the object with the SFM ground plane.

The total object cost $\mathcal{E}_{obj}$ is a weighted sum of the bounding box and detection costs. The bounding box cost seeks the object poses and dimensions whose projection through the estimated camera poses are most consistent with the detected 2D bounding boxes. Further, we model detection scores as a sum of Gaussians, that can be evaluated during continuous optimization without evaluating the detector model. This allows the detection cost to undo any tracking errors and seek object poses that correspond to high detection scores as well as good alignment with other 3D cues. Finally, the energy $\mathcal{E}_{prior}$ encourages object trajectories to be smooth and object sizes to be close to the category mean.

We evaluate our system on the KITTI dataset [2]. Our ablation studies demonstrate the effectiveness of various components such as ground plane estimation, object bounding boxes, detection score modeling and 3D points from epipolar-guided optical flow using intensity-aligned poses. The position accuracy in depth is 8.3% for near objects and 10.4% for far objects, compared to 13.9% and 26.9%, repsectively, for a baseline method. Our approach can also improve the accuracy of existing scene understanding frameworks, demonstrated by a 7% improvement in position accuracy over the results of [1]. A sample output from our system is shown in Figure 2.

To summarize, we propose a novel framework for 3D object localization, designed for autonomous driving applications. It recognizes and exploits the complementary strengths of SFM cues (3D points and ground plane) and object cues (bounding boxes and detection scores), to achieve good 3D localization accuracy in both near and far fields.

[1] W. Choi and S. Savarese. Multi-target tracking in world coordinate with single, minimally calibrated camera. In *ECCV*, 2010.

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

[3] S. Song and M. Chandraker. Robust scale estimation in real-time monocular SFM for autonomous driving. In *CVPR*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.