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

Shiyu Song and Manmohan Chandraker

NEC Laboratories America
Relentless passion for innovation

CVPR 2015 BOSTON JUNE 7-12

Intuition

3D reconstruction and object detection are complementary.

Object SFM
- ✗ Far objects "invisible"
- ✓ Near objects accurate
- ✗ Moving objects
- ✓ Moving camera

Object Detection
- ✓ Far objects accurate
- ✗ Near objects "distorted"
- ✓ Handles moving objects
- ✗ Camera motion cue

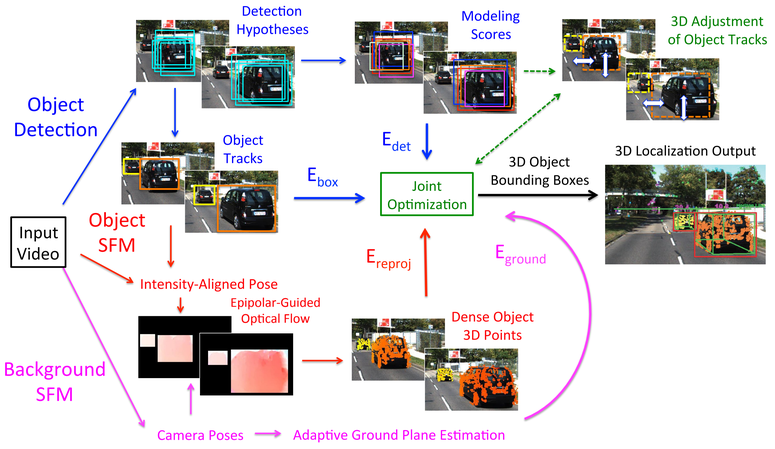
Localization: Joint optimization

SFM Camera Poses, Ground Plane → Joint optimization ← 2D Object Position, Object Identities

→ 3D object position and orientation

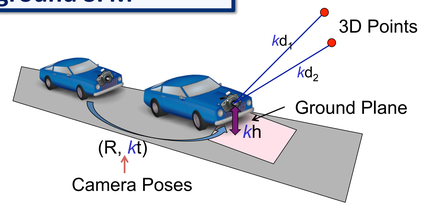
- ✓ Accurate for both near and far objects
- ✓ Uniformly handles camera and object motions

Overall Framework



Background SFM

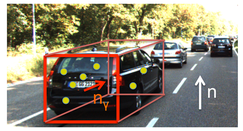
- Estimate camera poses and the ground plane.
- Ground plane provides the absolute scale of the translation.
- A novel data-driven framework that estimates the ground plane.
- The framework combines multiple cues based on per-frame observation covariance.



S. Song and M. Chandraker. Robust scale estimation in real-time monocular SFM for autonomous driving. In CVPR 2014, pages 1566–1573, 2014.

Initialization and Optimization

Initialization



- Heading angle: nonholonomic motion assumption.
- 3D bounding box: Fit to the 2D tracks assuming it is on the ground.
- 3D points: lie on the plane n_γ.

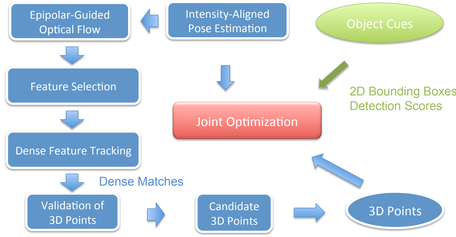
Optimization

$$\varepsilon_{sfm} + \lambda_o \varepsilon_{obj} + \lambda_p \varepsilon_{prior}$$

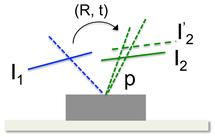
- Minimize over all object poses in a window.
- Can be posed as extension of bundle adjustment that includes object cues
- Use Levenberg-Marquardt like traditional bundle adjustment.

Object SFM

Overall Pipeline



Intensity-based Pose Alignment

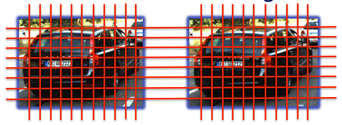


- Photo-consistency assumption.
- Pose estimation without feature matching.
- Avoid PnP failure when feature tracks are not plentiful.

Epipolar Guided Optical Flow

- Optical flow within sub-image defined by object bounding box.
- Optical flow with epipolar constraints.
- Faster speed and better accuracy.

Dense Feature Tracking

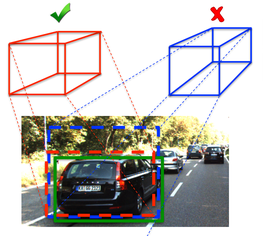


- 8x8 bucketing with highest Harris Score.
- Quality control of the tracks.

- C. Kerl et al. ICRA 2013; N. Slesareva et al. Pattern Recognition 2005; C. Zach et al. DAGM 2007; N. Sundaram et al. ECCV 2010.

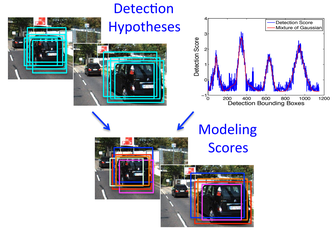
Object Cues and Priors

Bounding Box Fitting



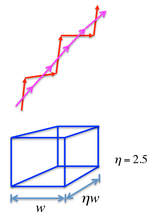
Use the edges of the 2D bounding box as constraints.

Detection Scores



Detection score modeling allows 3D cues to adjust suboptimal 2D tracks.

Priors

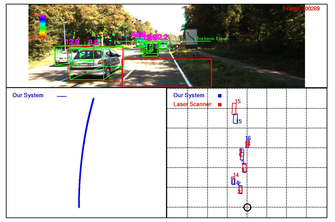


Trajectory smoothness and length ratio prior.

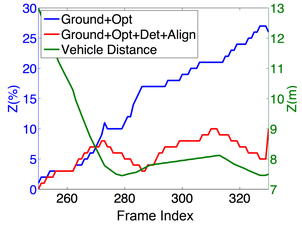
Results

Method	Ground truth tracks						Tracked bounding boxes*					
	Near Object			Far Object			Near Object			Far Object		
	Z(%)	X(m)	Size(%)	Z(%)	X(m)	Size(%)	Z(%)	X(m)	Size(%)	Z(%)	X(m)	Size(%)
CalibGround	10.2	0.53	14.8	25.3	0.79	12.3	13.9	0.58	16.1	26.9	0.75	12.0
Adaptive Ground	9.0	0.38	14.8	9.8	0.35	12.3	13.3	0.50	16.1	10.2	0.33	12.0
Ground+Opt	6.4	0.26	9.3	8.9	0.35	13.3	9.5	0.33	13.5	9.4	0.34	13.6
Ground+Opt +Det	6.1	0.25	9.1	8.6	0.33	12.1	9.4	0.32	12.4	9.5	0.33	12.5
Ground+Opt +Det+PnP	5.9	0.24	8.1	8.5	0.34	11.8	9.4	0.30	10.9	11.2	0.37	14.2
Ground+Opt +Det+Align	5.5	0.24	7.3	8.3	0.33	12.0	8.3	0.28	8.0	10.4	0.36	13.9

Visualization



Relative Benefits of Cues



*A. Geiger et al. PAMI 2014; A. Geiger et al. CVPR, 2012.