VL-SLAM: Real-Time Visual-Inertial Navigation and Semantic Mapping

Jingming Dong, Xiaohan Fei, Nikolaos Karianakis, Konstantine Tsotsos, Stefano Soatto

UCLAVISIONLAB

Motivation

- Objects exist in the scene, not in images: Images provide *evidence* in support of object hypotheses in 3D space.
- An object's geometric, photometric and semantic attributes persist across multiple observations.
- Deep convolutional neural networks (CNNs) do not enforce continuity in detections across images
- Even when occluded, once seen we remain aware of objects' presence in the scene and can *predict* their re-appearance.
- Objects have characteristic size (scale) and shape.
- Gravity, through inertial sensors, provides a persistent orientation reference for *objects.*

Key Ideas

- Exploit inertial reference and structure from motion for explicit reasoning of objects in the scene. e.g., can enforce size and shape priors, canonize rotations in images using gravity alignment.
- There are no objects in images, just pixels. CNNs do not detect objects in the scene.
- Interpret output of a CNN as a likelihood function to score hypotheses of objects in the scene.
- Perform causal, real-time detection and localization along with state-of-the-art visual inertial fusion and mapping.

Acknowledgments

- Supported by AFOSR, AFRL, ARO, ONR.

Object Representation

- Object attributes z (shape, pose, ID) given images up to the current time x^t:

$$p(z|x^t)$$

- Condition on an attributed point cloud s, a minimal sufficient statistic for localization [Tsotsos et al., 2015], and sensor pose g_t, given images x^t and inertials u^t up to the current time:

$$p(g_t, s|x^t, u^t)$$

- Marginalize object representation over viewpoint estimate from SLAM:

$$p(z|x^t) = \int p(z|g_t, s, x^t) dP(g_t, s|x^t, u^t)$$

SLAM

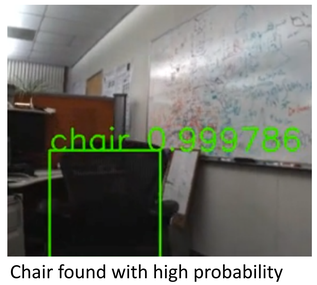
- Causal update of object hypotheses:

$$p(z|g_{t+1}, s, x^{t+1}) \propto p(x_{t+1}|z, \hat{g}_t, u^t, s) p(z|g_t, s, x^t)$$

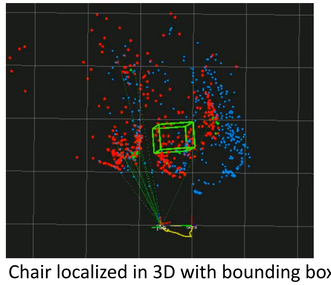
CNN

Bayesian Filter: (weighted mixture of EKFs)

- Captures joint distribution of object shape (including scale) and identities, and geometric relations in the scene.
- Represent objects with 3D bounding boxes:



Chair found with high probability

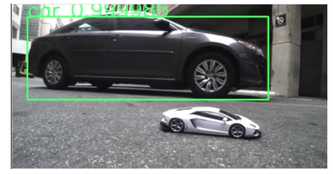


Chair localized in 3D with bounding box

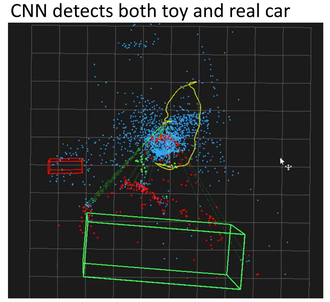
Takeaway Message

Toy Car or Real Car?

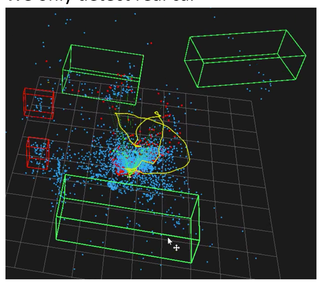
CNN detects both toy and real car



We only detect real car



Real car localized in 3D



Multiple cars and people in 3D

- Causal, real-time object detection and localization in the scene.
- Employs state-of-the-art real-time visual-inertial fusion/geometric mapping [Tsotsos et al., 2015] and off-the shelf CNN (YOLO).
- Captures identities and geometric relations.
- Handles scale and occlusion.
- Future Work: Dynamic objects, topology through dense reconstruction.

References

- S. Soatto and A. Chiuso, "Visual Scene Representations: Defining Properties and Deep Approximations", Proc. Of ICLR 2016 (ArXiv 1411.7676)
- K. Tsotsos, A. Chiuso, S. Soatto, "Robust Filtering for Visual-Inertial Sensor Fusion", Proc. Of ICRA 2015.
- US Pat. Appl. 14932899
- Redmon et al., "You only look once: Unified real-time object detection", ArXiv 1506.02640.