Localization and Mapping using Instance-specific Mesh Models

UC San Diego
JACOBS SCHOOL OF ENGINEERING
Electrical and Computer Engineering

Qiaojun Feng, Yue Meng, Mo Shan, Nikolay Atanasov

Department of Electrical & Computer Engineering, Contextual Robotics Institute, University of California San Diego

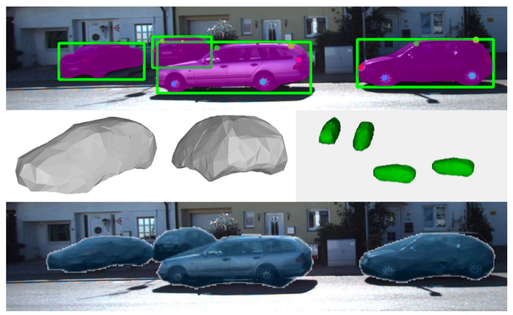
CONTEXTUAL ROBOTICS INSTITUTE

ERL

ABSTRACT

This paper focuses on building semantic maps, containing object **poses** and **shapes**, using a monocular camera. This is an important problem because robots need rich understanding of geometry and context for higher level tasks. Our contribution is an instance-specific mesh model of object shape that can be optimized online based on semantic information extracted from camera images. Multi-view constraints on the object shape are obtained by detecting objects and extracting category-specific keypoints and segmentation masks. We show that the errors between projections of the mesh model and the observed keypoints and masks can be differentiated in order to obtain accurate instance-specific object shapes.

PROBLEM FORMULATION



States to estimate:

- Camera poses: $\mathcal{C} \triangleq \{c_t\}_{t=1}^{T}, c_t \in SE(3)$
- Object shapes(meshes vertices & faces) and poses
  $\mathcal{O} \triangleq \{o_n\}_{n=1}^{N}, o_n = (\mu_n, T_{o_n})$
  $\mu_n = (V_n, F_n), T_{o_n} \in SE(3)$

Given observations:

- A sequence of images from the camera pose:
  $\mathrm{I} \triangleq \{i_t\}_{t=1}^{T}$
- Object categories, instance masks and semantic keypoints from an image:
  $\mathcal{Z}_t \triangleq \{z_{lt} = (\xi_{lt}, s_{lt}, y_{lt})\}_{l=1}^{L_t}$
  - Category $\xi_{lt} \in \Xi$
  - Instance mask(binary) $s_{lt} \in \{0,1\}^{W \times H}$
  - Semantic keypoints(2D) $y_{lt} \in \mathbb{R}^{2 \times |K_{lt}|}$
- Data association $n = \pi_t(l)$

METHODOLOGY

1. Form the problem as a loss-minimization task

- Observation model:
  - Instance mask $\hat{s}_{lt} = \mathcal{R}_{\text{mask}}(\hat{c}_t, \hat{o}_n)$
  - Semantic keypoints $\hat{y}_{lt} = \mathcal{R}_{\text{kps}}(\hat{c}_t, \hat{o}_n, A_n)$
    keypoints association matrix A_n

- Loss function:
  - Instance mask: negative Intersection over Union

$$\mathcal{L}_{\text{mask}}(s, \hat{s}) = -\frac{\|s \odot \hat{s}\|_1}{\|s + \hat{s} - s \odot \hat{s}\|_1}$$

  - Semantic keypoints: square Euclidean distance
    $\mathcal{L}_{\text{kps}}(y, \hat{y}) = \|y - \hat{y} \cdot vis(\hat{y})\|_F^2$
    $vis(\hat{y}_{lt})$ discards unobservable object keypoints

- Find camera poses and object states to minimize the loss

$$\arg\min_{\mathcal{C},\mathcal{O}} \sum_{t=1}^{T} \sum_{l=1}^{L_t} w_{\text{mask}} \mathcal{L}_{\text{mask}}(s_{lt}, \mathcal{R}_{\text{mask}}(c_t, o_{\pi_t(l)})) + w_{\text{kps}} \mathcal{L}_{\text{kps}}\left(y_{lt}, \mathcal{R}_{\text{kps}}(c_t, o_{\pi_t(l)}, A_{\pi_t(l)})\right)$$

2. Make the observation model differentiable

- Semantic keypoints: standard perspective projection

$$\mathcal{R}_{\text{kps}}(c, o, A) := K \pi R_c^T (R_o V A + (p_o - p_c)\mathbf{1}^T)$$

- Instance mask:
  A rasterization function projects the mesh faces F to the image frame to generate mask.
  Though the origin function is not differentiable, Kato et al. [1] show a method to obtain an approximate gradient for the rasterization function.

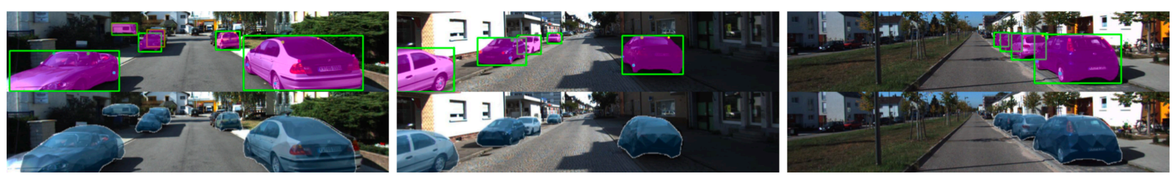
$$\mathcal{R}_{\text{mask}}(c, o) := Raster\left(\mathcal{R}_{\text{kps}}(c, o, I), F\right)$$

3. Gradient-based optimization and its initialization

- Since the gradient w.r.t. the unknown variables can be calculated explicitly, we can use gradient-based method to minimize the loss to find the desired states
- For the initialization of the optimization(mapping task), we estimate the 3D coordinates of the keypoints using Levenberg-Marquardt algorithm and the pose using Kabsch algorithm.

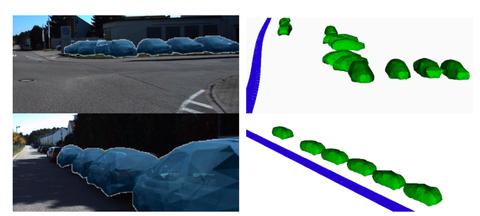
RESULTS

Experiments on KITTI dataset for mapping task.



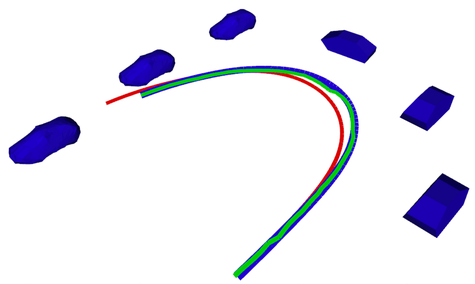
Top: the semantic observations. Bottom: the projection of reconstructed mesh models.

Top: category-level model before shape optimization.
Bottom: instance-level model after shape optimization.

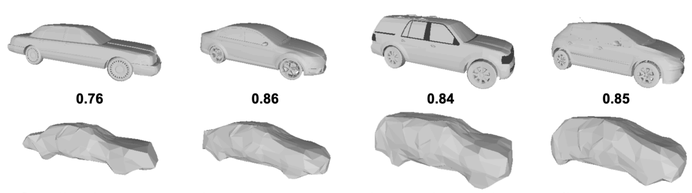


Left: 2D observation of mesh models.
Right: corresponding 3D configuration.
Trajectory in blue.

Experiments for localization task and mapping task on simulation dataset.



Localization results from a simulated dataset, showing car poses (blue), the ground truth camera trajectory (blue), the inertial odometry used for initialization (red), and the optimized camera trajectory (green).



Qualitative comparison between estimated car shapes (bottom row) obtained from a simulation sequence and ground truth object meshes (top row). The numbers indicate 3D IoU. 10 views are used for each object.

CONCLUSION

This work demonstrates that object categories, shapes and poses can be recovered from visual semantic observations. The key innovation is the development of differentiable keypoints and segmentation mask observation models that allow object shape to be used for simultaneous semantic mapping and camera pose optimization. Our ultimate goal is to develop an online SLAM algorithm that unifies semantic, geometric, and inertial measurements to allow rich environment understanding and contextual reasoning.

REFERENCES

[1]. H. Kato, Y. Ushiku, and T. Harada, “Neural 3d mesh renderer,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3907–3916.

ACKNOWLEDGEMENTS

We gratefully acknowledge support from ARL DCIST CRA W911NF-17-2-0181.

Email: qjfeng@ucsd.edu, natanasov@ucsd.edu