Localization and Mapping using Instance-specific Mesh Models*

Qiaojun Feng Yue Meng Mo Shan Nikolay Atanasov

***Abstract*— This paper focuses on building semantic maps, containing object poses and shapes, using a monocular camera. 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.**

I. Introduction

The foundations of artificial perception lie in the twin technologies of inferring geometry (e.g., occupancy mapping) and semantic content (e.g., scene and object recognition). Simultaneous Localization And Mapping (SLAM) are approaches capable of tracking the pose of a robotic system while simultaneously reconstructing a sparse or dense geometric representation of the environment. A major research challenge today is to exploit information provided by deep learning, such as category-specific object keypoints , semantic edges, and segmentation masks , in VIO and SLAM algorithms to build rich models of the shape, structure, and function of objects.

This paper addresses camera localization and object-level mapping, incorporating object categories, poses, and shapes. Our main **contribution** is the development of an instance-specific object shape model based on a triangular mesh and differentiable functions that measure the discrepancy in the image plane between projections of the model and detected semantic information. We utilize semantic keypoints [1], [2] and segmentation masks [3] as observations for optimizing the error functions. Initialized from a pre-defined mean category-level model, the optimization steps are inspired by the recently proposed differentiable mesh renderer [4], which allows back-propagation of mask errors measured on a rendered image to update the mesh vertices.

II. Problem Formulation

We consider the problem of detecting, localizing, and estimating the shape of object instances present in the scene, and estimating the pose of a camera over time. The states we are interested in estimating are the camera poses $\mathcal{C} \triangleq \{c_t\}_{t=1}^{T}$ with $c_t \in SE(3)$ and the object shapes and poses $\mathcal{O} \triangleq \{o_n\}_{n=1}^{N}$. An object state $o_n = (\mu_n, R_{o_n}, p_{o_n})$ consists of a pose $R_{o_n} \in SO(3)$, $p_{o_n} \in \mathbb{R}^3$ and shape μ_n, specified as a 3-D triangular mesh $\mu_n = (V_n, F_n)$ in the object canoncial frame with vertices $V_n \in \mathbb{R}^{3\times|V_n|}$ and faces $F_n \in \mathbb{R}^{3\times|F_n|}$. We define a keypoint association matrix $A_n \in \mathbb{R}^{|V_n|\times|K_n|}$ that generates $|K_n|$ keypoints $V_n A_n$ from all mesh vertices.

*We gratefully acknowledge support from ARL DCIST CRA W911NF-17-2-0181. The authors are with Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093, USA {qjfeng,yum107,moshan,natanasov}@ucsd.edu

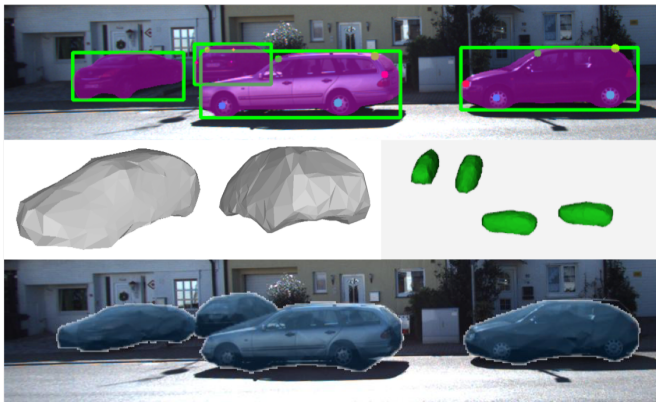

Fig. 1: Our objective is to build detailed environment maps incorporating object poses and shapes. The figure from KITTI [5] in the top row shows the kind of information that our method relies on: bounding boxes (green), segmentation masks (magenta) and semantic keypoints (multiple colors). The middle row includes the reconstructed mesh models and 3D configuration. The last row shows the projection result.

Suppose that a sequence $\mathcal{I} \triangleq \{i_t\}_{t=1}^{T}$ of T images $i_t \in \mathbb{R}^{W\times H}$, collected from the corresponding camera poses $\{c_t\}_{t=1}^{T}$, are available for the estimation task. From each image i_t, we extract a set of object observations $\mathcal{Z}_t \triangleq \{z_{lt} = (\xi_{lt}, s_{lt}, y_{lt})\}_{l=1}^{L_t}$, consisting of a detected object's category $\xi_{lt} \in \Xi$, a segmentation masks $s_{lt} \in \{0,1\}^{W\times H}$ and the pixel coordiantes of detected keypoints $y_{lt} \in \mathbb{R}^{2\times|K_{lt}|}$. See Fig. 1 for example object observations.

We can predict expected semantic mask $\hat{s}_{lt}$ and semantic keypoint observations $\hat{y}_{lt}$ using a perspective projection model: $\hat{s}_{lt} = \mathcal{R}_{\text{mask}}(\hat{c}_t, \hat{o}_n), \hat{y}_{lt} = \mathcal{R}_{\text{kps}}(\hat{c}_t, \hat{o}_n, A_n)$

The camera and object estimates can be optimized by reducing the error between the predicted $\hat{\mathcal{Z}}_{1:T}$ and the actual $\mathcal{Z}_{1:T}$ observations measured by loss functions $\mathcal{L}_{\text{mask}}$ and $\mathcal{L}_{\text{kps}}$

Problem. Given object observations $\mathcal{Z}_{1:T}$, determine the camera poses $\mathcal{C}$ and object states $\mathcal{O}$ that minimize the mask and keypoint losses:

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

where w_{mask}, w_{kps} are scalar weight parameters specifying the relative importance of the mask and keypoint loss functions.

III. Technical Approach

For each frame, we first use [3] to get object detection results represented with bounding boxes and instance segmentations inside the boxes. Each object is assigned to one of the class labels in Ξ. Then we extract semantic keypoints y_{lt} within the bounding box of each detected object using the stacked hourglass model of [2].

Next, we develop the observation models $\mathcal{R}_{\text{mask}}$ and $\mathcal{R}_{\text{kps}}$ that specify how a semantic observations $z = (\xi, s, y)$ is generated by a camera with pose $(R_c, p_c) \in SE(3)$ observing an object of class $\xi \in \Xi$ with pose $(R_o, p_o) \in SE(3)$ and mesh shape $\mu = (V, F)$ with keypoint association matrix A. Let K be the intrinsic matrix of the camera. Let $x := VAe_k \in \mathbb{R}^3$ be the coordinates of the k-th object keypoint in the object frame, where e_k is a standard basis vector. The projection of x onto the image frame can be determined by first projecting it from the object frame to the camera frame using (R_o, p_o) and (R_c, p_c) and then the perspective projection $\pi(\cdot)$. In detail, this sequence of transformations leads to the pixel coordinates of x as follows:

$$y^{(k)} = K\pi(R_c^T(R_o x + p_o - p_c)) \in \mathbb{R}^2 \tag{2}$$

Applying the same transformation to all object keypoints VA simultaneously leads to the keypoint projection model:

$$\mathcal{R}_{\text{kps}}(c, o, A) := K\pi\left(R_c^T(R_o VA + (p_o - p_c)\mathbf{1}^T)\right) \tag{3}$$

where $\mathbf{1}$ is a vector whose elements are all equal to 1.

To define $\mathcal{R}_{\text{mask}}$, we need an extra rasterization step, which projects the object faces F to the image frame. Kato et al. [4] show how to obtain an approximate gradient for the rasterization function $Raster(\cdot)$, which is used here. We can define the mask projection model:

$$\mathcal{R}_{\text{mask}}(c, o) := Raster\left(\mathcal{R}_{\text{kps}}(c, o, I), F\right) \tag{4}$$

Since all the steps are differentiable, we can use gradient-based method to solve the optimization in (1).

We implemented the localization and mapping tasks separately. In the localization task, we initialize the camera pose using inertial odometry obtained from integration of IMU measurements [6]. The camera pose is optimized sequentially between every two images via (1), leading to an object-level visual-inertial odometry algorithm.

To initialize the object model in the mapping task, we collect high-quality keypoints (according to q_{lt} defined in Sec. III) from multiple frames until an object track is lost. The 3-D positions of these keypoints are estimated by optimizing $\mathcal{L}_{\text{kps}}$ only using the Levenberg-Marquardt algorithm. Using a predefined category-level mesh model(mean model) with known keypoints, we apply the Kabsch algorithm to initialize the object pose (i.e., the transformation from the detected 3-D keypoints to the category-level model keypoints). To improve the deformation optimization and obtain a smooth mesh model, we add regularization using a discretization of the continuous Laplace-Beltrami operator [7]. Constraints from symmetric object categories can be enforced by directly defining the mesh shape model to be symmetric.

Fig. 2: Top: category-level model before shape optimization. Bottom: instance-level model after shape optimization.

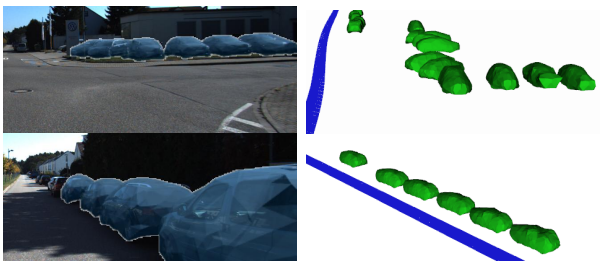

Fig. 3: Left: 2D observation of mesh models. Right: corresponding 3D configuration. Trajectory in blue.

IV. Experiments and Conclusion

We evaluate the ability of the proposed localization and mapping technique to optimize the camera trajectory and reconstruct object poses and shapes using real-world KITTI data. Our experiments use images from a monocular camera and inertial odometry information and focus on detecting, localizing and reconstructing cars. Fig. 2 and 3 show some qualitative results.

This work demonstrates that object categories, shapes and poses can be recovered from visual semantic observations. The key innovation is the development of differentiable keypoint and segmentation mask projection models that allow object shape to be used for simultaneous semantic mapping and camera pose optimization. In contrast with existing techniques, our method generates accurate instance-level reconstructions of multiple objects, incorporating multi-view semantic information.

References

[1] G. Pavlakos, X. Zhou, A. Chan, K. G. Derpanis, and K. Daniilidis, "6-DoF Object Pose from Semantic Keypoints," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[2] X. Zhou, A. Karpur, L. Luo, and Q. Huang, "Starmap for category-agnostic keypoint and viewpoint estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 318–334.

[3] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn." *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[4] 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.

[5] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[6] A. Mourikis and S. Roumeliotis, "A Multi-State Constraint Kalman Filter for Vision-aided Inertial Navigation," University of Minnesota, Tech. Rep., 2006.

[7] O. Sorkine, "Differential representations for mesh processing," in *Computer Graphics Forum*, vol. 25, no. 4. Wiley Online Library, 2006, pp. 789–807.