Towards a Hybrid Scene Representation in VSLAM

Georges Younes
System Design Engineering
University of Waterloo
Waterloo, Ontario, Canada
Email: gyounes@uwaterloo.ca

Daniel Asmar
Mechanical Engineering
American University of Beirut
Beirut, Lebanon
Email: da20@aub.edu.lb

John Zelek
System Design Engineering
University of Waterloo
Waterloo, Ontario, Canada
Email: jzelek@uwaterloo.ca

***Abstract*—In their basic form, Visual Simultaneous Localization and Mapping (VSLAM) systems rely on metric maps to produce meaningful camera pose measurements. As researchers shift their focus towards integrating higher levels of data representations, we argue that the performance of such systems requires a mutually beneficial, tight integration between metric, topological and semantic maps. Therefore, we emphasize the need for a hybrid locally semi-dense and globally sparse metric map and explore the multi-level hybrid map representation interactions for VSLAM and how they can be mutually beneficial.**

I. Introduction

The use of cameras, as sensory input for Visual Simultaneous Localization and Mapping (VSLAM), provides a wide range of input data, resulting in a multitude of map representation possibilities. VSLAM, in its basic form, evolve around generating a 3D representation of the environment using geometric primitives (keypoints, edges, etc.)[7, 8] in what is referred to as *metric maps*. Realizing that such representation becomes computationally intractable in large environments, researchers attempted to forfeit geometric information in favor of connectivity information through *Topological maps*[2]; however, metric measurements were still needed to localize the camera in a meaningful way, i.e. relative metric distances estimation, obstacle avoidance, low-level motor control, etc. Unfortunately, the conversion from a topological to a metric map is not a trivial process, and hybrid maps were introduced [11, 6, 10] where a scene is locally metric and globally topological.

With the recent advancements in machine learning, researchers are moving towards adopting another level of data abstraction: semantic information. Semantic information allows for VSLAM to reason about the nature of the observed scene, paving the way for basic cognitive abilities such as object detection, structural support information, scene understanding, etc. all of which can be capitalized on for increased VSLAM robustness and exploited to broaden VSLAM range of capabilities.

While the incorporation of semantic data into VSLAM is undoubtedly the next step in the right direction, we argue, drawing from the example of topological maps, that such integration requires a hybrid approach that tightly integrates metric, topological and semantic maps such that it is mutually beneficial for all three. Furthermore, we suggest that it is imperative to make sure that the metric representation is as robust and accurate as possible.

In the aforementioned mindset, we suggest to fuse the two main data sources for VSLAM used for metric map reconstructions, namely feature-based and direct frameworks into an intermediate scene representation, where a scene is locally semi-dense and globally sparse. By analyzing their corresponding traits, which are different but complementary, we realize that an intermediate metric representation can overcome the limitations of each framework alone, for a robust, scalable, and accurate VSLAM pipeline. Moreover, we argue that the ability to vary the density of the metric map is a crucial step towards integrating semantic information: feature-based methods are too sparse, maintaining a global semi-dense map becomes computationally intractable; a metric map that is locally semi-dense and globally sparse becomes a desirable setup for which semantic labels can be encoded. We develop this idea further by suggesting a set of multi-level map interactions between semantics, feature-based and semi-dense representations.

II. Metric based VSLAM paradigms

A. Feature based methods

Feature based methods process 2D images to extract *salient* geometric primitives such as Keypoints (corners), edges, lines, etc. The pixel patterns surrounding these features are manipulated to generate descriptors as a quantitative measure of similarity to other features, through what is referred to as *data association*. ORB SLAM, currently considered the defacto in feature based VSLAM, relies on ORB descriptors; as such, ORB SLAM can handle relatively large baseline between frames and tolerate, to an extent, illumination changes. When compared to the direct framework, it has a relatively compact scene representation that allows for failure recovery, loop closure, and allows for global/local optimizations. On the other hand, the density of the features used is inversely correlated with the data association performance: as the density of the extracted features increases, their distinctiveness decreases, leading to ambiguous feature matches. As such, feature-based methods are limited to sparse representations. Their performance is brittle in low textured or self repeating environments and are unstable for far-away features or features that have been observed with a small parallax. Most importantly, data association in feature based methods is established between features extracted at discretized locations in the images, resulting in an inferior accuracy and larger drift than the direct framework.

B. Direct methods

In contrast to feature-based methods, direct methods use raw pixel intensities as inputs: no intermediate image representation is computed. The direct pipeline proceeds by minimizing a **photometric** error defined over pixels with a gradient in the image using image alignment methods [1]. As such the direct framework can make use of virtually all image pixels with an intensity gradient. It naturally handles points at infinity and points observed with small parallax using the inverse depth parametrization [3], and since no explicit data association is required, it can have semi-dense or sparse scene representations. Most importantly, the photometric error can be interpolated over the image domain resulting in an image alignment with sub-pixel accuracy and relatively less drift than feature-based methods as shown in DSO [5]. However, it requires an accurate prior on the relative pose between frames (limited to small baseline motions), does not naturally allow for loop closures nor failure recovery (probabilistic appearance based methods are required [4]), is brittle against any sources of outliers in the scene (dynamic objects, occlusions), and requires a spatial regularization when a semi-dense representation is employed. Most importantly, it becomes intractable for very large scene exploration scenarios.

III. Map interactions

A. feature-based - direct maps interactions

When the corresponding pros and cons of both feature-based and direct frameworks are placed side by side, a pattern of complementary traits emerges. For example, direct methods require small baseline motions to ensure convergence, whereas feature-based methods are more reliable at relatively larger baselines. Furthermore, due to sub-pixel alignment accuracy, direct methods are relatively more accurate but suffer from an intractable amount of data in large environments; on the other hand feature-based methods suffer from relatively lower accuracies due to the discretization of the input space but have a suitable scene representation for a SLAM formulation, which enables feature-based methods to easily maintain a reusable global map, perform loop closures and failure recovery. Therefore an ideal pipeline should exploit both direct and feature-based advantages to benefit from the direct formulation accuracy and robustness while making use of feature-based methods for large baseline motions, maintaining a reusable global map, and reducing drifts through loop closures. Furthermore, a hybrid feature-based-direct framework allows for the metric representation to be locally semi-dense and globally sparse, facilitating interactions with other types of representations such as topological and/or semantic, while maintaining scalability and computational tractability.

B. sparse metric maps and semantic labels interactions

In a feature-based framework, the 2D features are formed by sparsely sampling the 3D world surface into 2D geometric primitives, significantly abstracting the properties of the surface structure itself; after which the geometric primitives are treated as conditionally independent entities. Integrating semantic labels in such formulation is particularly challenging as the geometric primitives are usually defined over small areas in the image domain (*e.g.* keypoints are defined at a single pixel), whereas semantic labels are associated with objects. Nevertheless, semantic labels can be integrated in the sparse map in many ways: Semantic labels can be used to identify potentially dynamic objects in the scene, allowing VSLAM to properly handle image regions that violate the static scene assumption; furthermore the set of features associated with such regions can be grouped together to form the notion of dynamic objects, which can be tracked separately from the traditional static map, allowing VSLAM to perform path collision detection and dynamic object tracking within the same framework. Also, descriptors of 2D features, associated with semantic labels of rigid objects in the static map, can be grouped together to form a global based descriptor for observed objects in the scene. Such formulation allows for an object-based VSLAM framework that preserves the rigidity of the reconstructed objects, as opposed to the current conditionally independent 3D features reconstructions.

C. semi-dense metric map and semantic labels interactions

Semi-dense metric maps are highly susceptible to outliers from dynamic objects and occlusions; semantic labels can significantly reduce the outlier's spurious effects by flagging out potential dynamic objects and by providing consistency checks over occluded pixels, in different images, that belong to different semantic labels. Furthermore, semantic labels can benefit the semi-dense regularization process by tightly integrating the labels into the regularization, such that local points belonging to the same labels with discontinuous 3D reconstructions are penalized [9]. Also, semantic labels can help increase the density of semi-dense maps through planar detection (*e.g.* a 3D surface of a road is a continuously smooth surface, and support information *e.g.* a building has to be supported by a ground at its bottom.

On the other hand, semantic labeling can benefit from integrating structural and temporal measurements, provided from the semi-dense map, to provide classification priors on subsequent images. Furthermore, the semi-dense map can provide geometric consistency checks on the 2D classifications procedure, reducing outliers and increasing prediction accuracy.

IV. Conclusion

We have presented the different interactions between multi-level hybrid maps and discussed how such representations can mutually benefit one another. We also came to the conclusion that a fusion between a feature-based and a direct metric representation is an important first step towards multi-level hybrid maps in VSLAM as it increases the accuracy and robustness of the metric map as well as gains control over the sparsity of the metric representation.

REFERENCES

[1] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *Int. J. Comput. Vision*, 56(3): 221–255, February 2004. ISSN 0920-5691. doi: 10.1023/B:VISI.0000011205.11775.fd. URL http://dx.doi.org/10.1023/B:VISI.0000011205.11775.fd.

[2] Jaime Boal, Álvaro Sánchez-Miralles, and Álvaro Arranz. Topological simultaneous localization and mapping: a survey. *Robotica*, 32(5):803–821, Aug 2014.

[3] J. Civera, A. J. Davison, and J. M. M. Montiel. Inverse depth parametrization for monocular slam. *IEEE Transactions on Robotics*, 24(5):932–945, Oct 2008. ISSN 1552-3098. doi: 10.1109/TRO.2008.2003276.

[4] Mark Cummins and Paul Newman. Appearance-only slam at large scale with fab-map 2.0. *The International Journal of Robotics Research*, 30(9):1100–1123, 2011. doi: 10.1177/0278364910385483. URL http://dx.doi.org/10.1177/0278364910385483.

[5] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2017.2658577.

[6] Jakob Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: Large-Scale Direct Monocular SLAM. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014 SE - 54*, volume 8690 of *Lecture Notes in Computer Science*, pages 834–849. Springer International Publishing, 2014.

[7] Georg Klein and David Murray. Parallel tracking and mapping for small AR workspaces. In *Proc. Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'07)*, Nara, Japan, November 2007.

[8] Georg Klein and David Murray. Improving the agility of keyframe-based SLAM. In *Proc. 10th European Conference on Computer Vision (ECCV'08)*, pages 802–815, Marseille, October 2008.

[9] Xuanpeng Li and Rachid Belaroussi. Semi-dense 3d semantic mapping from monocular SLAM. *Arxiv*, abs/1611.04144, 2016. URL http://arxiv.org/abs/1611.04144.

[10] Hyon Lim, Jongwoo Lim, and H J Kim. Real-time 6-DOF monocular visual SLAM in a large-scale environment. In *Robotics and Automation (ICRA), IEEE International Conference on*, pages 1532–1539, May 2014.

[11] Raul Mur-Artal, J. M. M. Montiel, and Juan D. Tardos. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics*, PP(99): 1–17, 2015.