Directed Research: Semantic Visual Odometry

Sai Ramana Kiran Pinnama Raju
Email: spinnamaraju@wpi.edu

I. Introduction

Visual Odometry (VO) and Simultaneous Localization and Mapping (SLAM) are an integral component in several robotic and wearable technologies such as Unmanned Aerial Vehicles (UAV), Autonomous Ground Vehicles (AGV), Augmented (AR) and Virtual Reality (VR) headsets. With the advancements in silicon chips and faster NN computation, data driven approaches to these fields have gained more traction among research communities. Classical approaches to VO or SLAM usually involves tracking pixel intensities using optical flow or extracting low level features. Although these approaches perform decently well in several datasets, the features are short-term and often needs to be recomputed under different exposures or camera settings. These limitations can be reduced by tracking high level information instead. This write-up is a summary of related work, current problems and some potential ways to solve.

II. Towards "Robust" Visual Odometry

A visual odometry component is vital in several applications and hence it is important to clearly define what is considered as "Robust" Visual odometry. A philosophical non-exhaustive list towards this definition would be the following

1) Robust towards partial or blinded sensor measurements
2) Tracking across different environment conditions and motion conditions (example: high speed)
3) As close to real time as possible in execution
4) Identifying and tracking in case of occlusions
5) Demarcating static and dynamic objects in the field
6) Minimal drift over a long run
7) Cheaper to compute on smaller robots and devices

In this work, I am trying to address points 2,3,4 using semantic features of the scene and deep neural networks. The following section expands on the related work that has been done so far on the points of interest.

III. Related Work

Most of the traditional approaches, are classified into direct and indirect methods. Direct methods use the pixel intensities and construct optical flow optimization problems. These methods are computationally expensive and hence slower since the density of data needs to be processed is high, however the odometry decisions are based on more information making it more reliable. Some of the popular works like [1] take this method a bit further by sampling the pixels across frames and optimize the photometric error, thereby decreasing the computational requirements.

On the other hand, indirect methods like [2] primarily track corners and other geometric features for tracking. These methods are computationally inexpensive compared to direct methods but suffer from viewpoint variances. Moreover, their odometry information is based on very selective data points from the frames and might not be as reliable as direct methods output.

Irrespectively, both of these methods need hand tailored feature selection and needs different parameter or optimizer selection based on different scenarios. Moreover, these algorithms are working based off of low-level features making them unreliable for aspects mentioned in 2,4. Modern techniques approach this problem by making use of data points collected over the years and generalize the solution by adding more "intelligence" to the system. Some of the current research use DNNs in the following ways,

1) Enhancing low-level feature tracking with semantic information [3],[4],[5]
2) Dynamic object masks using semantic networks [8]
3) End-to-end pose estimation of camera [9],[10],[11]
4) Tracking using semantic networks alone [13],[14]
5) Embedding memory into the pipelines [15],[16]
6) Using semantic maps to construct full 3D maps [6],[7]

IV. Problems of Interest

On a high level, current research in 1 strike a good balance of enhacing robustness of features by mapping them to a semantic label. This improves their view invariance and tracking. However, some of these methods do not perform well in real time and are highly dependent on semantic labelling. Methods in 3 are interesting note, since they perform the entire camera pose estimation using neural networks. Inspiring by this methodology, we can train different aspects like geometric features, optical flow independently and combine them together for better explainability and robustness. Works like [12] expand on this end-to-end approach by incorporating both stereo and inertial measurements in network training. Another way to expand this could be is integrating visual transformer based networks along with inertial sensors to further improve the odometry accuracy

References

[1] Engel, J., Koltun, V. and Cremers, D., 2017. Direct sparse odometry. IEEE transactions on pattern analysis and machine intelligence, 40(3), pp.611-625.

[2] R. Mur-Artal, J. D. Tardos, "ORB-SLAM2: an Open-Source SLAM system for monocular stereo and RGB-D cameras", 2016.

[3] H. -J. Liang, N. J. Sanket, C. Fermüller and Y. Aloimonos, ”SalientDSO: Bringing Attention to Direct Sparse Odometry,” in IEEE Transactions on Automation Science and Engineering, vol. 16, no. 4, pp. 1619-1626, Oct. 2019, doi: 10.1109/TASE.2019.2900980.

[4] Yu, C., Liu, Z., Liu, X.J., Xie, F., Yang, Y., Wei, Q. and Fei, Q., 2018, October. DS-SLAM: A semantic visual SLAM towards dynamic environments. In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 1168-1174). IEEE.

[5] K.-N. Lianos, J. L. Schonberger, M. Pollefeys, and T. Sattler, “VSO: Visual semantic odometry,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 234–250.

[6] McCormac, J., Handa, A., Davison, A. and Leutenegger, S., 2017, May. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In 2017 IEEE International Conference on Robotics and automation (ICRA) (pp. 4628-4635). IEEE.

[7] Hermans, A., Floros, G. and Leibe, B., 2014, May. Dense 3d semantic mapping of indoor scenes from rgb-d images. In 2014 IEEE International Conference on Robotics and Automation (ICRA) (pp. 2631-2638). IEEE.

[8] J. Liu, X. Li, Y. Liu and H. Chen, ”RGB-D Inertial Odometry for a Resource-Restricted Robot in Dynamic Environments,” in IEEE Robotics and Automation Letters, vol. 7, no. 4, pp. 9573-9580, Oct. 2022, doi: 10.1109/LRA.2022.3191193.

[9] Qian, Z., Patath, K., Fu, J. and Xiao, J., 2021, May. Semantic slam with autonomous object-level data association. In 2021 IEEE International Conference on Robotics and Automation (ICRA) (pp. 11203-11209). IEEE.

[10] Wang, S., Clark, R., Wen, H. and Trigoni, N., 2017, May. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In 2017 IEEE international conference on robotics and automation (ICRA) (pp. 2043-2050). IEEE.

[11] Li, R., Wang, S., Long, Z. and Gu, D., 2018, May. Undeepvo: Monocular visual odometry through unsupervised deep learning. In 2018 IEEE international conference on robotics and automation (ICRA) (pp. 7286-7291). IEEE.

[12] Han, L., Lin, Y., Du, G. and Lian, S., 2019, November. Deepvio: Self-supervised deep learning of monocular visual inertial odometry using 3d geometric constraints. In 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 6906-6913). IEEE.

[13] H. Mahé, D. Marraud and A. I. Comport, ”Semantic-only Visual Odometry based on dense class-level segmentation,” 2018 24th International Conference on Pattern Recognition (ICPR), 2018, pp. 1989-1995, doi: 10.1109/ICPR.2018.8545263.

[14] M. Herb et al., ”Semantic Image Alignment for Vehicle Localization,” 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021, pp. 1124-1131, doi: 10.1109/IROS51168.2021.9636517.

[15] Xue, F., Wang, X., Li, S., Wang, Q., Wang, J. and Zha, H., 2019. Beyond tracking: Selecting memory and refining poses for deep visual odometry. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8575-8583).

[16] Xue, F., Wang, X., Wang, J. and Zha, H., 2020. Deep visual odometry with adaptive memory. IEEE Transactions on Pattern Analysis and Machine Intelligence.