Semantic and Geometric Features in SLAM

Jialin Li
Supervisor: Simon Julier, Ziwen Lu

June 2022

1 Motivation

Traditional SLAM system uses image intensity to track features across frames. However, image intensity could change a lot in different day time, weather condition and seasons. Flares and shadows could also interrupt image intensity consistency. On the other hand, semantic information is more robust and time-invariant compared to intensity. Intensity values range from 0 to 255 and usually has 3 channels (RGB). Semantic values usually only have a dozen of values and 1 channel. Hence, semantic information takes less storage and computation source than intensity information. This paper presents a feature tracking method using semantic information.

2 Challenges and relative work

With the advancement of semantic segmentation algorithms, more and more researches have been conducted to include semantic information in feature tracking. Tong et al. [5] applied semantic features in visual SLAM system for autonomous driving in parking lot. They synthesized an IPM image from the images captures by four cameras surrounding a car. Then they performed semantic segmentation on the IPM image and used this semantic information to build a local map. The semantic information, together with an odometry consisted with IMU and wheel encoders are used to localise the vehicle in the map. The accuracy is at centimeter-level. However, the high demand on sensors makes this proposed method a bit hard to be applied widely. And the system's application is limited in parking lot. Xiao et al. [6] combined semantic information with geometric features for feature tracking. The algorithm is similar with traditional feature-based tracking in which reprojection error is minimised, except the features will only be matched with features in the same semantic channel. This could make the feature tracking more robust compared to traditional feature tracking methods, but it makes the procedures even more complex. In other words, it adds semantic information to traditional feature-based tracking method rather than replace the traditional method with semantic method. Pauls et al. [3] and Petek el al. [4] employs semantic features to localize vehicle

in HD maps by matching semantic features with features in HD maps. The accuracy is relatively high but it requires a pre-built HD map. The same goes for [1]. They proposed a direct model for camera pose estimation based on mutual information. They optimise camera pose such that the mutual information between the image captured by a camera and the virtual image reprojected from a 3D model from that camera pose is maximised. Mutual information is calculated based on semantic labels instead of image intensity. Not only this method requires a pre-build 3D model of the environment, it needs to render lots of virtual images which is very computationally expensive.

3 Proposed method

I propose a 2D-2D template matching method based on semantic mutual information. Instead of matching features across the entire image, it focuses on part of the image at one time. It matches a patch of a image with another image (consecutive images). There is no need of a 3D model. The equation to calculate mutual information is adapted from [2]:

$$I(A,B) = \sum p_{AB}(a,b) log \frac{p_{AB}(a,b)}{p_A(a)p_B(b)}$$

$$p_{AB}(a,b) = \frac{h(a,b)}{\sum_{a,b} h(a,b)}$$

$$p_A(a) = \sum_a p_{AB}(a,b)$$

$$p_B(b) = \sum_b p_{AB}(a,b)$$

where A, B are a pair of pixels from reference image and template image, a, b are possible semantic labels of A and B, $h(a,b)$ is the number of time A has semantic label a, and B has semantic label b. $\sum_{a,b} h(a,b)$ is the total number of pixels in template image.

Sliding the template image across the reference image, we calculate I for every sliding position. This would give us a matrix full of I values with the size of (reference image rows - template image rows + 1)x(reference image columns - template image columns + 1). The index of the maximum element in this matrix is the coordinate of the best matching position.

However, there are several problems. First, the maximum I may not be the best matching position. The best matching position may have I slightly lower than the maximum I. To solve this issue, the highest 3 I are kept, and multiple template images are matched. Each I position would give us a camera pose estimation. The most common pose estimation is kept. Another problem is lack of texture. The texture information in semantic image is usually not as rich as in raw image. If we take a template at an inappropriate position, there would

be infinite number of best matches. A method must be developed to pick template at appropriate locations. Another problem is with semantic segmentation itself. The accuracy and precision of semantic segmentation network may affect the tracking accuracy and robustness. For example, a pixel may be wrongly classified, or the same pixel may be given different semantic labels if we classify it twice. All of these must be taken into consideration for the system to work well.

References

[1] Guillaume Caron, Amaury Dame, and Eric Marchand. Direct model based visual tracking and pose estimation using mutual information. *Image and Vision Computing*, 32(1):54–63, 2014.

[2] Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 1997.

[3] Jan-Hendrik Pauls, Kürsat Petek, Fabian Poggenhans, and Christoph Stiller. Monocular localization in hd maps by combining semantic segmentation and distance transform. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4595–4601. IEEE, 2020.

[4] Kürsat Petek, Kshitij Sirohi, Daniel Büscher, and Wolfram Burgard. Robust monocular localization in sparse hd maps leveraging multi-task uncertainty estimation. *arXiv preprint arXiv:2110.10563*, 2021.

[5] Tong Qin, Tongqing Chen, Yilun Chen, and Qing Su. Avp-slam: Semantic visual mapping and localization for autonomous vehicles in the parking lot. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5939–5945. IEEE, 2020.

[6] Zhongyang Xiao, Kun Jiang, Shichao Xie, Tuopu Wen, Chunlei Yu, and Diange Yang. Monocular vehicle self-localization method based on compact semantic map. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3083–3090. IEEE, 2018.