Vol.2022-CVIM-229 No.6
2022/3/10

Robust Outdoor Panoramic View SLAM using Semantic View Filter

Xiangyu Chu[1,a)] Ryoichi Ishikawa[1,b)] Takeshi Oishi[1,c)]

Abstract: Outdoor visual SLAM is easily affected by dynamic objects, such as moving cars and pedestrians. Moreover, panoramic view tracking has the problem of unstable feature extraction because of the image distortions. We propose a V-SLAM framework with multiple virtual cameras robust to the dynamic environment and computationally efficient while achieving stable camera tracking to solve these issues.

Keywords: Visual-SLAM, Panoramic image tracking, Sensor fusion

1. Introduction

Simultaneous Localization and Mapping (SLAM) aims to simultaneously construct a map of an unknown environment while localizing a mobile platform. When cameras are used as the external perception sensor, we call it visual SLAM. Many researchers study V-SALM based on monocular camera [17], stereo cameras [5], panoramic cameras [9] and so on. Panoramic vision can provide a wider view, more feature points, and rich texture information. V-SLAM systems using panoramic cameras still face these issues like moving vehicle interference, some views being too far, image distortion, no real scale, etc.

In terms of outdoor visual navigation, Y. shi et al. concentrate on BA-SLAM systems using GPS information [14], they use GPS information to optimize pose estimation. Similarly, S. Shen et al. used multi-sensor fusion technology on UAVs, and they mainly used GPS information for global pose optimization [13]. But their GPS information is all pre-calibrated; they don't consider the case when GPS has significant error as shown in Fig. 2. For moving vehicle issues, K, Masaya, et al. [7] proposed a method of using semantic segmentation technology to extract the mask of the vehicle in the field of view and then remove the feature points on the mask. But they are using simulated data and not tested in the real world.

To get more accurate outdoor navigationwe propose a panoramic V-SLAM with semantic view filter. The semantic view filter method has three main properties as follows:

(1) Multi-virtual camera system
(2) GPS fusion in local bundle adjustment
(3) Semantically guided feature selection and pose optimization

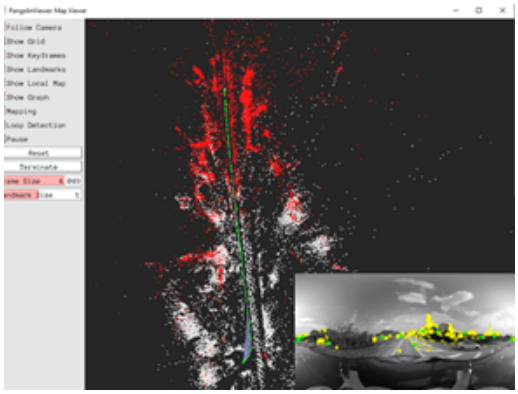

Fig. 1 Running Visual-SLAM.

2. Related work

2.1 GPS loose coupling

How to use GPS cleverly in outdoor navigation is a hot research topic. As we described earlier in Fig. 2, GPS can sometimes have large errors. However, GPS also has advantages such as providing actual geographic information, data errors do not accumulate, sensors are cheap and easy to integrate, etc. Larnaout, D.etc.[8] proposed a penalty term that takes into account the GPS to the convergence of the BA. To avoid with aberrant data has a very bad effect on the BA, they proposed an original BA with inequality constraint IBA. GPS constrain works only if it does not degrade too much the re-projection error.It means that when individual GPS errors occur, the system will not use GPS for attitude estimation.This method has good results when there are only individual GPS errors, but when there are very many consecutive wrong GPS like we go through a tunnel, the results are often less than ideal.

2.2 Semantic segmentation

In the process of Visual SLAM tracking, it tends to perform current pose estimation on all the feature points in the field of

[1] Institute of Industrial Science, The University of Tokyo, Komaba, Tokyo 153–8505, Japan
[a)] xy@cvl.iis.u-tokyo.ac.jp
[b)] ishikawa@cvl.iis.u-tokyo.ac.jp
[c)] oishi@cvl.iis.u-tokyo.ac.jp

Fig. 2 Wrong GPS information in tunnel.

view, but among these feature points collected outdoors, there are many low-quality feature points (such as distant scenes) and wrong feature points (like moving objects). Using these problematic feature points for pose estimation is prone to errors. So many people are committed to using semantic segmentation technology to find these problematic areas, and then do not use the above feature points. As K, Masaya, et al. [7] did ,they removed areas of sky and vehicles from view,This will make the remaining points more credible, and the pose estimation will be more accurate.But we are using panoramic images, which cannot be easily migrated. Xu.etc[18] proposed a semantic segmentation model of panoramic images. They used synthetic data such as SYNTHIA and expand the panoramic image like a cylinder in perspective.Greatly reduces distortion issues in panoramic views.They have achieved good results on simulated datasets, but the problem is that it is still uncertain whether they can perform equally well in the real world.But the way they expands the panoramic image gives us some inspiration.

3. Proposed V-SLAM framework

As shown in Fig. 4. We followed the ORB-SLAM framework [10][11][2] and made changes in the tracking module and optimization module.

3.1 View Filter for Panoramic SLAM

In the equirectangular projection image, we can find apparent distortion. We build multiple virtual cameras to convert the panoramic image into multiple perspective images to solve this problem. Our multi-virtual camera framework maps the equirectangular projection image to 3D spherical space. Then we choose some directions that we are interested in and render multiple perspective images at each frame as shown in Fig. 3.

3.1.1 Virtual Camera Generation Module

Generating virtual cameras relies on the camera center determined by the 3D points in space. These 3D points provide direction and depth information which are the key Information for rendering. It is important to avoid selecting the 3D points with too large or too small depths as the virtual camera centers.

In practice, there are two corresponding situations. One is the initialization phase. Because there is no virtual camera yet, virtual cameras need to be generated in multiple directions. After the initialization in the tracking phase, we only need to generate a virtual camera in the forward direction area.

Fig. 3 Rendered multiple perspective images. Multiple perspective views are rendered from a panoramic frame.

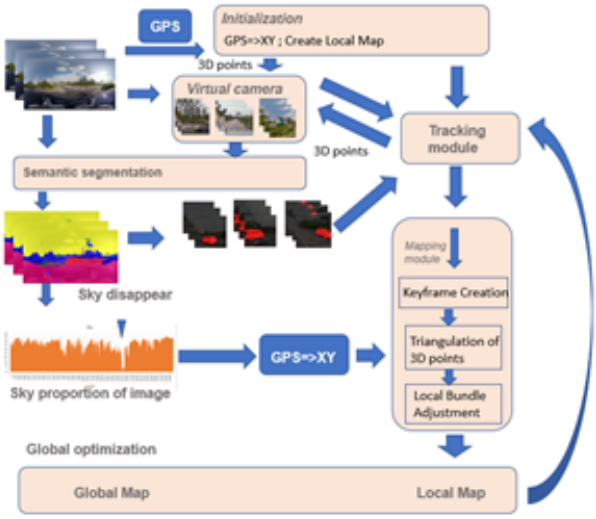

Fig. 4 Overview of proposed Visual-SLAM.

Fig. 5 Virtual camera generation area. Left is in initialization, right is in tracking.

3.1.2 Virtual Camera Removal Module

Many selected 3D points will move farther away from the platform as the mobile platform moves forward. These distant 3D points are very detrimental to the accuracy of pose estimation. We will remove those points whose depth exceeds a certain threshold in practice. Also, as the number of virtual cameras increases, more and more virtual cameras will have overlapping parts. It is

Fig. 6 Vehicle mask by semantic segmentation.

Fig. 7 ORB Feature points mapping from rendered perspective images to panoramic image.

necessary to remove some virtual cameras promptly to reduce repeated calculations. The distance between 3D points between two virtual camera centers and the angle observed from the platform is used to evaluate the overlap.

3.1.3 Semantic view filter

Semantic segmentation [16] is the process of clustering image regions that belong to the same object class for a given image. We extracted the mask of the vehicle using semantic segmentation as shown in Fig. 6. After getting the mask of the vehicles, we can choose not to extract feature points from them. This process ensures that our pose estimation is not affected by the vehicles.

3.1.4 Feature extraction

ORB feature [12] points are extracted in the perspective image, and then the feature points are mapped from each perspective image back to the panoramic space. The feature points extracted from the vehicles, as shown in Fig. 7.

3.2 GPS fusion

As we all know, Global Positioning System (GPS) is a satellite-based radio navigation system. Since GPS height information is not reliable, only xy coordinates are used.

3.2.1 Map initialization

The map in SLAM system requires absolute direction and scale. The GPS positions are used for pose estimation in the initialization phase. The algorithm sets the initial frame at the origin of the coordinates and starts checking the x-y changes of the next frames. In this phase, some frames with a small x-y change are omitted. Then it can obtain the rotation and translation of the n-th frame from the x-y position as the initial frame. The initial direction and scale are fixed by making the map from the origin and the initial frame.

3.2.2 Reliability of GPS measurement

As shown in Fig. 2, GPS sometimes has large errors. We evaluate the reliability of GPS data. GPS measures the positions by receiving satellite signals. When the platform is in a tunnel or the center of a city with many tall buildings, there is signal ob-

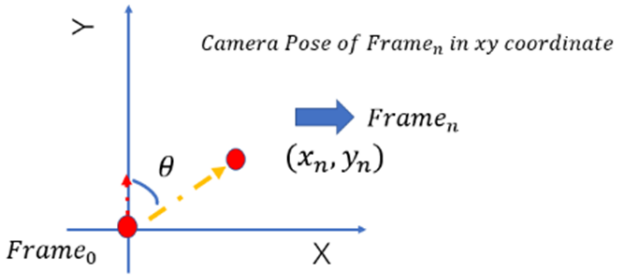

Fig. 8 Camera pose estimation by GPS.

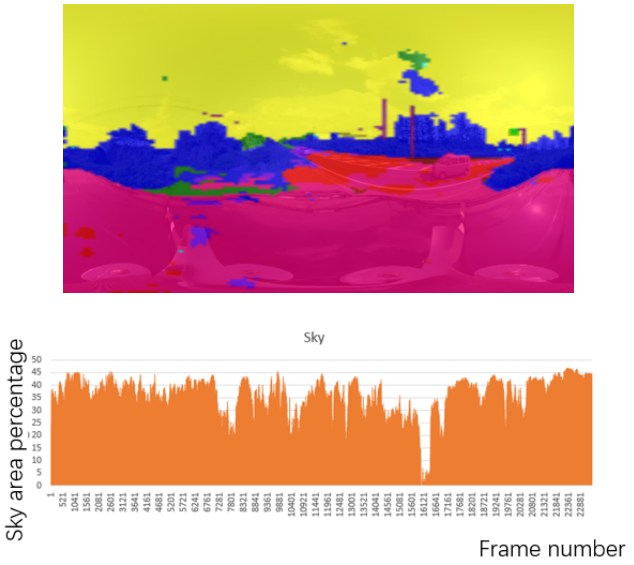

Fig. 9 Semantic segmentation on panoramic image and sky proportion information.

struction, GPS measurement is often inaccurate. Therefore, by checking the proportion of the sky in the field of view, we can infer whether the camera is currently in an open area that is easy to receive GPS signals or in the areas that are difficult to receive GPS signals.

As shown in the upper part of Fig. 9, We can extract sky areas separately using semantic segmentation. Then we can get a proportion of the sky at different frames. We can see that when passing through the tunnel, the sky completely cannot be observed.

3.2.3 GPS fusion in local BA

In several SLAM systems, the bundle adjustment is applied to past sequential frames: local Bundle Adjustment (BA). We use the GPS information in the local BA to constrain the pose estimation in the global coordinate. Pose graph optimization (PGO) [3] is widely used for nonlinear optimization in local BA. The vertices are poses to be optimized, and edges are constraints in PGO.

As shown in Fig. 10, the pose a keyframe is a vertex, and the relationship between keyframes in the 3d space and the relative pose constitutes a constraint. Usually, the constraint is optimized in the fastest descent of bundle adjustment [1].

$$Loss_{all} = Loss_{original} + W * \frac{w}{1 + e^{-20p_s+8}}(P_e - P_{gps}) \qquad (1)$$

Here we add GPS information as a unary edge; the constraint is the difference between the x-y position of the pose and the GPS-transformed x-y position. We use a nonlinear weight to control

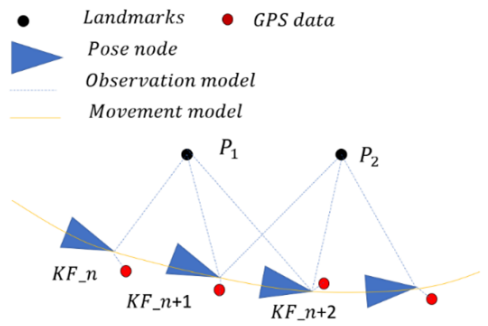

Fig. 10 Local bundle adjustment with GPS measurement.

Fig. 11 Sample frames.

the effect of GPS reliability on PGO.

4. Implementation and experiments

4.1 Dataset

We used sequential street-view videos of national roads as the dataset. These videos have been taken with cheap and lightweight panoramic cameras: Gopro fusion. The lengths are more than a few kilometers apart, and there are both rural and urban roads.

4.2 Semantic Segmentation

We finally chose the DDRNET model [6] because DDRNET employs a lightweight architecture reasoning on low-resolution images and realizes very fast scene parsing. Since we did not have a labeled training dataset, we fine-tuned the model using the Cityscapes dataset [4]. Since some of our rendered perspective images only contain a part of the image, we used a multi-scale data augmentation method when training the model.

4.3 Experimental results

We show the performance comparison with OpenVSLAM [15] and GPS information. OpenVSLAM employs several key features of SLAM in ORB-SLAM2. As shown in Fig.12, OpenVSLAM lost tracking in halfway during outdoor long-distance tracking. Our proposed system ran the whole course and maintained a certain accuracy.

Since OpenVSLAM does not have a real scale, it is impossible to evaluate it by the scale. However, the proposed VSLAM obtains the real scale from GPS information to be compared with GPS in more detail. In Fig. 13, on longer journeys, we can see that the GPS information entirely coincides with the trajectory we generated. We can only see the subtle differences between the trajectory generated by the proposed VSLAM and GPS data in the zoomed images.

The low proportion of the sky has an almost negligible weight of GPS in optimization. Therefore, there was no interference from the wrong GPS. It successfully passed the tunnel, and the generated map is also quite correct. It clarifies the robustness of our proposed VSLAM system.

Fig. 12 Trajectory comparison. Left: generated trajectory by OpenVSLAM. Middle: GPS data from GSI. Right: generated trajectory by proposed VSLAM .

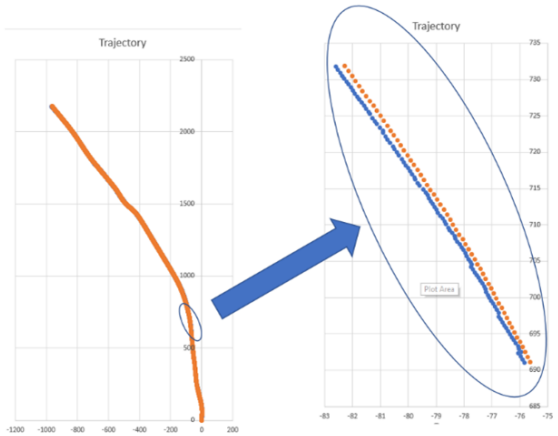

Fig. 13 Trajectory comparison in detail. Orange paths are from GPS. Blue paths are from proposed VSLAM.

Fig. 14 Trajectory details around tunnel. Left: path is from GPS. Right: path is from proposed VSLAM.

4.3.1 Tracking quality

Since we use multiple virtual cameras as the semantic view filter, as shown in Fig. 15, the proposed method did not extract feature points from the moving vehicles. Accordingly, the pose estimation is more accurate than the conventional method.

Since our pose estimation is more accurate, we can extract more high-quality feature points. As shown in Fig. 16. Compared with OpenVSLAM, the number of feature points that are tracked on the map has increased. It also shows that our proposed VSLAM is more robust than OpenVSLAM.

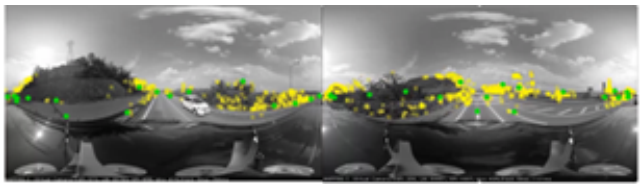

Fig. 15 Tracking frame by proposed VSLAM.

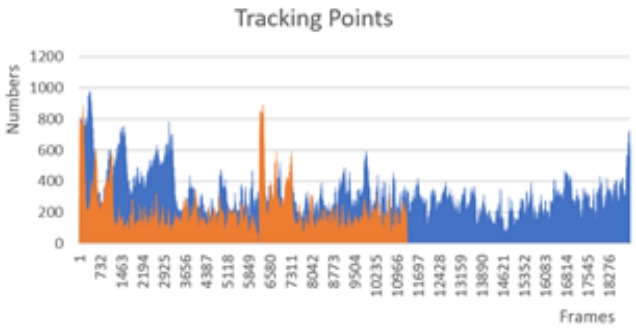

Fig. 16 Tracking points comparison.

5. Conclusion

The authors believe that the VSLAM system should have stronger adaptability and a good accuracy rate on some large outdoor maps, rather than only good results on indoor or outdoor small maps. This paper presents a VSLAM framework based on multiple virtual cameras with a semantic view filter. The proposed method also uses GPS information combined with the semantic view filter. In the experiment, we have shown that the proposed VSLAM achieved good results on outdoor data. It still has very good performance in various complex environments, including urban or rural scenes, reflecting the robustness of the model.

Due to the semantic segmentation inference of multiple perspectives in each frame, it's difficult to work in real-time. We are looking for a deep combination method of semantic information and feature points, which we hope could reduce the computation time while ensuring tracking quality.

References

[1] Agarwal, S., Snavely, N., Seitz, S. M. and Szeliski, R.: Bundle Adjustment in the Large, *Computer Vision – ECCV 2010* (Daniilidis, K., Maragos, P. and Paragios, N., eds.), Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 29–42 (2010).

[2] Campos, C., Elvira, R., Rodríguez, J. J. G., Montiel, J. M. M. and Tardós, J. D.: ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM, *IEEE Transactions on Robotics*, Vol. 37, No. 6, pp. 1874–1890 (online), DOI: 10.1109/TRO.2021.3075644 (2021).

[3] Carlone, L., Tron, R., Daniilidis, K. and Dellaert, F.: Initialization techniques for 3D SLAM: A survey on rotation estimation and its use in pose graph optimization, *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4597–4604 (online), DOI: 10.1109/ICRA.2015.7139836 (2015).

[4] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223 (online), DOI: 10.1109/CVPR.2016.350 (2016).

[5] Engel, J., Stückler, J. and Cremers, D.: Large-scale direct SLAM with stereo cameras, *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1935–1942 (online), DOI: 10.1109/IROS.2015.7353631 (2015).

[6] Hong, Y., Pan, H., Sun, W. and Jia, Y.: Deep Dual-resolution Networks for Real-time and Accurate Semantic Segmentation of Road Scenes, *arXiv preprint arXiv:2101.06085* (2021).

[7] Kaneko, M., Iwami, K., Ogawa, T., Yamasaki, T. and Aizawa, K.: Mask-SLAM: Robust Feature-Based Monocular SLAM by Masking Using Semantic Segmentation, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 371–3718 (online), DOI: 10.1109/CVPRW.2018.00063 (2018).

[8] Larnaout, D., Gay-Bellile, V., Bourgeois, S. and Dhome, M.: Vehicle 6-dof localization based on slam constrained by gps and digital elevation model information, *2013 IEEE International Conference on Image Processing*, IEEE, pp. 2504–2508 (2013).

[9] Lemaire, T. and Lacroix, S.: SLAM with panoramic vision, *Journal of Field Robotics*, Vol. 24, pp. 1–2 (2007).

[10] Mur-Artal, R., Montiel, J. M. M. and Tardós, J. D.: ORB-SLAM: A Versatile and Accurate Monocular SLAM System, *IEEE Transactions on Robotics*, Vol. 31, No. 5, pp. 1147–1163 (online), DOI: 10.1109/TRO.2015.2463671 (2015).

[11] Mur-Artal, R. and Tardós, J. D.: ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras, *IEEE Transactions on Robotics*, Vol. 33, No. 5, pp. 1255–1262 (online), DOI: 10.1109/TRO.2017.2705103 (2017).

[12] Rublee, E., Rabaud, V., Konolige, K. and Bradski, G.: ORB: An efficient alternative to SIFT or SURF, *2011 International Conference on Computer Vision*, pp. 2564–2571 (online), DOI: 10.1109/ICCV.2011.6126544 (2011).

[13] Shen, S., Mulgaonkar, Y., Michael, N. and Kumar, V.: Multi-sensor fusion for robust autonomous flight in indoor and outdoor environments with a rotorcraft MAV, *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4974–4981 (online), DOI: 10.1109/ICRA.2014.6907588 (2014).

[14] Shi, Y., Ji, S., Shi, Z., Duan, Y. and Shibasaki, R.: GPS-Supported Visual SLAM with a Rigorous Sensor Model for a Panoramic Camera in Outdoor Environments, *Sensors*, Vol. 13, No. 1, pp. 119–136 (online), DOI: 10.3390/s130100119 (2013).

[15] Sumikura, S., Shibuya, M. and Sakurada, K.: OpenVSLAM: A Versatile Visual SLAM Framework, *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA, Association for Computing Machinery, p. 2292–2295 (online), DOI: 10.1145/3343031.3350539 (2019).

[16] Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X. and Cottrell, G.: Understanding Convolution for Semantic Segmentation, *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Los Alamitos, CA, USA, IEEE Computer Society, pp. 1451–1460 (online), DOI: 10.1109/WACV.2018.00163 (2018).

[17] Weiss, S., Scaramuzza, D. and Siegwart, R.: Monocular-SLAM–Based Navigation for Autonomous Micro Helicopters in GPS-Denied Environments, *J. Field Robot.*, Vol. 28, No. 6, p. 854–874 (online), DOI: 10.1002/rob.20412 (2011).

[18] Xu, Y., Wang, K., Yang, K., Sun, D. and Fu, J.: Semantic segmentation of panoramic images using a synthetic dataset, *Artificial Intelligence and Machine Learning in Defense Applications*, Vol. 11169, International Society for Optics and Photonics, p. 111690B (2019).