IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9554752

A PROPOSAL TO INTEGRATE ORB-SLAM FISHEYE AND CONVOLUTIONAL NEURAL NETWORKS FOR OUTDOOR TERRESTRIAL MOBILE MAPPING

Thaisa Aline Correia Garcia[1], Mariana Batista Campos[2], Letícia Ferrari Castanheiro[1], Antonio Maria Garcia Tommaselli[1]

[1] São Paulo State University, Unesp, Presidente Prudente, SP, Brazil.
[2] Finnish Geospatial Research Institute, FGI, Kirkkonummi, Finland

ABSTRACT

SLAM methods, such as ORB-SLAM, can build a map of an unknown environment (sparse point cloud) with optical images. The sensor motion provides image sequences over which keypoints are extracted and matched, enabling the simultaneous computation of sensor locations and 3D coordinates of points. In the last years, enormous progress has been done to solve the SLAM problem, especially focusing on computational efficiency and accurate sensor trajectory estimation. However, the auto-detection of incorrect or undesired match points (outliers) to support the auto-decision of include or not an image observation in the estimation process is still an open problem. ORB-SLAM fisheye is applied in this study to estimate sensor trajectory based on dual-fisheye images acquired with Ricoh Theta S omnidirectional camera in a terrestrial mobile mapping system carried by a backpack. This preliminary study demonstrated the possible effects of image observation outliers in the sensor trajectory estimation (planimetric and altimetric accuracy of 0.381m and 0.26m, respectively). A proposal to combine semantic segmentation using CNN in the photogrammetric process workflow to cope with this problem and detect potential image observation outlier areas is presented.

Index Terms— fisheye images, image matching, Convolutional Neural Networks, ORB-SLAM fisheye

1. INTRODUCTION

Sensor orientation (or pose) is a fundamental task for autonomous platform navigation. Nowadays, imaging sensors can automatically estimate their position in an arbitrary reference system without a prior location or exterior data source. The simultaneous estimation of sensor poses and a 3D sparse point cloud, based on multiple images of a scene acquired during sensor motion, can be achieved using similar techniques developed by the Photogrammetry, Computer Vision and Robotics communities, such as Simultaneous Localization and Mapping (SLAM) [1] and Structure from Motion (SfM) [2]. Sensor orientation approaches based only on images require a strong network configuration of image observations to achieve an accurate solution, which relies on high accurate image measurements and a suitable geometric distribution of homologous points in the overlapping images. Therefore, the main challenge is related to the automatic detection of image observations to support a consistent network of matches and the auto-decision of whether including or not an observation in the sensor orientation estimation process. The presence of outliers and the weak geometry of match points have been mentioned in related works as the main reasons for the drift of the sensor trajectory solution with SLAM or SfM approaches [3].

The ability to detect outliers in SLAM and SfM methods is still an open problem. The failure in outlier detection can lead to a wrong sensor pose solution, consequently reducing these methods' capability to recover the subsequent sensor position. Outliers can be avoided or/and rejected from the set of matches using different methods, such as RANSAC [3-4], epipolar restriction [5], and match point geometry [6], for instance, as a function of the reprojection error or the number of intersection rays. However, these methods can be time-consuming and computationally intensive. Recently, deep learning methods have proven to be a powerful tool for image classification, segmentation, and detection, which can be integrated with the image matching process to support the auto-decision of include or not an image observation depending on the image region.

A preliminary study on the use of SLAM methods (ORB-SLAM fisheye), followed by an example of Convolutional Neural Networks (CNN) is presented. CNN can be applied as an alternative to detecting image observations outliers in fisheye images during the image matching process.

2. MATERIALS AND METHODS

This study was performed using a dataset obtained with a terrestrial mobile mapping system (MMS) [7], composed of an omnidirectional camera (dual-fisheye lenses, Figure 1) embedded in a backpack (Section 2.1). The MMS trajectory was estimated with the state-of-art ORB-SLAM [8] method adapted for fisheye lenses [9], which is named in this work as ORB-SLAM fisheye.

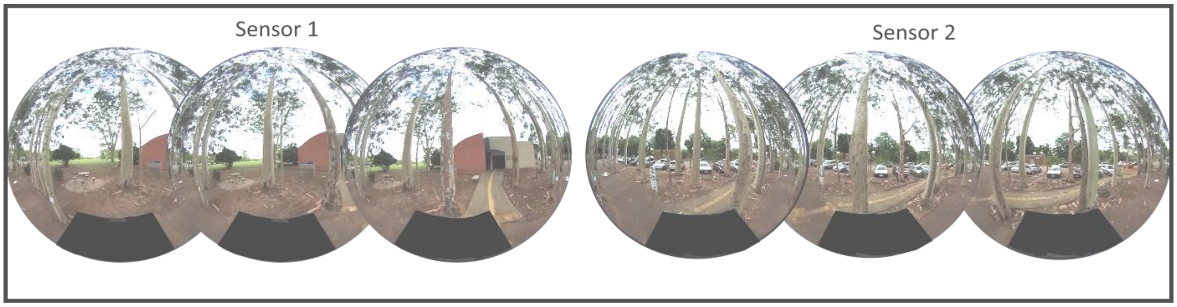


Figure 1. Image dataset: a sequence of dual-fisheye images acquired by Ricoh Theta S embedded in a backpack.

2.1. Dataset

Omnidirectional systems have been widely used in SLAM approaches. The use of an omnidirectional camera (Ricoh Theta S) in this work was encouraged by its 360° coverage around the sensor, which allows more features to be tracked in a single image shot. Recent works have also shown that larger FoV and larger keyframe overlap of fisheye images can improve the solution's accuracy, even if fewer keyframes are used [10]. The Ricoh Theta camera is composed of two sensors that record 190° FoV each. The Ricoh Theta S camera was embedded in a backpack mobile platform with sequential data acquisition of 29 frames per second. This terrestrial mobile platform is also composed of a single frequency GPS receiver (Ublox NEO-6M) that provides the platform position (latitude, longitude, and altitude). The sequential data acquisition with the mobile platform was performed along a trajectory of 10 m. The area of study is covered with sparse vegetation areas (Eucalyptus trees and ground vegetation) and urban features (small building, light poles, and traffic signs), as presented in Figure 1. The pixel size in object space units in the images dataset range from 1 to 80 cm. A set of 28 fisheye images (14 images of each sensor) was selected for the experimental assessments from the complete dataset.

2.2. ORB-SLAM fisheye framework

ORB-SLAM fisheye is applied in this study to estimate sensor trajectory based on dual-fisheye images acquired with Ricoh Theta S omnidirectional camera. The ORB-SLAM method [8] has been recently adapted for fisheye images [9] using a generic model for fisheye lenses named Enhanced Unified Camera Model (EUCM) [11]. The data processing using ORB-SLAM method was performed in two steps: camera calibration and trajectory estimation. Each sensor composing the Ricoh Theta S camera was previously calibrated independently, using the technique and software released by Khomutenko et al. [11]. Thus, the interior orientation parameters estimated in the camera calibration process are compatible with the EUCM model. In the second step, three initial frames were used in the process to enable a consistent ORB-SLAM initialization. The match points were automatically detected using the feature-based matching (FBM) operator ORB [9]. After initialization, the number of frames increased from one frame per second to five frames per second for sensor 1 (55 images in this sequence) and eight frames per second for sensor 2 (105 images), aiming an acceptable solution with ORB-SLAM fisheye for this preliminary study. Each sensor's set of images was used in the monocular estimation process separately, obtaining a trajectory and a sparse point cloud for each sensor.

The EOPs were estimated in a local reference system. 3D Helmert transformation parameters were calculated to convert the trajectory coordinates from the local reference system, defined by ORB-SLAM, to global coordinates (in this case, Universal transverse Mercator - UTM), aiming comparative analyses with a ground reference. Ground control points (using σ = 0.10 m as weight) and the Ublox GPS receiver trajectory (using σ = 2 m as the weight for these points) were used to compute the parameters considering different weights with least squares method. The similarity transformation parameters were estimated considering a total of 9 GCPs and 13 Ublox GPS receiver trajectory positions for sensor 1 and a set of 4 GCPs and 11 Ublox GPS receiver trajectory positions for sensor 2.

2.3. Semantic segmentation using CNN

Finding suitable observation matches with FBM operator, such as ORB, in omnidirectional images with a consistent geometric distribution for the SLAM sensor trajectory estimation can be considered a difficult task, especially in outdoor mobile mapping applications. In general, urban environments are uncontrolled environment, in which many image features result in undesired observations. Some match points can be detected in moving objects or in the sky, producing errors in the estimated SLAM trajectory. These complex outliers are difficult to detect using the most common outlier filters based on RANSAC algorithm, epipolar restriction and match point geometry, with a reasonable computational cost. This limitation open room to studies focusing on the use of deep learning semantic segmentation (e.g. CNN) as an alternative to detect moving objects, as well as other areas in which keypoints should be avoided, such as sky and clouds.

In this study, we apply the CNN of semantic segmentation DeepLab [14] to label these undesired pixels.

These labels make it possible to infer if a match point is an outlier candidate or not, considering the labelled class. This information enables remove outliers according to the pixel classification, such as sky class, before the image matching process, improving and accelerating the image matching process and consequently the estimation SLAM trajectory. Furthermore, the semantic information can be assigned to the point cloud for future segmentation and 3D classification.

3. RESULTS

In this section, the effects of the image observations distribution in ORB-SLAM fisheye solution for outdoor scenes (Section 3.1 and 3.2) will be analyzed. Additionally, the use of deep learning methods, such as CNN, as an alternative to label regions and optimize the detection of outliers in image observation by FBM operators, e.g., ORB (Section 3.3) will be assessed.

3.1 Sensor trajectory estimation

We assessed the estimated camera positions (trajectory) with ORB-SLAM fisheye by comparing the results with a reference trajectory. This reference was obtained from a consistent bundle block adjustment performed with the Agisoft Metashape with the same set of dual-fisheye images and GCPs well distributed along the trajectory. The reprojection error of this solution was less than 1 pixel, and the RMSE of the GCPs was 0.096 m. Figure 2 presents the reference trajectory (green) and the trajectories estimated with ORB-SLAM fisheye for sensor 1 (orange) and sensor 2 (blue). The ORB-SLAM fisheye method resulted in planimetric errors of 0.382 m for sensor 1 (0.34 m and 0.16 m for E and N) and 0.380 m for sensor 2 (0.18 m and 0.29 m for E and N). Altimetric errors were higher with sensor 1 (0.35 m) than sensor 2 (0.17 m).

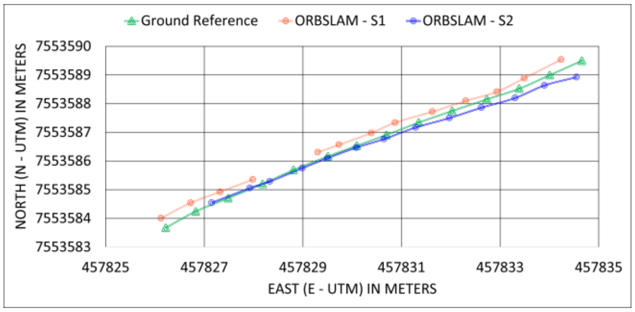
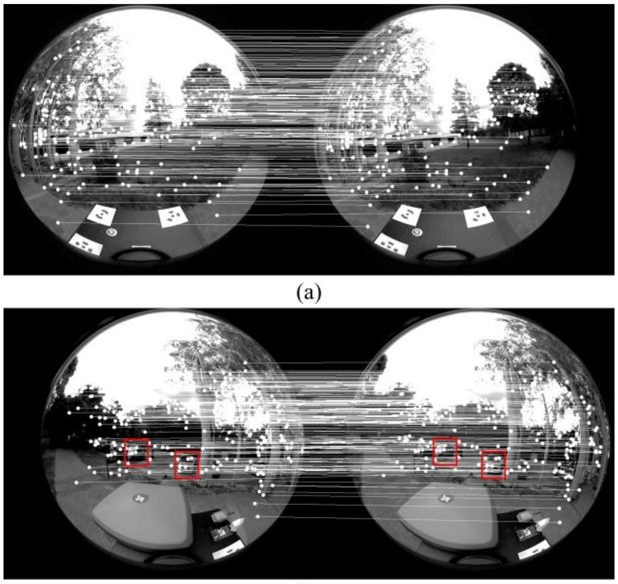


Figure 2. Sensor trajectory estimated with ORB-SLAM fisheye for sensor 1 (orange) and sensor 2 (blue), compared to the reference sensor trajectory (green).

Usually, ORB-SLAM method can achieve an accuracy of around 5 to 15 cm in indoor scenes [9]. However, outdoor mobile mapping applications can be more challenging, especially due to the huge depth and scale variation in the scenes. The sensor trajectory estimation, presented in Figure 2, is compatible with the results obtained in related works using only fisheye images acquired with low-cost cameras [12]. However, this result could be improved if a better set of image observations from the image matching process was considered as input to the sensor pose estimation.

The drift of the SLAM solution along the trajectory can be mainly attributed to outliers' occurrence and match points' weak geometry. For instance, the lack of suitable matching points in the first images of the trajectory using ORB operator delayed the initialization of the trajectory estimation and affected ORB-SLAM fisheye solution, mainly when estimating the initial sensor pose for sensor 2.

Another main problem to be mentioned is the drift of the solution between sensors. The positional discrepancies between sensor 1 and sensor 2 using ORB-SLAM fisheye at the end of the trajectory were 0.314 m and -0.613 m for E and N, respectively, which was not compatible with the physical reality between Ricoh Theta S camera perspective centres (offset of 0.019 m [13]). This problem can be related to the effect of the distribution of image observations in each image to estimate the trajectory. The features detected by ORB in sensor 2 (Figure 3.b) were weakly distributed when compared to sensor 1 (Figure 3.a). Furthermore, more moving objects, such as car and people, were observed in the images acquired with sensor 2. These moving objects in the scene, mainly those with slow speed, are often detected in the image matching process and accepted as suitable observations, resulting in low residuals and affecting the whole solution.



(a)

(b)

Figure 3. Examples of match points detect by ORB used as image observation in the ORB-SLAM fisheye solution for (a) sensor 1 and (b) sensor 2.

3.2. Outliers Detection

A preliminary assessment was performed with Tensorflow, applying DeepLab [14] with a trained model using the

Cityscapes database for urban perspective images. According to previously defined classes, semantic segmentation using CNN enables the pixel labelling of such features in the images. Figure 4 shows an example of the semantic segmentation applied to the fisheye image of sensor 2. The highlighted part in black indicates the location of the platform that was excluded from the solution. It can be seen that even using a model originally trained with perspective images, cars, people, and sky were correctly detected, with respective segments being depicted in blue, dark pink, and light blue.

Due to the large distortions in fisheye images, the model trained for images captured by perspective cameras is not optimal to be extended to fisheye images. Therefore, a network model for fisheye images should be trained to improve the network accuracy. Currently, there is no dataset of fisheye images for outdoor urban environments because the acquisition and processing of such dataset are expensive and laborious. Therefore, our future goal is to conduct network training transforming the images from the Cityscapes database into fisheye images, using the strategy defined by Ye [15].

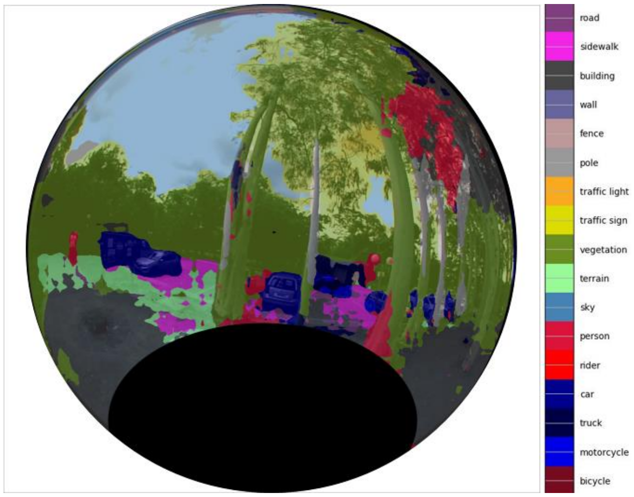


Figure 4. Example of semantic segmentation applied to a fisheye image of sensor 2.

4. CONCLUSIONS

This study discussed the feasibility of integrating deep learning techniques as a labelling process to constraint photogrammetric image matching for mobile mapping and earth observation applications. Recently developed deep learning techniques can improve the solution of some photogrammetric problems, such as the automatic detection of incorrect or undesired match points during the sensor pose estimation. In future works, the combination of ORB-SLAM fisheye method with a semantic segmentation (e.g. CNN) will be integrated and assessed to show the improvements in the accuracy and robustness of ORB-SLAM fisheye for outdoor mobile mapping applications.

5. REFERENCES

[1] H. Durrant-Whyte, T. Bailey, Simultaneous localization and mapping: part I. *IEEE robotics & automation magazine*, vol. 13, no. 2, p. 99-110, 2006

[2] S. I. Granshaw, Structure from motion: origins and originality. *The Photogrammetric Record*, vol. 33, no. 161, pp.6-10, 2018

[3] D. Scaramuzza, F. Fraundorfer, R. Siegwart, Real-time monocular visual odometry for on-road vehicles with 1-point ransac. *In Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, Kobe, Japan, pp. 4293-4299, 2009.

[4] R. C. Bolles, M. A. Fischler, A RANSAC-Based Approach to Model Fitting and Its Application to Finding Cylinders in Range Data. *In Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, University of British Columbia, Vancouver, pp. 637-6431981.

[5] D.Valiente, A. Gil, Ó. Reinoso, M. Juliá, M. A. Holloway, Improved omnidirectional odometry for a view-based mapping approach. *Sensors*, vol. 17, no. 2, pp. 325, 2017

[6] R. Lukierski, S. Leutenegger, A. J. Davison, Rapid free-space mapping from a single omnidirectional camera. *In Proceedings of European Conference on Mobile Robots (ECMR)*, IEEE, Lincoln, pp. 1-8, 2015.

[7] M. B. Campos, A. M. G. Tommaselli, E. Honkavaara, F. D.S. Prol, H. Kaartinen, A. El Issaoui, T. Hakala, A Backpack-Mounted Omnidirectional Camera with Off-the-Shelf Navigation Sensors for Mobile Terrestrial Mapping: Development and Forest Application. *Sensors*, vol. 18, no. 827, pp. 1-18, 2018

[8] R. Mur-Artal, J. M. M. Montiel, J. D. Tardos, ORB-SLAM: A versatile and accurate monocular SLAMsystem. *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147-1163, 2015

[9] S. Liu, P. Guo, L. Feng, A. Yang, Accurate and Robust Monocular SLAM with Omnidirectional Cameras. *Sensors*, vol.19, no.20, pp. 4494, 2019

[10] H. Matsuki, L.Von Stumberg, V. Usenko, J. Stückler, D. Cremers. Omnidirectional DSO: Direct sparse odometry with fisheye cameras. *IEEE Robotics and Automation Letters*, vol.3, no.4, pp. 3693-3700, 2018

[11] B. Khomutenko, G. Garcia, P. Martinet, An enhanced unified camera model. *IEEE Robotics and Automation Letters*, vol. 1, no. 1, pp.137-144, 2016

[12] X. Chen, W. Hu, L. Zhang, Z. Shi, M. Li, Integration of Low-Cost GNSS and Monocular Cameras for Simultaneous Localization and Mapping. *Sensors*, vol.18, no. 7, pp. 2193-2210, 2018.

[13] M. B. Campos, A. M.G Tommaselli, J. Marcato-Junior, E. Honkavaara, Geometric model and assessment of a dual-fisheye imaging system. *The Photogrammetric Record*, vol. 33, no.162, pp. 243-263, 2018.

[14] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp.834-848, 2017.

[15] Y. Ye, K. Yang, K. Xiang, J. Wang, K. Wang, Universal Semantic Segmentation for Fisheye Urban Driving Images. *In Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, Toronto, Canada, pp. 648-655, 2020