Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

A Robust Stereo Camera Localization Method with Prior LiDAR Map Constrains

Dong Han[1,2], Zuhao Zou[1], Lujia Wang[1] and Cheng-Zhong Xu[3]

***Abstract*— In complex environments, low-cost and robust localization is a challenging problem. For example, in a GPS-denied environment, LiDAR can provide accurate position information, but the cost is high. In general, visual SLAM based localization methods become unreliable when the sunlight changes greatly. Therefore, inexpensive and reliable methods are required. In this paper, we propose a stereo visual localization method based on the prior LiDAR map. Different from the conventional visual localization system, we design a novel visual optimization model by matching planar information between the LiDAR map and visual image. Bundle adjustment is built by using coplanarity constraints. To solve the optimization problem, we use a graph-based optimization algorithm and a local window optimization method. Finally, we estimate a full six degrees of freedom (DOF) pose without scale drift. To validate the efficiency, the proposed method has been tested on the KITTI dataset. The results show that our method is more robust and accurate than the state-of-art ORB-SLAM2.**

***Index Terms*— Global localization, point cloud, sensor fusion, stereo vision, SLAM**

I. Introduction

High-precision localization is necessary for autonomous vehicles. In past research, many kinds of sensors have been adopted for localization in a complex environment. The LiDAR is often considered the most reliable sensor in mapping and localization due to its accurate range measurements. However, its high cost is an obstacle for applications. On the other hand, GPS can perform well in the intense signal area, but it may fail to provide accurate localization when in urban areas and indoor environment. Cameras have been proposed as a substitute for LiDARs because of its low cost, small size, and ability to get color information. However, a monocular camera suffers from a fatal weakness, scale uncertainty , which can cause angular drift. Stereo camera systems overcome the problem of scale uncertainty. However, their accuracy and robustness still do not catch up with those of LiDARs.

[1] Dong Han, Zuhao Zou and Lujia Wang are with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. {dong.han, zh.zou, lj.wang1}@siat.ac.cn
[2] Dong Han is also with the University of Chinese Academy of Sciences.
[3] Cheng-Zhong Xu is with the University of Macau. czxu@um.edu.mo
* This research is supported by the Shenzhen Science and Technology Innovation Commission (Grant Number JCYJ20170818535187 89), the Guangdong Science and Technology Plan Guangdong-Hong Kong Cooperation Innovation Platform (Grant Number 2018B050502009) and the National Natural Science Foundation of China (Grant Number 61603376) awarded to Dr. Lujia Wang.

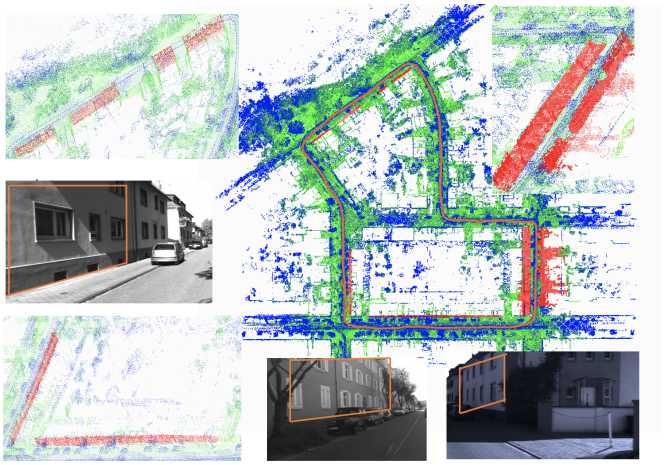

Fig. 1: LiDAR map is produced by G-loam(GPS based Loam) using a Velodyne HDL-64E LiDAR. As shown on the LiDAR map, there are at least six planes; the green line marks plane information in the visual image and the red point is the plane in the LiDAR map. The the red trajectory shows the camera position in the point cloud map.

Different situations have different requirements of precision and system cost [1]. Mapping requires high-precision sensors, in this work, it is less sensitive to price because the device can be reused. For autonomous vehicles localization, as the number of vehicles increases, the number of sensors also increases. Considering accuracy and cost, high-precision and low-cost sensors are needed, LiDAR is not considered at this time because of its price. An effective method to decrease cost and maintain precision is to combine the advantages of cameras and LiDARs. Generally, there are two ways to fuse LiDARs and cameras [2]. The first is to synthesize images from the LiDAR map [3], [4]. This method requires solving the relative extrinsic parameters of the camera and LiDAR, but the computation is heavy for registering images to the point cloud. Therefore, it is not suitable for a localization system, which has a strong requirement of real-time performance. The other way is to make landmarks both from the LiDAR map and visual image [5], [6]. In particular lane lines are the most common landmark to aid visual localization. However, this method only satisfies some special scenarios. Moreover, the lane line extraction is also a difficult problem. So, the above mentioned methods can not work well in the

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

complex environment. These methods also need the LiDAR and camera to be calibrated, and the calibration result can affect the accuracy of the localization result.

To avoid sensor calibration, we propose a robust localization method, which only needs a stereo camera. Unlike fusing the point cloud and image directly, we extract geometric features from a prior LiDAR map generated by some algorithms like G-LOAM [7], then pick out points with the same geometric properties in the visual image. These points satisfy bundle adjustment(BA) constraints as well as satisfy geometric constraints. Overall, the contributions of this paper are as follows:

- We propose an accurate and robust stereo visual localization method, which only relies on a camera and a prior LiDAR map.
- We design a new visual optimization model based on bundle adjustment.
- We propose a new framework for fusing camera and LiDAR, which greatly reduces the dependence on the LiDAR.

The rest of the paper is organized as follows: We discuss related work in Section II , describe our method in Section III, and present the experimental results in Section IV. Conclusions are given in Section V.

II. Related Works

Camera- [8], [9] and LiDAR- [10] based methods are common for mobile robot localization. Vision sensors capture the appearance of the surrounding environment, while LiDARs provide accurate range information and are mostly invariant to lighting conditions. In the past decades, great progress has been made in both vision-based localization and LiDAR-based localization.

For visual localization, there are two main approaches, feature-based methods [11], [12] and direct methods [13], [14], [15]. Among feature-based methods, ORB-SLAM2 [16] is a classical framework. In this method, ORB features are extracted from the image and BA is used, which minimizes the reprojection error over multiple image frames, to solve optimization problems. Direct methods, on the other hand, use an optical flow model to track motion points. In [17], the authors proposed to minimize the photometric error by using a sliding window. For LiDAR localization, the most common approach is to rely on the intensity information of the surrounding environment, and with the help of lanes to obtain the position information [18]. Meanwhile, [19] proposes a generic probabilistic method for localization. This algorithm uses Gaussians to model the world, which stores the z-height and intensity distribution of the environment.

Because the traditional vision-based methods fails to meet the accuracy requirements of localization, while LiDAR-based method have a high cost map-based visual localization has been an active field of research in recent years. [20] proposes to use a single monocular camera within a 3D prior ground-map. The map is generated by LiDAR and the height information is removed. The novelty of this work is using a GPU to generate several synthetic views from different poses, then calculating the normalized mutual information between the real camera measurements and these synthetic views, and finally finding the maximized NMI of the synthetic view. This view is the pose we need. [21] presents a monocular vision-based approach for localization in urban environments using road markings. First, a random forest-based edge detector is employed to detect road lanes. Then the Chamfer distance is computed between the detected edges and the projected road marking points in the image space. Epipolar geometry constraints and odometry are taken into account to formulate a non-linear optimization problem to estimate the six-DoF camera pose. In [22], SGBM is used for estimating disparity, and recover the depth from stereo images. Depth from the stereo camera is matched with the prior LiDAR map. A full six degree of freedom camera pose is estimated via minimizing the depth residual. This method is based on ORB-SLAM2 which adds depth constraint when tracking. We are also interested in cloud robotic systems [23], [24], [25], [26] and we will apply our work to cloud robots in the future.

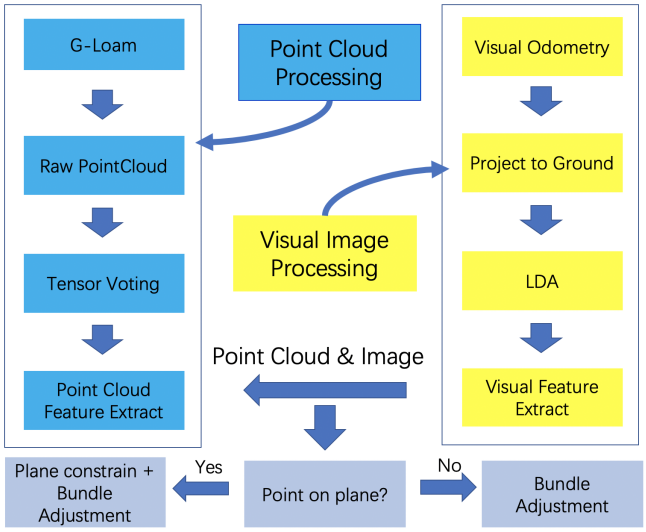

Fig. 2: The system structure, which includes three parts. The blue box shows the framework of point cloud processing, the yellow box shows the framework of visual image processing, and the purple box shows the framework of visual localization.

III. METHOD

A. Problem Definition

In this paper, the camera frame is denoted as $\mathbf{F}_{camera}$ and the prior map from the LiDAR as $\mathbf{M}_{LiDAR}$. We represent

the pose of the camera as $\mathbf{T} \in \mathbf{SE(3)}$, which transforms a point $\mathbf{p} \in R^3$ in the current frame to world coordinate . For the sake of convenience in the computation, we map $\mathbf{T}$ to $\xi \in \mathbf{se(3)}$ by using operators $\mathbf{Log}(.)$. $\mathbf{R} \in \mathbf{SO(3)}$ and $\mathbf{t} \in \mathbf{R}^3$ represent the rotation matrix and translation vector respectively. We use a stereo camera model, and the intrinsic of the camera $\mathbf{K}$ is given. The problem is: input $\mathbf{M}_{LiDAR}$ and $\mathbf{T}_{camera}$ and output a more accurate $\mathbf{T}_{current}$.

B. System Overview

The framework of our method contains three parts, a stereo online visual localization, and an offline point cloud map processing method. The high-accuracy prior map is generated by using GPS and LiDAR. In the framework of offline point cloud map processing, we use Tensor Voting to extract plane features form the map $\mathbf{M}_{LiDAR}$ and use K-means to get the normal of the surface.

As stereo images input, we can get the depth information by using triangulation measurement. So, our system first initializes visual localization with a frame of the picture. After initialization, the ORB-SLAM2 tracking thread is employed in visual localization. During back- end optimization, ORB-SLAM2 mainly uses bundle adjustment to minimize the reprojection error. Now, since the plane information has been obtained, the point on plane satisfied BA constrain, but plane constrains. As shown in Fig. 3, the all point fall into two categories, one class is not on the plane, because we use pinhole camera model, these points satisfy pinhole camera constraint, the other class point satisfy pinhole camera constraint and plane constrain.

C. Point Cloud Map Processing

Using LiDAR and GPS to generate high-precision maps, it is necessary to extract useful geometric information from the map to provide constraints for visual positioning. From observation, we find that the plane feature is a piece of beneficial information in the urban environment, and the plane is relatively easy to extract in point cloud maps.

The point cloud map processing is divided into two steps. The first step is to use the Tensor Voting [27], [28], [29] algorithm to solve the normal vector of each point in the point cloud. In the second step, the k-means algorithm is used to cluster points with the same normal, and the plane equation is solved according to the results of clustering.

Tensor Voting is an effective method for solving the surface normal. The essential idea is to extract implicit geometric features from a large amount of scattered point cloud data by transferring tensors between adjacent points. With the increase of distance, the influence coefficient of points in voting field decays gradually. According to this principle, we set the tensor kernel as

$$Decay(d, \sigma) = e^{-\frac{d}{\sigma^2}}, \tag{1}$$

where $d = (x_i - x)^2 + (y_i - y)^2$, (x, y) represents the coordinates of the votee point, and (x_i, y_i) represents the coordinates of the voter point, and σ is the kernel size of the sparse voting field. The input point $\mathbf{P}$ can be expressed by the second-order symmetric semi-positive definite tensor $\mathbf{T}$. Because of the equivalent relation between tensor and matrix, $\mathbf{T}_{3\times3}$ can be expressed by a positive semidefinite matrix. and be decomposed into three parts.

$\mathbf{T}_{3\times3}$ can be decomposed into the following forms:

$$\begin{aligned} T &= \lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T + \lambda_3 e_3 e_3^T && (2)\\ &= (\lambda_1 - \lambda_2) e_1 e_1^T && (3)\\ &+ (\lambda_2 - \lambda_3)(e_1 e_1^T + e_2 e_2^T) && (4)\\ &+ \lambda_3 (e_1 e_1^T + e_2 e_2^T + e_3 e_3^T). && (5) \end{aligned}$$

In (2), T is decomposed into three parts,(3) describes a stick, (4) describes a plate, and (5) describes a ball. $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$ is the eigenvalue of $\mathbf{T}$, $\mathbf{e_1}, \mathbf{e_2}, \mathbf{e_3}$ is corresponding eigenvectors.

According to the saliency of each tensor component, scattered point clouds are divided into three categories:

(1) if $(\lambda_1 - \lambda_2) \geq (\lambda_2 - \lambda_3)$ and $(\lambda_1 - \lambda_2) \geq \lambda_3$, the point belongs to the stick, and the orientation is e_1;
(2) if $(\lambda_2 - \lambda_3) \geq (\lambda_1 - \lambda_2)$ and $(\lambda_2 - \lambda_3) \geq \lambda_3$, the point belongs to the plate, and the orientation is e_3.
(3) if $(\lambda_3 >> (\lambda_1 - \lambda_2)$ and $\lambda_3 >> (\lambda_2 - \lambda_3)$, the point belongs to the ball, and the orientation is uncertain.

After the tensor voting framework, we can get the normal of each point. From (2), we can find the points on the wall conform to the following characteristics:

$$\mathbf{T_p} = (\lambda_2 - \lambda_3)(\mathbf{e_1 e_1^T} + \mathbf{e_2 e_2^T}). \tag{6}$$

K-means algorithm is employed for clustering point clouds. The surface normal can be easily obtained, and then the plane equation can be acquired.

D. Visual Image Processing

For point cloud data generated by the camera, due to the sparsity of the point cloud, we are searching for line information in two-dimensional space. We have experimented with a variety of methods for line detection. The first method uses the Hough transform for line detection. However, the Hough transform depends on parameter adjustment and is not suitable for the complex and changeable environments. Another method is to use LSD [30] for line detection. The algorithm does not depend on parameter changes, and the detection speed is faster than the Hough transform [31].

As we know, linear detection mainly depends on detecting the pixels with large gradient changes, so LSD is mainly used to detect the local straight contours in images, in which there are sharp changes from black to white or from white to black. Firstly, the image gradient is calculated, then the gradient of each pixel is sorted, and the gradient of points

is used for local region growth. Finally, the similar gradient points are clustered, and the straight-line part of the image map is obtained.

E. Visual Localization

Traditionally, visual localization is divided into three stages. Firstly, the ORB feature between two frames is matched to calculate the pose between two frames. Then, the depth information is restored by triangulation measurement. Finally, Bundle Adjustment is done according to the map point and corresponding frames [32], [33].

In this paper, our method is to add projection constraints for the points on the wall in addition to bundle adjustment when optimizing. In the initialization stage of localization, the original BA constraints are only relied on because there is too little valid information to be relied on. With the increase of input valid information, the number of optimization variables increases gradually, so the output pose is more accurate.

In the visual image processing framework, we can get the line feature of the local map. For the current observation points, the observed points are used to provide a prior constraint for the current point. This constraint function can be expressed by:

$$E = \sum_{k=1}^{K} \sum_{j \in F(k)} \lambda E_{ba} + \sum_{p \in \pi_i} (1-\lambda) E_{projection} \tag{7}$$

Here, we use the incremental sliding window method to connect the observed points with the current point. In each incoming frame, the target points in this frame are optimized simultaneously with the observed images to provide a maximum six-DOF pose. The most easily lost location information is at the corner, which corresponds to the end of each street. Therefore, at each corner, the sliding window will reach the maximum value, which provides more information for the constraints of the next frame, and also eliminate the problem of location drift.

1) No Additional Constraints: In optimization problems without additional constraints, bundle adjustment is mainly used to solve optimization problems. We take the Lie algebra corresponding to $\mathbf{R}$ and $\mathbf{t}$ as ξ. $\mathbf{p}$ represents the observed map point. The above formula represents the error caused by the observation of the k^{th} point in the j^{th} frame. The cost function is as follows:

$$E_{ba} = ||(u_i - Kexp(\xi^\wedge)p_k\ )^T Q_{k,j}^{-1}(u_i - Kexp(\xi^\wedge)p_k\ )||_2^2. \tag{8}$$

2) Additional Constraints: For the points on the wall, the sliding window algorithm is used to add plane constraints. In the first part of the algorithm, only the BA algorithm is used to constrain it. With the increase of input plate information, the points on the wall gradually increase, and the constraints of the points on the wall are added. Because of the increase in the points, the precise pose can be obtained. The plane

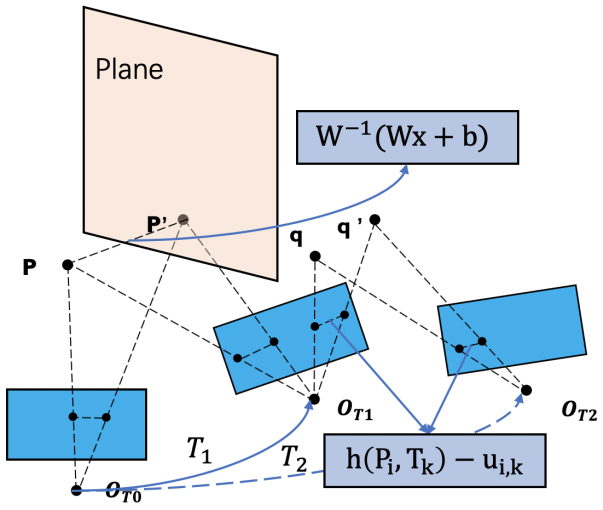

Fig. 3: Visual points can be divided into two categories, point $\mathbf{p}$ is on the plane and point $\mathbf{q}$ is not on the plane. In the visual localization framework, point $\mathbf{p}$ needs to minimize the reprojection error and coplanarity error and point $\mathbf{q}$ is required to minimize the reprojection error.

equation is $Wx + b = 0$, so the projection cost function is shown as follows:

$$E_{projection} = ||(\frac{1}{|W|}(Wp_j + b))^T R_j^{-1}(\frac{1}{|W|}(Wp_j + b))||_2^2. \tag{9}$$

To solve the optimization problem in visual localization, we adopt a graph-based method, which is generally used in solving the SLAM problem. We establish a graph-model based on an optimization problems, and in a graph, each edge represents different constraints. BA constraint, and plane constraints are increased to our system.

We let $\mathbf{P}$ be the node of visual point, and $\mathbf{X}$ represent the pose of frame, $\mathbf{P_L}$ is the point on plane which we extract form LiDAR map. We also define the error function $\mathbf{e_k}(\mathbf{p_k}, \mathbf{T})$ between $\mathbf{p_k}$ and observation point, we use a edge to represent it. $\mathbf{e_l}(\mathbf{p_k}, \mathbf{p_l})$ is defined as the cost function which project point to the plane, same as $\mathbf{e_k}(\mathbf{p_k}, \mathbf{T})$, we also use a edge to represent it. We can use (10) to describe optimization problem:

$$\begin{aligned} F(p) = &\sum_{k=1}^{K} \sum_{j \in F(k)} e_l(p_k, p_l)^T Q_{k,j}^{-1} e_l(p_k, p_l) \\ &+ \sum_{p \in \pi_i} e_l(p_k, p_l)^T R_j^{-1} e_l(p_k, p_l). \end{aligned} \tag{10}$$

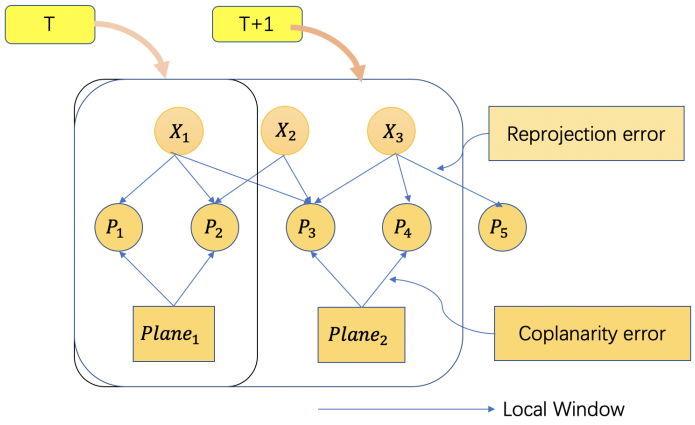

Fig. 4: Graph-structure optimization. $\mathbf{P}$ is the node of visual point, and $\mathbf{X}$ represent the pose of frame, $\mathbf{Plane}$ is the plane information extracted from LiDAR map.

After get (10),first we can expand it by using taylor expansion,

$$F(p+\delta p) = F(p) + J(p)\delta p, \tag{11}$$

$J(p)$ is Jacobian matrix of $F(p)$ which is a sparse matrix, We use Levenberg-Marquadt algorithm to solve the problem, our aim is to chose appropriate p and T to minimize error and make $F(p) = \alpha$. So the problem will turn into solve (10):

$$(J^T J + \lambda I)\delta x = -J^T \alpha. \tag{12}$$

To solve this graph optimization problem, Ceres is employed to solve the equation, which is an open-source C++ library for modeling and solving large, complicated optimization problems.

IV. Experiment

We test our algorithm on the KITTI dataset and compare our framework with ORB-SLAM2. KITTI-07 is an outdoor image sequence that includes 1101 stereo images, and this dataset is based on the urban scene. The experiment is divided into two parts, which include mapping and localization.

Our localization algorithm is based on the prior map generated by LiDAR and GPS. So the first step is to acquire a high-precision map. The KITTI odometry dataset provides a sequence of Velodyne HDL-64E LiDAR scans; we use this dataset and G-Loam mapping algorithm to produce a 3D LiDAR Map. The GPS-INS system provides the ground truth of the camera pose. As shown in Fig. 1, the red line is the trajectory of ground true.

For localization, we test ORB-SLAM2 on the same dataset, and because we only test the localization model, the loop closing function in ORB-SLAM2 is closed.

As shown in Fig. 5. Our method is more accurate than ORB-SLAM2 localization, especially in the corner of the

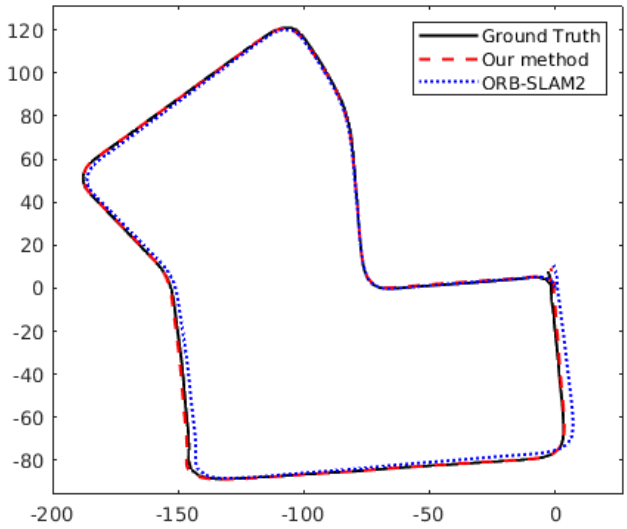

Fig. 5: Camera trajectory on the KITTI-07 dataset. The blue line is the result of ORB-SLAM2, the red line is that of our method, and the black line is the ground truth.

TABLE I: Translation error result from KITTI-07

Method	Our Method			ORBSLAM2		
	Mean	RMSE	Std	Mean	RMSE	Std
KITTI 07	0.53264	0.5643	0.2031	0.7535	0.8433	0.3786

street, ORB-SLAM2 easily lost the position and is not able to correct the pose; it will cause error accumulation. And our method is more robust, even if in the corner because we use the plane constraint to aid visual localization, it eliminates error.

In order to evaluate our algorithm, the statistics of absolute trajectory error(ATE) are computed, and ATE figure is drawn. We compute RMSE(root-mean-square error), Mean, Std(Standard Deviation), and each indicator is better than ORB-SLAM2. As shown in the table, our method is more robust than ORB-SLAM2. We calculate the error by this follow form $e(t) = \sqrt{e(t)_x^2 + e(t)_y^2}$. From the result, we can get the average error is 0.5326m, and the maximum error is 0.9664m, minimum error is 0.1884. About ORB-SLAM2, the average error is 0.8060, and the maximum error is 2.0300m, minimum error is 0.1276m.

V. Conclution And Future Work

In this paper, we propose a stereo visual localization method based on the prior LiDAR map. We only use a stereo camera to acquire decimeter-level localization. Compare with localization based on LiDAR, we cost low and reach the same level of precision. Our method performs robust in a complex environment, which can provide the accurate estimation of a

six-DoF pose in urban without GPS signal. It can also work well in a sunlight change environment and without scale drift.

To combine the advantages of LiDAR and camera and cut costs, we design a novel visual optimization model by matching planar information between LiDAR map and visual image. To achieve the real-time and robust localization, LDA and Tensor Voting are employed to extract geometric features in visual image and point cloud map, respectively. We use the coplanarity constrains to build bundle adjustment and solve it using graph-based optimization algorithm a local window optimization method.

In the experiment part, we test our approach in the KITTI urban environment. The result shows our method is more robust and accurate than ORB-SLAM2. This result proves that our method has a great advantage in this environment. In the future, we will try to extract more geometry features from the stereo image, increase line constrain to localization. Furthermore, we will implement evaluation in the long term localization application.

REFERENCES

[1] I. Deutsch, M. Liu, and R. Siegwart, "A framework for multi-robot pose graph slam," in *Real-time Computing and Robotics (RCAR) 2016 IEEE International Conference on*, (Angkor Wat, Cambodia), June 2016.

[2] M. Liu, L. Wang, and R. Siegwart, "DP-Fusion: A generic framework for online multi sensor recognition," in *IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, IEEE, 2012.

[3] B. Steder, M. Ruhnke, S. Grzonka, and W. Burgard, "Place recognition in 3d scans using a combination of bag of words and point feature based relative pose estimation," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1249–1255, IEEE, 2011.

[4] G. Pascoe, W. Maddern, and P. Newman, "Direct visual localisation and calibration for road vehicles in changing city environments," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 9–16, 2015.

[5] R. Wang, "3d building modeling using images and lidar: A review," *International Journal of Image and Data Fusion*, vol. 4, no. 4, pp. 273–292, 2013.

[6] X. Ding, Y. Wang, D. Li, L. Tang, H. Yin, and R. Xiong, "Laser map aided visual inertial localization in changing environment," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4794–4801, IEEE, 2018.

[7] L. Zheng, Y. Zhu, B. Xue, M. Liu, and R. Fan, "Low-cost gps-aided lidar state estimation and map building," *arXiv preprint arXiv:1910.12731*, 2019.

[8] M. Liu, C. Pradalier, Q. Chen, and R. Siegwart, "A bearing-only 2d/3d-homing method under a visual servoing framework," in *2010 IEEE International Conference on Robotics and Automation*, pp. 4062–4067, May 2010.

[9] M. Liu, C. Pradalier, F. Pomerleau, and R. Siegwart, "Scale-only Visual Homing from an Omnidirectional Camera," in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, 2012.

[10] M. U. M. Bhutta and M. Liu, "Pcr-pro: 3d sparse and different scale point clouds registration and robust estimation of information matrix for pose graph slam,"

[11] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 1052–1067, 2007.

[12] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pp. 1–10, IEEE Computer Society, 2007.

[13] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *2011 international conference on computer vision*, pp. 2320–2327, IEEE, 2011.

[14] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*, pp. 834–849, Springer, 2014.

[15] C. Forster, M. Pizzoli, and D. Scaramuzza, "Svo: Fast semi-direct monocular visual odometry," in *2014 IEEE international conference on robotics and automation (ICRA)*, pp. 15–22, IEEE, 2014.

[16] R. Mur-Artal and J. D. Tardós, "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, 2017.

[17] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[18] A. Hata and D. Wolf, "Road marking detection using lidar reflective intensity data and its application to vehicle localization," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pp. 584–589, IEEE, 2014.

[19] R. W. Wolcott and R. M. Eustice, "Robust lidar localization using multiresolution gaussian mixture maps for autonomous driving," *The International Journal of Robotics Research*, vol. 36, no. 3, pp. 292–319, 2017.

[20] R. W. Wolcott and R. M. Eustice, "Visual localization within lidar maps for automated urban driving," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 176–183, IEEE, 2014.

[21] Y. Lu, J. Huang, Y.-T. Chen, and B. Heisele, "Monocular localization in urban environments using road markings," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, pp. 468–474, IEEE, 2017.

[22] Y. Kim, J. Jeong, and A. Kim, "Stereo camera localization in 3d lidar maps," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–9, IEEE, 2018.

[23] L. Wang, M. Liu, M. Q.-H. Meng, and R. Siegwart, "Towards real-time multi-sensor information retrieval in cloud robotic system," in *Proceedings of the IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, 2012.

[24] L. Wang, M. Liu, and M. Q. H. Meng, "A hierarchical auction-based mechanism for real-time resource allocation in cloud robotic systems," *IEEE Transactions on Cybernetics*, vol. 47, pp. 473–484, Feb 2017.

[25] B. Liu, L. Wang, and M. Liu, "Lifelong federated reinforcement learning: A learning architecture for navigation in cloud robotic systems," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4555–4562, 2019.

[26] L. Wang, M. Liu, and M. Q. H. Meng, "Real-time multisensor data retrieval for cloud robotic systems," *IEEE Transactions on Automation Science and Engineering*, vol. 12, pp. 507–518, April 2015.

[27] G. Medioni, C.-K. Tang, and M.-S. Lee, "Tensor voting: Theory and applications," in *Proceedings of RFIA*, vol. 2000, 2000.

[28] M. Liu, "Efficient segmentation and plane modeling of point-cloud for structured environment by normal clustering and tensor voting," in *2014 IEEE International Conference on Robotics and Biomimetics (ROBIO 2014)*, pp. 1805–1810, IEEE, 2014.

[29] M. Liu, F. Pomerleau, F. Colas, and R. Siegwart, "Normal Estimation for Pointcloud using GPU based Sparse Tensor Voting," in *IEEE Int. Conf. on Robotics and Biomimetics (ROBIO)*, 2012.

[30] R. G. Von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall, "Lsd: A fast line segment detector with a false detection control," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 4, pp. 722–732, 2008.

[31] D. H. Ballard, "Generalizing the hough transform to detect arbitrary shapes," *Pattern recognition*, vol. 13, no. 2, pp. 111–122, 1981.

[32] J. Zhang, L. Tai, P. Yun, Y. Xiong, M. Liu, J. Boedecker, and W. Burgard, "Vr-goggles for robots: Real-to-sim domain adaptation for visual control," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1148–1155, 2019.

[33] M. Liu, C. Pradalier, F. Pomerleau, and R. Siegwart, "The role of homing in visual topological navigation," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012.(IROS 2012)*, 2012.