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

LMVI-SLAM: Robust Low-Light Monocular Visual-Inertial Simultaneous Localization and Mapping

Luoying Hao[1,2,3,§], Hongjian Li[1,2,3,§] , Qieshi Zhang[2,3], Xiping Hu[2,3], Jun Cheng[2,3,*]

[1] Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Beijing, China
[2] Guangdong Provincial Key Laboratory of Robotics and Intelligent System, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[3] The Chinese University of Hong Kong, Hong Kong, China
{ly.hao, hj.li1, qs.zhang, xp.hu, jun.cheng}@siat.ac.cn

***Abstract*— Visual-inertial simultaneous localization and mapping (SLAM) shows significant progress in recent years due to the complementary nature of the visual and inertial sensor but still, challenges remain for low-light environments. Recent visual-inertial SLAMs often drift or even fail in low-light conditions due to insufficient 3D-2D correspondences for bundle adjustment. To address the issue, this paper performs image preprocessing firstly with a united image enhancement method involving adaptive gamma correction and contrast limited adaptive histogram equalization, which could ameliorate the brightness and contrast of the image greatly. Moreover, we track features using optical flow for adequate point correspondences in dim-light environments, and supplement the corresponding map points continually by insert keyframe and triangulation to keep tracking. Finally, we construct a tightly-coupled nonlinear optimization model, which combines a feature reprojection error on point correspondences and IMU measurement by pre-integration to constrain and compensate each other for more accurate pose estimation. We validate the performance of our algorithm on public dataset and real-word experiments with a mobile robot, including dark laboratory, etc., and compare against existing state-of-the-art visual-inertial algorithms. Experimental results indicate our algorithm outperforms other state-of-the-art SLAMs in accuracy and robustness, and works reliably well for both general and low-light environments.**

***Index Terms*— Low light, Sensor fusion, Visual-inertial SLAM**

I. Introduction

SLAM is undoubtedly the most indispensable module for a wide range of applications, such as unmanned aerial vehicles (UAVs), robots, autonomous driving, airborne navigation [1], and augmented reality (AR). Among different sensor modalities, visual-inertial setups provide a cheap solution with great potential [2]. Visual information and inertial information are highly complementary. The drift of inertia can be corrected by visual information, while inertial information can be used to locate in the environment where vision-based methods fail, such as fast motion, lack of texture or low illumination. Furthermore, inertial sensor can recover metric scale for monocular vision, rendering absolute pitch and roll observable, which is a limitation for monocular vision. Therefore, the fusion of camera and inertial sensor can greatly improve the robustness and accuracy of the algorithm.

The visual-inertial fusion approaches can be categorized to filter-based and optimization-based ones generally. A popular filter-based VIO is MSCKF [3], [4] which maintained several previous camera poses in the state vector in the form of sliding window, and made a feature located at several camera views, to establish multi constraint updates. Huai et al. [5] developed a lightweight, high-precision, robocentric visual-inertial odometry (R-VIO) algorithm. The key idea of the proposed approach is to deliberately reformulate the 3D visual-inertial navigation system (VINS) with respect to a moving local frame. While the filtering-based system have shown to exhibit effective state estimation, they theoretically suffer from a limitation, that is, nonlinear measurements must have a onetime linearization before processing, possibly introducing large linearization errors into the estimator and degrading performance [6].

Recent advances of pre-integration have also allowed for efficient inclusion of high-rate IMU measurements in graph optimization-based formulations [7]. In particular, Leutenegger et al. [1] introduced a keyframe-based optimization approach (OKVIS), whereby a set of non-sequential past camera poses and a series of recent inertial states, connected with inertial measurements, was used in nonlinear optimization for accurate trajectory estimation. Raul et al. [2] presented a tightly-coupled visual-inertial SLAM system (VIORB) that is able to close loops and reuse the map to achieve zero-drift localization in already mapped areas, achieving a typical scale factor error of 1% and centimeter precision. Qin et al. [8] recently presented an optimization-based monocular VINS that can incorporate loop closures in a non-real time thread.

§These two authors contributed equally.
*Jun cheng is the corresponding author.

This work was supported by National Key RD Program of China (2018YFB1308000), Key Research and Development Program of Guangdong Province [grant numbers 2019B090915001], National Natural Science Funds of China (U1813205) and Shenzhen Technology Project (JCYJ20180507182610734, JCYJ20170413152535587) CAS Key Technology Talent Program.

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

However, even with these most advanced methods, the robustness to illumination variations is still poor. Our experimental results show that the algorithms are prone to pose drift or tracking failure in dim-light scenes when using VINS-Mono [8], VIORB [2] and OKVIS [1] algorithms to test. For SLAM algorithm in low-light scenes, only few researches have been done so far. Soumyadip et al. [9] proposed edge based slam framework, which tracks edge point using optical flow for point correspondences and then uses a robust method for two-view initialization for bundle adjustment. But it has limited effect in the case of low illumination. [10], [11] proposed the visual odometry (VO) method based the combination of point and line, which improves the accuracy of pose estimation, but increases the computational cost greatly.

With the aim of robust and accurate visual-inertial algorithm, we advocate optimization-based and tightly-coupled fusion, rather than filter-based one in order to reduce suboptimality due to linearization. Our algorithm refers to [2], and the main contributions of this paper are summarized as three folds:

- Image preprocessing is carried out firstly for dim-light images. The effective enhancement algorithms, contrast limited adaptive histogram equalization (CLAHE) [12] and adaptive gamma correction [13] are combined to improve the brightness and contrast of images. Then we can extract enough features to keep track.
- Bi-directional sparse iterative and pyramidal version of the Lucas-Kanada optical flow is utilized to generate sufficient feature correspondences, and map points are added continually by inserting keyframes and triangulation at the same time. Then IMU error term is combined with the reprojection error based on feature correspondences and map points to construct nonlinear Optimal model for the incremental motion of the sensor with high accuracy and robustness.
- We validate the performance of our algorithm on public dataset and real-world experiments, including low-light yoga studio, laboratory and rest room, and compare with other state-of-the-art algorithms.

II. The Proposed Method

The pipeline of proposed monocular visual-inertial state estimator is shown in Fig. 1, including three threads: tracking, local mapping, and loop closing. We mainly modified the tracking thread by adding the image preprocessing module, optical flow tracking and map points supplementing algorithm.

A. Image Preprocessing in Low-light Conditions

For low-light conditions, it is difficult to get adequate visual information. By image preprocessing, some important structural information of images can be restored. In this

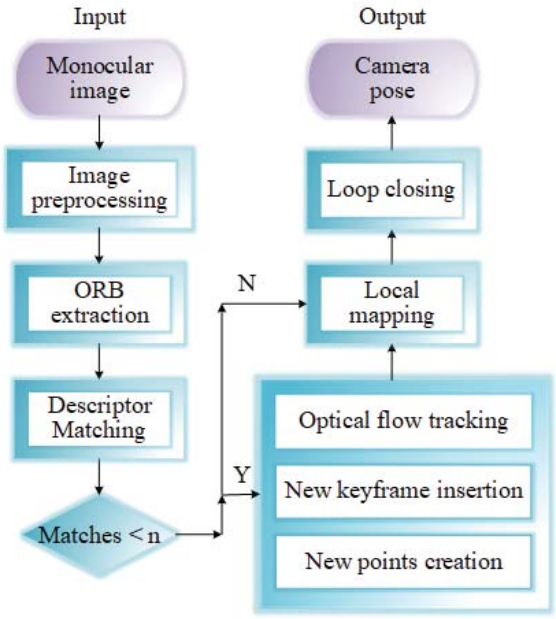


Fig. 1: The pipeline of our system.

paper, we apply gamma correction and CLAHE [12] to improve the quality of dim-light images. Gamma correction can improve the brightness of the images effectively by using a varying adaptive parameter γ. The simple form of the transform-based gamma correction is derived by

$$f(I) = I_{max} I^{\gamma} \tag{1}$$

Where I_{max} is the maximum intensity of the input image. The intensity I of each pixel in the input image is transformed as $f(I)$. In the common case of $I_{max} = 1$, inputs and outputs are normalized typically in the range $(0, 1)$. As expected, when $\gamma < 1$, the overall brightness of the image is reinforced and the contrast at low gray level is increased, which is more conducive to the resolution of image details at low gray level [13].

The histogram equalization (HE) can further improve the contrast of images, make the image structure and texture clearer, as shown in Fig. 2. Its basic idea is to transform the histogram of random distribution of the original image into a uniform distribution form, which means increase in the dynamic range of pixel gray [14].

For a given image $\mathbf{X}$, the pixel gray value is normalized and represented by $x(0 \leq x \leq 1)$. To make process image more easily, the normalized and discrete gray level x_k is introduced. $p_x(x_k)$ presents the probability density function of x_k and is defined as:

$$p_x(x_k) = \frac{n_k}{n} \tag{2}$$

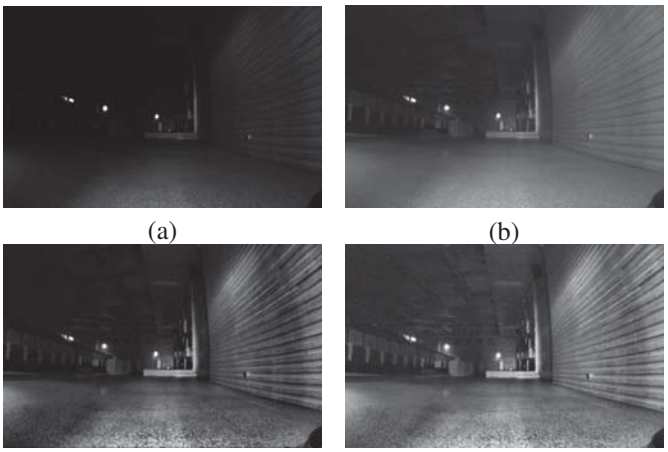


Fig. 2: The results of image preprocessing. (a) is the original image captured by our mobile robot in dark conference rooms. (b) is the result of gamma correction. (c) is the image processed by CLAHE [12]. (d) is the final image processed by the two methods above.

where $k = 0, 1, 2, ..., L-1$, n_k represents the number of times that the level x_k appears in the input image $\mathbf{X}$ and n is the total number of samples in the input image. Based on the probability density function, the cumulative density function is defined as

$$r_k = \sum_k^{i=0} p_x(x_k) \tag{3}$$

Note that r_k = 1 by definition. HE is a scheme that maps the input image into the entire dynamic range, (x_0, x_{L-1}), by using the cumulative density function as a transform function [12].

For adaptive histogram equalization (AHE) [12], it improves on HE by transforming each pixel with a transformation function derived from a neighbourhood region. Therefore, AHE is more suitable for enhancing the local contrast of the image and obtaining more image details. However, AHE has the problem of over-enlarging the noise in the same area of the image. Another adaptive histogram equalization algorithm , the CLAHE algorithm [12], can limit the enlargement of this disadvantage, which is primarily achieved by limiting the degree of contrast enhancement of AHE algorithm.

In Fig. 2, the experimental image is collected by our robot in the dark conference room. Fig. 2 (a) is the image processed by gamma correction which become brighter. Fig. 2 (b) shows the contrast of the image is enhanced compared with the original image (Fig. 2 (a)). In particular, Fig. 2 (d) is the result of the combined action of the above two methods. In Fig. 2 (d), most of structural information and texture is restored, and its effect is the most remarkable.

B. Motion Estimation in Low-light Conditions

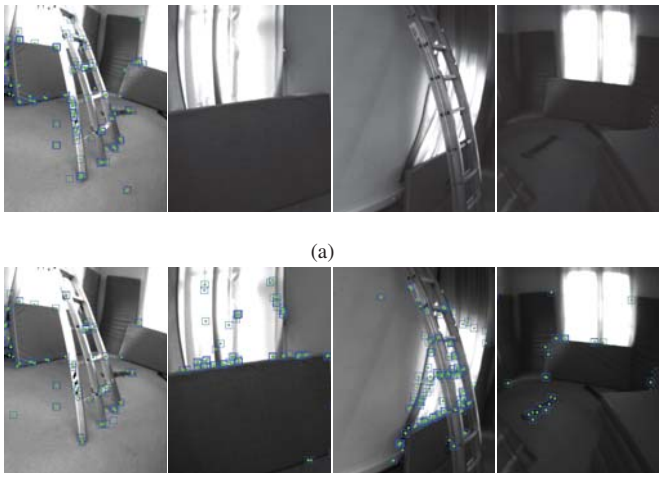
Track-loss is one of the major issues presented in the literature of visual SLAM where estimated camera track and reconstructed map breaks due to tracking failure during resection. This problem often arises in low-light environments due to the reduction in the number of 2D point correspondences and corresponding structure points reduction in bundle adjustment [9].

1) Optical flow tracking: For the most excellent feature-based SLAM algorithms at present, the tracking method predicts camera pose by features extraction, descriptor computing and features matching, which relies heavily on the number of extracted features and consumes a lot of computation [15]. Therefore the method can hardly work in the case of low light and less environmental texture due to lack of sufficient feature matches, which is easy to lead to tracking failure.

To address the issue, we estimate strong point correspondences of detected features using a bi-directional sparse iterative and pyramidal version of the Lucas-Kanade optical flow [16] from the sequential images. The point correspondences obtained by optical flow may only contain noise and tracking missing points and therefore can create drifts in the estimation of flow vectors. We remove those noisy correspondences using several filtering methods. We discard redundant pixels (pixels whose euclidean distance is very low) first and then remove the points having a higher bi-directional positional error [9]. With this approach, we can get sufficient feature matches to optimize pose estimation and keep tracking.

Fig. 3 indicates that our method with optical flow can deal with the problem of tracking loss in low-light condition better. It should be noted here that the features will only be displayed on the image when the algorithm tracking is successful. The four image of Fig. 3 (a) or (b) are four frames randomly extracted from a scene in V1_03_difficult sequence of euroc dataset [17]. In the scene of Fig. 3, the proposed algorithm achieved normal tracking when the environment was dim-light and the camera moved fast, while the original system failed to track immediately.

2) Adding map points: With the scene switching, although the feature matches are increased through optical flow, the reduction of map points leads to insufficient 3D-2D correspondences for the optimization model. So the corresponding map point of the current frame should be added to keep tracking. In our proposed method, when the number of map points fall below a threshold, current frame will be spawned as a new keyframe and matched with recent keyframe by bag of DBoW2 [18] to inherit the map points of the previous keyframe, as shown in Fig. 4. Furthermore, if the map points obtained by this measure still do not meet the demand, new map points will be created by triangulating features from connected keyframes in the covisibility graph [19]. In



(a)

(b)

Fig. 3: (a) is the tracking process of the original method in continuous scenes in V1_03_difficult sequence of euroc dataset [17]. (b) is the tracking process of the proposed method in the same scenarios. Features only appear on images when tracking is successful.

these ways, the optical flow tracking becomes more robust to challenging environment.

3) Pose optimization combining IMU: Pure visual tracking based on monocular camera can achieve pose estimation, but it sensitive to environmental condition and the scale of the camera cannot obtain. In the dim-light conditions, IMU can calculate position and rotation direction from the measurements of visual-independent angular velocity and acceleration by pre-integration, but they are affected by white noise and biases [7], which are denoted by η_a, η_g and b_a, b_g respectively. In our paper, we construct the tightly-coupled nonlinear optimization model to combine a feature reprojection error on point correspondences of optical flow and IMU measurement by pre-integration in order to constrain and compensate each other for more accurate pose estimation. The state of current frame j is optimized by linking the previous frame i [20]. The optimization function is as follows:

$$\arg\min_{\tau}\left(\sum\nolimits_k E_{proj}(k,j)+E_{IMU}(i,j)\right) \tag{4}$$

For given point correspondences k by optical flow tracking, the reprojection error is defined as follows [2]:

$$\begin{aligned} E_{proj}(k,j) &= \rho\left({M_k}^T \sum\nolimits_k M_k\right) \\ M_k &= \mathbf{x}_k - \pi\left(\mathbf{X}_k^C\right) = \mathbf{x}_k - \left(\frac{X_k f_u}{Z_k}+c_u, \frac{Y_k f_v}{Z_k}+c_v\right)^T \end{aligned} \tag{5}$$

where $\mathbf{x}_k$ is the position of the feature in image, $\sum_k$ is the information matrix, ρ is Huber robust cost function, and $\mathbf{X}_k^C = [X_k, Y_k, Z_k]^T \in \mathbb{R}^3$ is the 3D point in the camera coordinate system C. Furthermore, in the intrinsic parameters of a pinhole-camera mode used in our model, the focal length is denoted with $[f_u, f_v]^T$ and the principal point is denoted with $[c_u, c_v]^T$. The IMU residual model $E_{IMU}(i,j)$ refers to [7].

III. Experimental Result

To evaluate the proposed SLAM algorithm, our experiments are divided into two parts. In the first part, we compare the proposed method with other state-of-the-art methods on public dataset. We carry out a numerical analysis to show the accuracy and effectiveness of our algorithm. Then we test our algorithm in the low-light yoga studio, laboratory and rest area of our office building with the mobile robot, and compare with other excellent algorithm taking radar location data as ground truth.

A. Public dataset and Hardware Platform

Our experiments is carried out based on public dataset and the mobile robot platform primarily.

1) Public dataset: To test the validity of our proposed method, we employ the EuRoc dataset [17] to conduct the comparison experiment. The recent EuRoC dataset [17] contains 11 stereo sequences recorded from a micro aerial vehicle (MAV) flying around two different rooms and a large industrial environment [19]. The sequences are classified

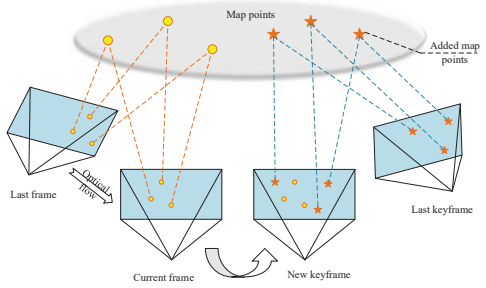
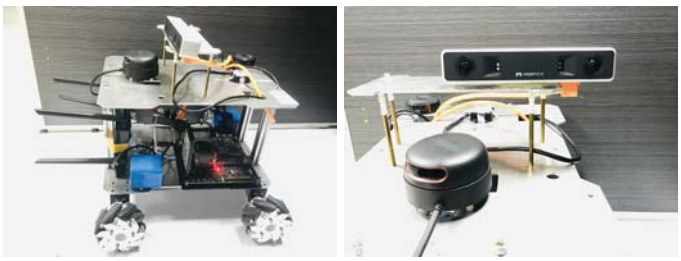
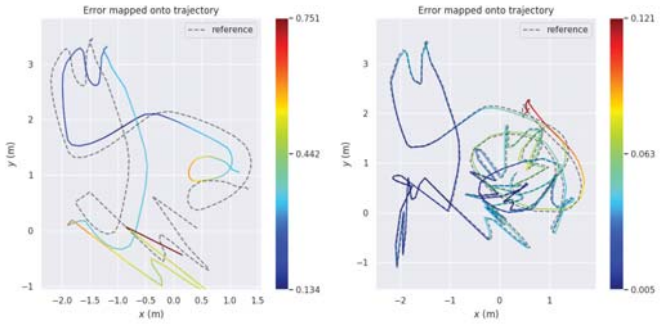


Fig. 4: Adding map points from keyframe.



(a) (b)

Fig. 5: Physical figure of mobile robot and camera.



(a) Trajectory of VI-ORB (b) Trajectory of our method

Fig. 6: Trajectory in V1_03_difficult, compared with VI-ORB.

as easy, medium, and difficult depending on MAVs speed, illumination, and scene texture [19].

2) Hardware platform and experimental scenes: The mobile robot is a highly integrated ROS robot with various hardware driving modules, including NVIDIA Jetson TX2 module, RPLIDAR-A2 360° laser scanning ranging radar, MYNTEYE binocular camera equipped with IMU and 752x480 resolution, STM32F1 large load driving board, etc. The physical figure of the mobile robot and binocular camera is shown in Fig. 5. We only utilise the left images of the camera in our experiments. Moreover, the calibration of camera intrinsic parameters, IMU intrinsic parameters and the joint calibration of IMU and camera have been completed in the experiment.

In the real-world experiments, we choose our nighttime and low-light yoga studio, laboratory and rest area as the experiment areas, as show in Fig. 7.

B. Experimental Results based on Public Dataset

All the experiments are carried out on a desktop, HP Z440 workstation, which is set up with Ubuntu 16.04 and ROS Kinetic. It owns a eight-core Intel Xeon CPU E5-1660 v3, operating at 3 GHz, and 32 GB of RAM.

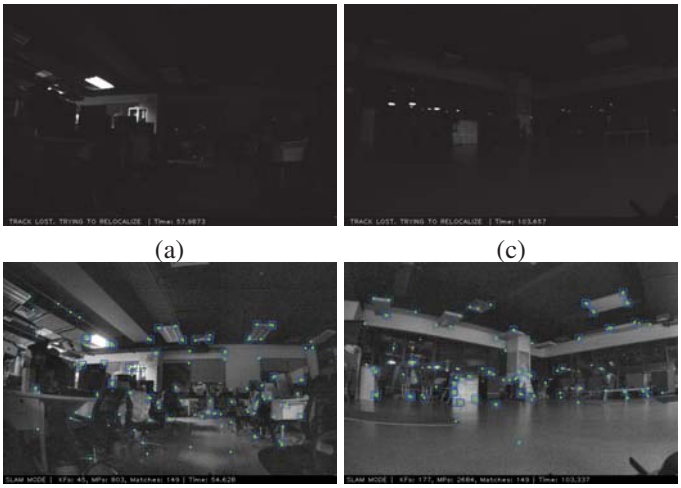
TABLE I: Root mean square error (RMSE) of different architectures (m).

Dataset	OKVIS [1]	VINS [8]	VI-ORB [2]	Ours
MH_01_easy	0.164	0.062	0.068	**0.029**
MH_02_easy	0.187	0.078	0.073	**0.030**
MH_03_medium	0.274	0.045	0.071	**0.035**
MH_04_difficult	0.375	0.134	0.087	**0.054**
MH_05_difficult	0.432	0.088	0.060	**0.052**
V1_01_easy	0.224	0.045	**0.023**	0.033
V1_02_medium	0.176	0.045	0.027	**0.009**
V1_03_difficult	0.193	0.088	X	**0.054**
V2_01_easy	0.176	0.057	0.018	**0.016**
V2_02_medium	0.187	0.114	**0.024**	0.028
V2_03_difficult	0.316	0.109	**0.047**	0.054

In this paper, our proposed method is compared with the state-of-the-art methods: OKVIS [1], VINS-mono [8] and VI-ORB [2]. For each trial, we performed sim3 trajectory alignment to the ground truth and then computed the root mean square error (RMSE) position error over the aligned trajectory. These results are shown in Table I, which indicate that our method outperforms the other approaches in most sequences. It should be noted that for V1_03_difficult sequence, our method could successfully complete tracking and obtain pose estimation results with low RMSE, while the original VIORB method failed to track. Fig. 6 are trajectory graphs drawn on VI-ORB and the method of this paper based on V1_03_difficult sequence, which also verify the robustness of our algorithm.

C. Experimental Results based on Mobile Robot

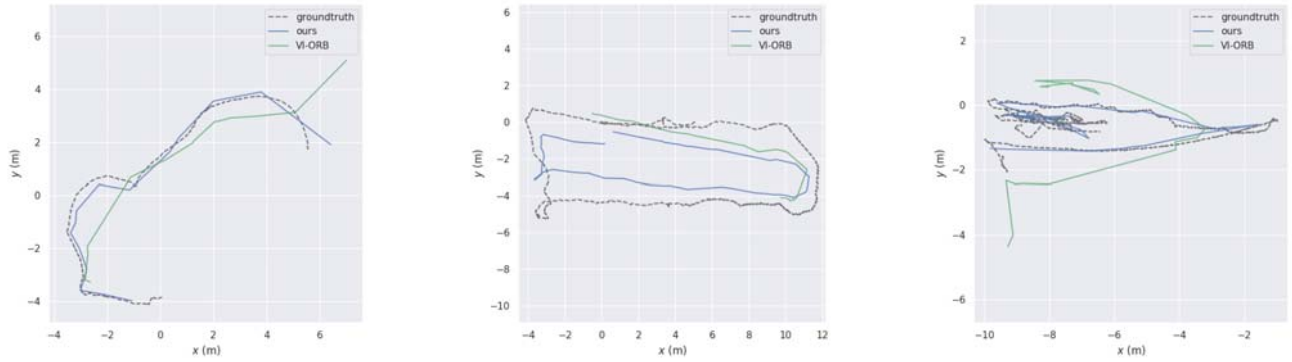
In the real-world experiment, in order to test robustness of our algorithm, we choose the challenging environments as the experiment areas. We control the robot to move at normal speed by a notebook in the same local networks, and output pose in real time using our algorithm, whose experimental image is shown in Fig. 7 (b) and (d). At the same time, we collect monocular image, IMU data and radar positioning results. We compare our result with VI-ORB, as shown in Fig. 8. The gray dashed lines marked "groundtruth" in Fig. 8 represents the trajectories of ground truth obtained by the radar location. Fig. 8 (a) shows the VI-ORB method has position drift, while our method can track with better accuracy. As shown in Fig. 8 (b), VI-ORB fails to track while the proposed method completes the whole experiment. Furthermore, for the scene of the rest area, which is not only a



(a) (c)

(b) (d)

Fig. 7: The original images collected by our mobile robot and corresponding experimental images from our method. (a) and (b) are images in the laboratory. (c) and (d) are images in the yoga studio.



(a) yoga studio (b) laboratory (c) rest area

Fig. 8: The three trajectories are the outputs from the proposed method and VI-ORB in the (a) yoga studio; (b) laboratory; (c) rest area. The gray dashed line marked "groundtruth" in the figure represents the ground truth trajectory obtained by the radar location.

low illumination but also a low texture environment, Fig. 8 (c) verifies the robustness and accuracy of our method compared with the other excellent method.

IV. Conclusions

In this paper, we propose a monocular visual-inertial SLAM algorithm for the low-light environments. Firstly, two image preprocessing methods are combined to enhance the brightness and contrast of images greatly. Then the optical flow and operation of inserting keyframe provide enough feature correspondences and map points to maintain tracking. Finally, the IMU error term is utilized to acquire accurate pose estimate with the feature reprojection error based on feature correspondences by constructing optimization function. We show superior performance by comparing against state-of-the-art open source implementations on public datasets and real-world experiments.

References

[1] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *The International Journal of Robotics Research (IJRR)*, vol. 34, no. 3, pp. 314–334, 2015.

[2] R. Mur-Artal and J. D. Tardós, "Visual-inertial monocular SLAM with map reuse," *IEEE Robotics and Automation Letters (RA-L)*, vol. 2, no. 2, pp. 796–803, 2017.

[3] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint kalman filter for vision-aided inertial navigation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2007, pp. 3565–3572.

[4] M. Li and A. I. Mourikis, "High-precision, consistent EKF-based visual-inertial odometry," *The International Journal of Robotics Research (IJRR)*, vol. 32, no. 6, pp. 690–711, 2013.

[5] Z. Huai and G. Huang, "Robocentric visual-inertial odometry," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 6319–6326.

[6] G. Huang, "Visual-inertial navigation: A concise review," *arXiv preprint arXiv:1906.02650*, 2019.

[7] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual–inertial odometry," *IEEE Transactions on Robotics (TRO)*, vol. 33, no. 1, pp. 1–21, 2017.

[8] T. Qin, P. Li, and S. Shen, "VINS-Mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics (TRO)*, vol. 34, no. 4, pp. 1004–1020, 2018.

[9] S. Maity, A. Saha, and B. Bhowmick, "Edge SLAM: Edge points based monocular visual SLAM," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2408–2417.

[10] Y. He, J. Zhao, Y. Guo, W. He, and K. Yuan, "PL-VIO: Tightly-coupled monocular visual-inertial odometry using point and line features," *Sensors*, vol. 18, no. 4, p. 1159, 2018.

[11] S.-J. Li, B. Ren, Y. Liu, M.-M. Cheng, D. Frost, and V. A. Prisacariu, "Direct line guidance odometry," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 1–7.

[12] G. Yadav, S. Maheshwari, and A. Agarwal, "Contrast limited adaptive histogram equalization based enhancement for real time video system," in *IEEE Advances in Computing, Communications and Informatics (ICACCI)*, 2014, pp. 2392–2397.

[13] S.-C. Huang, F.-C. Cheng, and Y.-S. Chiu, "Efficient contrast enhancement using adaptive gamma correction with weighting distribution," *IEEE Transactions on Image Processing*, vol. 22, no. 3, pp. 1032–1041, 2012.

[14] S.-D. Chen and A. R. Ramli, "Minimum mean brightness error bi-histogram equalization in contrast enhancement," *IEEE Transactions on Consumer Electronics*, vol. 49, no. 4, pp. 1310–1319, 2003.

[15] H. Li, L. Hao, Q. Zhang, X. Hu, and J. Cheng, "A lifted semi-direct monocular visual odometry," in *IEEE International Conference on Data Science and Computational Intelligence (DSCI)*, 2019, pp. 422–426.

[16] J.-Y. Bouguet *et al.*, "Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm," *Intel Corporation*, vol. 5, no. 1-10, p. 4, 2001.

[17] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The EuRoC micro aerial vehicle datasets," *The International Journal of Robotics Research (IJRR)*, vol. 35, no. 10, pp. 1157–1163, 2016.

[18] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics (TRO)*, vol. 28, no. 5, pp. 1188–1197, 2012.

[19] 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 (TRO)*, vol. 33, no. 5, pp. 1255–1262, 2017.

[20] L. Hao, H. Li, J. Cheng, and Q. Zhang, "EIP-VIO: Edge-induced points based monocular visual-inertial odometry," in *IEEE International Conference on Data Science and Computational Intelligence (DSCI)*, 2019, pp. 416–421.