2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

UVIP: Robust UWB aided Visual-Inertial Positioning System for Complex Indoor Environments*

Bo Yang, Jun Li, and Hong Zhang, *Fellow, IEEE*

***Abstract*— Indoor positioning without GPS is a challenge task, especially, in complex scenes or when sensors fail. In this paper, we develop an ultra-wideband aided visual-inertial positioning system (UVIP) which aims to achieve accurate and robust positioning results in complex indoor environments. To this end, a point-line-based stereo visual-inertial odometry (PL-sVIO) is firstly designed to improve the positioning accuracy in structured or low-textured scenarios by making use of line features. Secondly, a loop closure method is proposed to suppress the drift of PL-sVIO based on image patch features described by a CNN for handing the situation of a large environment and viewpoint variation. Thirdly, an accurate relocalization approach is presented for the case when the visual sensor fails. In this scheme, a top-to-down matching strategy from image to point and line features is presented to improve relocalization performance. Finally, the UWB sensor is combined with the visual-inertial system to further improve the accuracy and robustness of the positioning system and provide the results in a fixed reference frame. Thus, desirable real-time positioning results are derived for complex indoor scenes. Evaluations on challenging public datasets and real-world experiments are conducted to demonstrate that the proposed UVIP can provide more accurate and robust positioning results in complex indoor environments, even in the case when the visual sensor fails or in the absence of UWB anchors.**

I. Introduction

Indoor positioning technique which aims to compute in real-time accurate 6D poses of an agent in indoor environments plays an essential role in the field of robotics such as indoor service robots or logistics robots [1-3]. Recently, considerable progress has been made in indoor positioning, although challenges remain, especially with regard to achieving accurate and robust results in a complex scene where GPS signal is unavailable with significant illumination variation and structured, low-textured and repetitive spaces. Therefore, multi-sensors have been widely relied on including wheel encoder [4], light detection and ranging (LiDAR) [5], camera [6], ultra-wideband (UWB) [7] and inertial measurement unit (IMU) [8]. Among the above-mentioned sensors, 3D LiDAR obtains the highest accuracy, whereas it suffers from high costs in practice.

*Research supported by National Natural Science Foundation of China (Grant no. 51775110, 61703096).

Bo Yang is with the School of Artificial Intelligence, Nanjing University of Information Science & Technology, Nanjing, China (e-mail: ybseu@seu.edu.cn).

Jun Li is with the School of Computer and Electronic Information, Nanjing Normal University, Nanjing, China (e-mail: lijuncst@njnu.edu.cn)

Hong Zhang is with the Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, China, on leave from the University of Alberta, AB, Canada (e-mail: zhangh33@sustech.edu.cn).

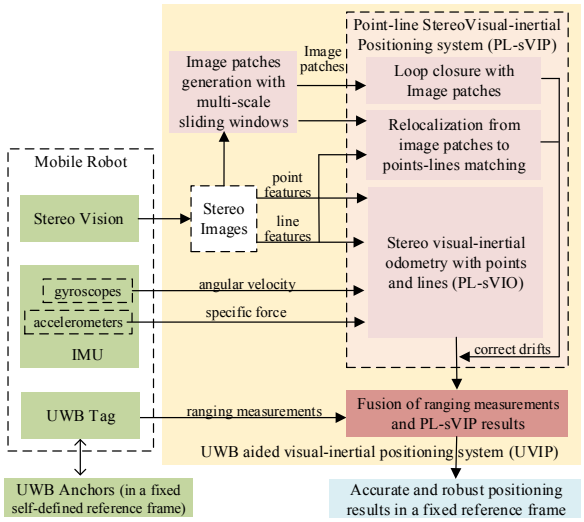

Figure 1. The flowchart of the UVIP system. Using the multi-sensors highlighted in green parts, we firstly design a PL-sVIO module which is subsequently combined with the presented loop closure and relocalization strategies leading to a PL-sVIP module. Furthermore, an effective data fusion method is proposed to combine the UWB data with the PL-sVIP to achieve the accurate and robust positioning performance in a fixed reference frame.

With the recent development in robot vision, accurate positioning systems are developed with the help of low-cost cameras in combination with inertial sensors [6, 9-14]. Many systems have been designed for indoor environments, although they suffer from several drawbacks that limit their applications in practice.

On one hand, most of these current visual-inertial systems focus on positioning accuracy in the common environment without considering some particular situations. For example, visual tracking will fail in the case of large illumination variation or in a dark scene, whilst the low-cost MEMS-IMU is uncertain for long-term positioning. Besides, single visual feature association is unreliable in a structured and repetitive indoor environment. Thus, the aforementioned cases lead to the deteriorating positioning performance or even unreliable results in complex indoor scenes.

On the other hand, a visual-inertial positioning system relies on relative localization mechanism, and thus it suffers from the accumulative drifts. In addition, due to this mechanism, it lacks a fixed reference frame, and hence, the positioning results depend on the starting position of the system which is inconvenient in practice.

Regarding the above-mentioned two limitations, we propose two-fold scheme to address these issues. On one hand, we design a more accurate and robust point-line-based stereo visual-inertial positioning system (PL-sVIP) for complex scenes. Specifically, line features are introduced to handle

structured and low-texture indoor scenes. Besides, a loop closure method is designed based on image patch features for a large illumination variation or repetitive space to correct the accumulated errors of VIO. Furthermore, we present a relocalization method with image blocks to handle the case when visual tracking fails in dark or low-textured scenes.

On the other hand, we introduce UWB sensor to positioning system based on an optimization framework. Although UWB is affected by the construction of anchors and unsmooth positioning track, it can achieve cm-to-dm positioning accuracy in a fixed coordinate system and is drift-free for environment changes [15]. Therefore, it is reasonable to integrate it with PL-sVIP to obtain drift-free, accurate and robust positioning results in a fixed reference frame. Besides, when visual sensor fails, UWB can be fused with inertial sensor to provide a high quality positioning result.

In summary, in this paper, we propose an UWB aided visual-inertial positioning system (UVIP) to achieve desirable and robustness positioning results in a complex indoor scene. Figure 1 illustrates the main system components of UVIP while the main contributions are summarized as follows:

1) Line and image patch features are developed for the visual-inertial positioning system, and thus improves the accuracy and robustness of system in complex indoor scenes.

2) An optimization-based fusion method is proposed to combine the UWB ranging measurements with visual-inertial positioning results, and improves positioning performance in complex indoor environments.

II. Related Work

The proposed positioning system is analogous to visual-inertial SLAM, whereas our system aims to record the real-time 6D estimated poses without the global optimization while the latter focuses on building a precise map using global or local Bundle Adjustment (BA) in common.

The first complete tightly-coupled visual-inertial system is MSCKF [9] which is constructed based on a multistate constraint Kalman filter. However, filtering-based VI-SLAM suffers from the linearization error caused by the linearization process of the Kalman Filter (KF). In recent years, with the development of IMU preintegration algorithm, keyframes and BA are utilized for the visual-inertial system such as OKVIS [10], ORB-SLAM-VI [11]. The most representative method VINS-Mono [6] achieves the state-of-the-art performance by using Shi-Tomasi points and Lucas-Kanade tracker in the front-end, and pose-graph optimization with DBoW2-based loop closure detection in the back-end. In addition, it has been extended to stereo vision [12] which is friendly for indoor wheeled robots.

All the aforementioned systems only use feature points which are unstable in some low-textured scenarios exemplified by man-made or indoor environments. To address this issue, massive efforts are devoted involving the introduction of the line features into VI-SLAM [13-14]. Trifo-VIO is an efficient stereo point-line-based VIO method proposed in [14]. However, the back-end component involved is built on EKF, whilst it lacks a loop closure and relocalization module without sufficient robustness to the complex environment. In this paper, in order to design an accurate and robust system, we extend line features to stereo visual-inertial systems based on tightly-coupled optimization-based framework. Besides, novel loop closure and relocalization methods are also used to handle complex scenarios.

Next, we review the loop closure methods in VI-SLAM. These methods are mostly based on hand-crafted features with the bag-of-words model such as DBoW2 [16]. However, this kind of methods is sensitive to the environment variations or repeated scenes and thus, cannot achieve desirable results in a complex indoor environment. Recently, deep CNN features have been exploited in loop closure [17, 18]. In these methods, the whole or local regions of an image are described by CNN features and the loop closure is reduced to feature matching followed by computing the overall similarity between images [19]. These studies reveal that using CNN features can improve the correction of feature matching, and thus, benefits the performance of loop closure. Therefore, in this paper, we further explore the benefits of CNN features for loop closure in complex scenes.

In addition, relocalization aiming to relocate the system when tracking fails, is another critical system component that can improve the robustness of a visual-inertial system. However, most of existing systems ignore this module [10, 13-14] or develop this algorithm based on the DBoW2 [6, 11, 12], which is prone to the complex indoor scenes as mentioned before. Hence, in this paper, following the loop closure methods with CNN features, we adopt a robust and accurate keyframe-based relocalization approach using image patch and point-line matching.

Finally, the UWB localization method is a common indoor positioning technology. Recently, the research focuses on integrating UWB with other sensors such as IMU [7, 8] or LiDAR [20] to achieve a high accuracy in a non-line-of-sight (NLOS) and complex environment. The current data fusion schemes are mainly built on KF [7-8, 21-22], and rely on the state equation of the integrated system. However, the state equation is highly nonlinear or is difficult to build up in some complex situations, especially for robotics. The simplified or linearized result from KF will bring the estimation error or make the system be sensitive to the outliers. Therefore, inspired by the Maximum a Posteriori estimation in a SLAM system which is independent of the state equation, in this paper, the nonlinear optimization method is explored to fuse the UWB and the visual-inertial system.

III. System Overview

As shown in Figure 1, we propose four main steps to process and fuse visual, IMU and UWB data. They are summarized as follows:

1) **Stereo VIO with points and lines** is designed to fuse the visual and IMU information in an optimization-based tightly-coupled manner and aims to provide a preliminary positioning result. The centerpiece of this approach is the introduction of line features. Compared with point features, line features provide more geometrical structural information of environments. Therefore, they are capable of improving the reliability of feature matching in human-made scenarios such as indoor scenes, and thus benefit the VIO precision. In

addition, a keyframe-based sparse map is also recorded real-time for the loop closure and relocalization.

2) **Loop Closure with image patches** can suppress VIO drifts that accumulate over time. In this method, image patches with their CNN feature descriptors are introduced to handle the situation of illumination and viewpoint changes, and thus yielding a better loop closure performance in complex scenes. In addition, we propose an image patches generation method using multi-scale sliding windows to obtain features with invariance to environment changes. These invariant features improve the accuracy of image patch matching in complex scenes, and thus benefit the loop closure performance.

3) **Relocalization from image patches to points-lines matching** is proposed for the case when visual sensor fails. It can compute an accurate pose when the visual sensor recovers, and thus correct the positioning errors caused by the failure of visual sensor. To achieve a precise result with recorded sparse map, the proposed loop closure with image patches method is used firstly to provide the nearest keyframes of current image with their pairwise matched image patches. Then, points and lines are matched between the pairwise patches and combined finally. This step makes the feature matching process robust to large viewpoint changes which is the main challenge of the relocalization based on sparse map. In the end, a 6D pose is calculated by OPnPL[23].

4) **Multi-sensor fusion method** is designed based on a graph-optimization framework to fuse UWB ranging measurements and visual-inertial positioning results in a fixed reference frame. In our cost function, UWB data provide global constraints while results from visual-inertial system provide relative constraints between two times. In addition, for the UWB untrusted measurement caused by NLOS, a simple but useful UWB outlier rejection method is utilized. It can improve the accuracy and robustness of the system.

TABLE I. POSITIONING MODES OF THE UVIP

Modes	UWB-visual-inertial	Visual-inertial	UWB-inertial	Inertial-only
Scenes	At least 4 useful UWB data; Visual tracking well	Lack useful UWB data; Visual tracking well	At least 4 useful UWB data; Visual sensor fails	Lack useful UWB data; Visual sensor fails
Precisi-on	Highest	Higher	Higher	High in short-term Low in long-term

Finally, different positioning modes of the UVIP in different scenes are summarized in Table I. These modes can provide valuable localization results in different complex environments, and thus indicate the robustness of the system. In addition, due to the designed optimization framework, the essential difference between various modes is the fluctuation of components in the cost function. Thus, different modes can seamlessly switch between each other. Besides, since the visual-inertial system relies on relative location mechanism, the relocalization method is presented to correct drifts when inertial-only mode is converted to visual-inertial mode.

IV. PROPOSED METHODS

A. Features

We use three kinds of features in the proposed system: points extracted by ORB [24], lines extracted by LSD [25] and described by LBD [26], and image patches. Among them, point and line features are extracted in each image while the image patches are only generated in keyframes. Next, we elaborate the image patches generation method with multi-scale sliding windows.

For an image, we define 'scales' as the ratio of the size of image patch over the entire image, and keep the aspect ratio of image patch similar to the original image. In addition, 4 scales of sliding window, referred as s=[0.16,0.25,0.36,0.49], are utilized in this method to obtain the better shift invariance. For each scale, for the sake of efficiency, 25 patches are extracted by a standard sliding window procedure [29], and thus a total of 100 image patches following spatial uniform distribution are generated. Thereafter, their high-level representations derived from the third convolutional layer of pre-trained AlexNet [27] are computed and the Gaussian Random Projection [28] is used to reduce the dimensions of CNN features to 1024. After normalization, the whole image is represented as a hundred 1024-dimensional vectors with high invariance to environment changes [29].

B. Stereo visual-inertial odometry with points and lines

For the PL-sVIO, the point is represented by the inverse depth, while the line is represented by Plücker coordinates[30] in 3D space and represented by Orthonormal [30] in optimization. The stereo and sequence frame matching are performed to triangulate features in 3D space and establish reprojection residuals, respectively. Besides, the IMU data is handled by the IMU preintegration approach [31].

For the data fusion, the estimated states at time k are defined as follows:

$$\boldsymbol{X}_{VI,k}=\left[\boldsymbol{q}_k^{VI},\boldsymbol{v}_k^{VI},\boldsymbol{p}_k^{VI},\boldsymbol{\varepsilon}_k^{I},\nabla_k^{I},\lambda_{k,i},...,\lambda_{k,i+n},\boldsymbol{O}_{l,i},...,\boldsymbol{O}_{l,i+m}\right] \quad (1)$$

where $\boldsymbol{q}_k^{VI}$, $\boldsymbol{v}_k^{VI}$ and $\boldsymbol{p}_k^{VI}$ are the rotation, velocity and position in the visual-inertial reference frame, respectively. $\boldsymbol{\varepsilon}_k^{I}$ and ∇_k^{I} are gyroscope and accelerometer bias, respectively. $\lambda_{k,i}$ and $\boldsymbol{O}_{l,i}$ are ith inverse depth of point features and Orthonormal representation of line features, respectively.

The estimation is formulated as follows:

$$\min_{\boldsymbol{X}_{VI}}\sum_{(k,j)\in P}\boldsymbol{\rho}\left(\left\|\boldsymbol{r}_{\boldsymbol{P}}\left(\boldsymbol{z}_{P_j}^{C_k},\boldsymbol{X}_{VI}\right)\right\|_{\boldsymbol{\Omega}_{P_j}}^2\right)+\sum_{(k,l)\in L}\boldsymbol{\rho}\left(\left\|\boldsymbol{r}_{\boldsymbol{L}}\left(\boldsymbol{z}_{L_l}^{C_k},\boldsymbol{X}_{VI}\right)\right\|_{\boldsymbol{\Omega}_{L_l}}^2\right)+\sum_{k\in IMU}\boldsymbol{\rho}\left(\left\|\boldsymbol{r}_{\boldsymbol{IMU}}\left(\boldsymbol{z}_{I_{k+1}}^{I_k},\boldsymbol{X}_{VI}\right)\right\|_{\boldsymbol{\Omega}_{I_k}}^2\right) \quad (2)$$

where $\boldsymbol{\rho}(\cdot)$ represents the Huber norm. $\boldsymbol{r}_{\boldsymbol{P}}(\cdot)$ and $\boldsymbol{r}_{\boldsymbol{L}}(\cdot)$ are the reprojection residuals of point and line features, respectively. $\boldsymbol{r}_{\boldsymbol{IMU}}(\cdot)$ is the inertial residual. $\boldsymbol{\Omega}_{P_j}$, $\boldsymbol{\Omega}_{L_l}$ and $\boldsymbol{\Omega}_{I_k}$ are related measurement covariances, respectively.

We leverage Levenberg–Marquardt (L-M) method for optimization of the above problem. Besides, the sliding window and marginalization strategy [6] are utilized for the sake of efficiency and accuracy. In addition, we follow [6] for the visual-inertial alignment, keyframe insertion and point

cloud build up, whereas we extend them with both point and line features along with stereo camera.

Finally, the PL-sVIO is combined with loop closure and relocalization methods leading to PL-sVIP. In the system, the loop closure approach is used to detect the loop for keyframes. If a loop is detected, a constraint between current keyframe and the loop is built with their position and rotation, and then it is added to the sliding window as a factor which can be optimized by L-M method. Besides, visual information is considered invalid when the number of point and line features extracted in one image is less than 100. After the visual sensor recovers, the relocalization procedure is executed.

C. Loop closure with image patches

The loop closure procedure is based on image patch features matching between the current keyframe and previous keyframes (excluding the nearest 10 frames of the current keyframe). To be specific, the nearest neighbor search based on cosine distance of features are utilized to match the features and only the reciprocal matches are identified as true matches. The overall distance between two images is calculated as the sum of distances of the matched image patches. The ratio of the second lowest to the lowest distance is compared with a pre-set threshold to produce loop closure results.

D. Relocalization from image patches to point-line matching

The relocalization procedure is executed when the visual sensor recovers. It mainly contains three steps. In the first step, the overall distances between the current image and previous keyframes are computed followed by the proposed loop closure scheme. Meanwhile, the matched pairwise image patches of the 6 nearest keyframes are returned.

In the second step which is the core of this method, related 3D points and lines of keyframe-based sparse map are matched with these 2D features in current image within the returned matched pairwise image patches. To be specific, firstly, we only consider the image patches of scales less than 0.4 for suppressing false positive matches, and discard the pairs with large distance if over 90% of the same 3D features are contained in two different image patches. Secondly, the feature matching is performed using the FLANN method combined with the RANSAC algorithm. Note that a line feature belongs to an image patch if it contains more than 70% of the line. Thirdly, all image-patch-specific matches are combined to produce the final 2D-to-3D correspondences for pose estimation. In this procedure, on one hand, for the redundant matched features, we only preserve one matched pair with the smallest distance. On the other hand, if the number of matched point and line features between pairwise matched image patches is less than a pre-set threshold, we discard all the matches between this matched pair to reduce the impact of false positive matches. This top-to-down matching strategy processing from image to point and line features is less vulnerable to the appearance and viewpoint variations, and thus provides a set of robust 2D-to-3D point and line matches for the accurate pose estimation [32].

In the final step, the OPnPL method is used to estimate the 6D pose with the set of 2D-to-3D correspondences.

E. Fusion of UWB data and PL-sVIP results

In the optimization-based UVIP method, UWB provides global measurements while visual-inertial system (PL-sVIP) provides local measurements for the whole system. The estimated states are defined as follows:

$$\boldsymbol{X}_{UVIP} = \left[\boldsymbol{p}_k, ..., \boldsymbol{p}_{k+n}, \boldsymbol{q}_k, ..., \boldsymbol{q}_{k+n}, \boldsymbol{P}_{U,0}^{VI} \right] \tag{3}$$

where $\boldsymbol{p}_k$ is the system position in time k, $\boldsymbol{q}_k$ is the system orientation represented as quaternion in time k, and $\boldsymbol{P}_{U,0}^{VI}$ is the transformation between UWB tag frame and visual-inertial frame.

The estimation is formulated as follows:

$$\boldsymbol{F}_{UVIP} = \min_{\boldsymbol{X}_{UVIP}} \sum_{\boldsymbol{k}} \left(\sum_{\boldsymbol{i}} \boldsymbol{\rho} \left(\left\| \boldsymbol{r}_{U,i} \right\|_{\boldsymbol{\Omega}_i^k}^2 \right) + \boldsymbol{\rho} \left(\left\| \boldsymbol{r}_{VI,k} \right\|_{\boldsymbol{\Omega}_{VI}^k}^2 \right) \right) \tag{4}$$

where k is the time series. i is the anchor sequence. $\boldsymbol{\Omega}_i^k$ and $\boldsymbol{\Omega}_{VI}^k$ are measurement covariances. $\boldsymbol{r}_{U,i}$ and $\boldsymbol{r}_{VI,k}$ are the UWB and visual-inertial factors, respectively defined as:

$$\boldsymbol{r}_{U,i} = d_{k,i} - \left\| \boldsymbol{p}_k - \boldsymbol{p}_{A,i} \right\|_2$$

$$\boldsymbol{r}_{VI,k} = \begin{bmatrix} \left(\boldsymbol{q}_k^{VI} \right)^{-1} \left(\boldsymbol{p}_{k+1}^{VI} + \boldsymbol{q}_{k+1}^{VI} \boldsymbol{P}_{U,0}^{VI} - \boldsymbol{p}_k^{VI} - \boldsymbol{q}_k^{VI} \boldsymbol{P}_{U,0}^{VI} \right) - \left(\boldsymbol{q}_k \right)^{-1} \left(\boldsymbol{p}_{k+1} - \boldsymbol{p}_k \right) \\ \left(\left(\boldsymbol{q}_k^{VI} \right)^{-1} \boldsymbol{q}_{k+1}^{VI} \right)^{-1} \left(\left(\boldsymbol{q}_k \right)^{-1} \boldsymbol{q}_{k+1} \right) \end{bmatrix} \tag{5}$$

where $d_{k,i}$ is the ith UWB measurement, $\boldsymbol{p}_{A,i}$ is the ith fixed known anchor position. For this formulation, L-M method is used to estimate the states and the self-defined UWB coordinate system is used as the reference frame.

The calibration of UWB and visual-inertial frames can also be finished by Equation 5. In addition, since the UWB sensor in our experiments only provides the range information, we use a simple method [33] to distinguish the outliers of UWB measurements caused by NLOS. Specifically, $d_{k,i}$ is considered as an outlier if the following condition is satisfied:

$$\left| \left\| \boldsymbol{p}_{k-1} - \boldsymbol{p}_{A,i} \right\|_2 - d_{k,i} \right| > s \times v \Delta T \tag{6}$$

where v is the system maximum velocity, ΔT is the time interval, and s is the adjustment factor.

V. Evaluation

A. EuRoC dataset for visual-inertial system

In this section, we evaluate the PL-sVIP on the public EuRoC dataset [34]. This dataset contains 11 sequences with stereo images and IMU measurements recorded by a micro air vehicle (MAV) in two types of indoor environments, machine hall (MH) and Vicon room (V). The two indoor environments exhibit significant variances in the speed of MAV, illumination and texture changes, motion blur and so forth. Besides, it also gives the ground-truth recorded by VICON.

In comparative studies, stereo-vision VINS [12] with point-only features, loop closure and relocalization module using DBoW2 obtains the state-of-the-art performance, and is thus used as the competing method. In terms of performance measure, the root mean square error (RMSE) is utilized in our experiments for evaluation metric.

Note that, since we evaluate the real-time positioning system, all results are recorded real-time without the global optimization in experiments and the pose of keyframes optimized by the BA are not used for performance evaluation.

Table II gives the RMSE of translation and rotation of several sequences. The errors of PL-sVIO and VINS without the loop are also provided. It shows that PL-sVIO achieves better positioning results than VINS without loop, especially in difficult sequences such as MH04, MH05. This performance improvement is attributed to the line features used in our system. Line features are abundant in indoor or man-made scenes. Thus, in complex structured low-textured scenes, PL-sVIO can still extract enough features to estimate the motion while VINS suffers from lacks of point features. In addition, our system reports comparable results in easy sequences such as MH02, due to enough high-quality points in rich-textured scenes of point-only system. However, the proposed PL-sVIO system demonstrates the advantageous performance in most scenarios.

In addition, both the PL-sVIP and VINS achieve the performance improvement when they are combined with the loop closure module. This indicates the effectiveness of loop closure detection approach. The slight performance decline results from false loop results caused by the complex scenes.

Finally, PL-sVIP system obtains the superior positioning results than VINS with loop. We attribute this improvement in two aspects. On one hand, it benefits from the line features. On the other hand, the proposed loop closure method contributes to the performance boost, especially for complex indoor scenes. For example, for difficult MH05 sequence, the VINS obtains comparable results with or without loop closure and relocalization, while the PL-sVIP achieves significantly superior performance over the PL-sVIO. This implies the degraded accuracy of DBoW2 in complex environments, and thus no performance improvement is observed due to the false loop results. By contrast, the proposed loop closure approach has better accuracy in difficult scenarios, and thus benefits the performance of the PL-sVIP. This further suggests the effectiveness of the proposed loop closure method.

To sum up, the EuRoC dataset experiment illustrates the advantage of the proposed PL-sVIP system as well as the presented loop closure method, especially in difficult scenes.

TABLE II. RELATIVE RMSE ERRORS IN THE EUROC DATASET. THE TRANSLATION (M) AND ROTATION (DEG) ERROR ARE LISTED AS FOLLOWS

Seq.	PL-sVIO (our)		VINS without loop		PL-sVIP (our)		VINS with loop	
	trans	rot	trans	rot	trans	rot	trans	rot
MH01	0.17	2.25	0.21	4.34	**0.15**	**1.56**	0.16	2.69
MH02	0.18	**1.97**	0.18	3.65	**0.17**	2.17	0.18	2.32
MH03	0.19	1.54	0.26	2.67	**0.18**	**1.38**	0.26	2.46
MH04	0.37	1.73	0.40	2.48	**0.34**	**1.29**	0.37	2.64
MH05	0.25	1.26	0.30	1.97	**0.14**	**1.07**	0.31	2.97
V102	0.10	**2.02**	0.09	3.04	**0.06**	3.62	0.08	2.47
V103	0.09	9.32	0.17	14.8	**0.08**	**8.87**	0.18	13.4
V201	0.09	2.25	0.08	1.86	0.08	**1.86**	**0.08**	2.24
V202	0.08	4.82	0.14	9.10	**0.07**	**3.54**	0.13	6.55
V203	0.32	**3.89**	0.34	4.61	**0.22**	4.75	0.27	4.72

B. Real-world experiments for UVIP system

In order to evaluate the proposed UVIP system in complex scenes, especially in the case when sensors do not work properly, we have conducted a real-world experiment. Specifically, a mobile robot is built using EAI wheeled robot base, MYNT EYE Depth 120 camera with 640×480 stereo images and BMI088 IMU, and Mini3sPlus UWB sensor with one tag and 4 anchors. The UVIP system is computed in real-time in Nvidia AGX Xavier with ROS at 10Hz. Besides, a RS-LiDAR-16 with the LINS method [35] which achieves the cm-level localization precision is used as the ground-truth.

We evaluate the UVIP in a 20m×5m office environment with significant variances in illumination changes and repeated, low-textured scenarios. Figure 2(a) illustrates the diagram of the office. Four UWB anchors are placed in fixed positions in Scene B with a self-defined reference frame. Figure 2(b), (c) show the images collected by the left camera of Scene A and Scene B with matched point and line features, respectively.

Four sequences with different conditions are built in experiment: **1) No sensors fail**. The robot moves in Scene B and all sensors work properly. **2) Visual sensor fails**. The robot moves in Scene B at night. During the movement, turn off the light for a few times to make the visual sensor fails and turn on the light in the end. **3) UWB fails**. The robot moves from Scene B to Scene A where UWB anchors do not exist and moves back to Scene B in the end. **4) Both visual sensor and UWB fail**. The robot moves from Scene B to Scene A and moves back to Scene B in the end. During the movement in Scene A, we occlude the stereo camera for a few times.

Since the mature open sources of visual-inertial-UWB systems are unavailable, and most existing visual-inertial systems such as VINS fail in sequence 2 and 4, we compare the UVIP with the proposed PL-sVIP.

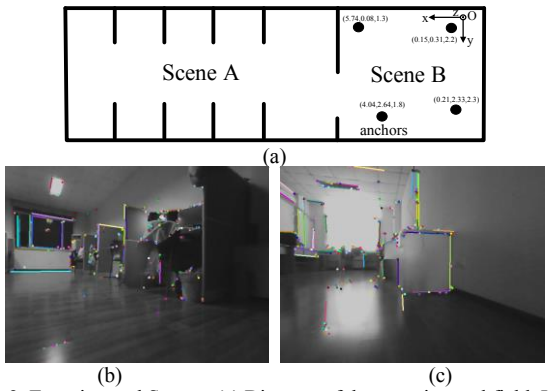

Figure 2. Experimental Scenes. (a) Diagram of the experimental field. Images of (b) Scene A and (c) Scene B with their matched point and line features.

Table 3 gives the results of the UVIP and the PL-sVIP. In all four sequences with different conditions, the UVIP significantly outperforms the PL-sVIP, which indicates the effectiveness of the integration with UWB sensors.

To demonstrate the results intuitively, Figure 3 shows several error curves of trajectories estimated by the UVIP and PL-sVIP. More specifically, Figure 3(a) illustrates the trajectory errors of PL-sVIP (left) and UVIP (right) in the case when the visual sensor fails. On one hand, for the PL-sVIP, it works on the inertial-only positioning mode when the visual sensor fails, and thus lead to accumulative and relatively large positioning errors (red cures in the left image). However, when visual sensor recovers, the drift is corrected by the

proposed relocalization approach and therefore, the system achieves desirable positioning results after that. By contrast, for most existing visual-inertial systems such as VINS which suppress drifts mainly based on loop closure or relocalization module built on the DBoW2, it is difficult to correct the large errors in complex scenes, making the localization fails in this sequence. Therefore, this indicates the advantage of the proposed relocalization method. On the other hand, for the UVIP, it achieves valuable results in whole trajectory even when the visual sensor fails, since the integration with UWB sensor can improve the positioning precision as suggested in condition 1. More importantly, it can help the IMU to obtain available positioning results when the visual sensor fails. This demonstrates the advantage of the fusion with UWB and the robustness of the UVIP in the case when visual sensor fails.

TABLE III. RELATIVE RMSE ERRORS IN REAL-WORLD EXPERIMRNTS. THE TRANSLATION (M) AND ROTATION (DEG) ERROR ARE LISTED

Seq.	PL-sVIP		UVIP	
	trans	rot	trans	rot
Condition 1	0.040	4.005	**0.036**	**3.923**
Condition 2	0.101	5.678	**0.051**	**4.427**
Condition 3	0.146	3.380	**0.109**	**3.06**
Condition 4	0.176	3.290	**0.141**	**3.215**

Figure 3(b) gives the trajectory errors of PL-sVIP and UVIP when the UWB fails. On one hand, in the trajectory of Scene B to Scene A, UWB tag cannot obtain enough useful measurements. In this case, the UVIP still obtains desirable positioning results because its working mode is switched to the visual-inertial mode (as same as the PL-sVIP) with higher positioning results. This indicates the effectiveness of outlier rejected scheme for UWB data and the robustness of UVIP in the case when UWB fails. On the other hand, in the trajectory of Scene A back to Scene B, both PL-sVIP and UVIP estimate the pose based on visual-inertial sensors, and results in drifts caused by large illumination variations and low-textured scenes. However, for PL-sVIP, these errors are accumulated until it detects a loop after entering Scene B for a few times. By contrast, for UVIP, the drifts are corrected when it enters Scene B, since UWB tag obtains enough useful ranging measurements, and then it is fused with the visual-inertial results. As a result, the drifts are corrected which indicates the effectiveness of the UVIP and UWB sensors.

Figure 3(c) illustrates the trajectory errors of PL-sVIP and UVIP in condition 4. It is clearly shown that the UVIP achieves the highest positioning performance based on three sensors in Scene B, while it works in visual-inertial mode with higher accuracy when entering Scene A. When the visual sensor fails in Scene A, the inertial-only mode produces available results in short-term situation whereas the drift increases in long-term situation. After visual sensor recovers, the relocalization method corrects the error immediately and the system obtains desirable positioning results. In the end, when system returns to Scene B, the UWB-visual-inertial mode is utilized to estimate accuracy poses again.

To sum up, this experiment with 4 sequences in various conditions shows the advantage of UVIP which can seamlessly switch among different working modes according to various scenes, and achieve accurate and robust positioning performance in complex scenes. Besides, the effectiveness of the proposed relocalization method is also verified.

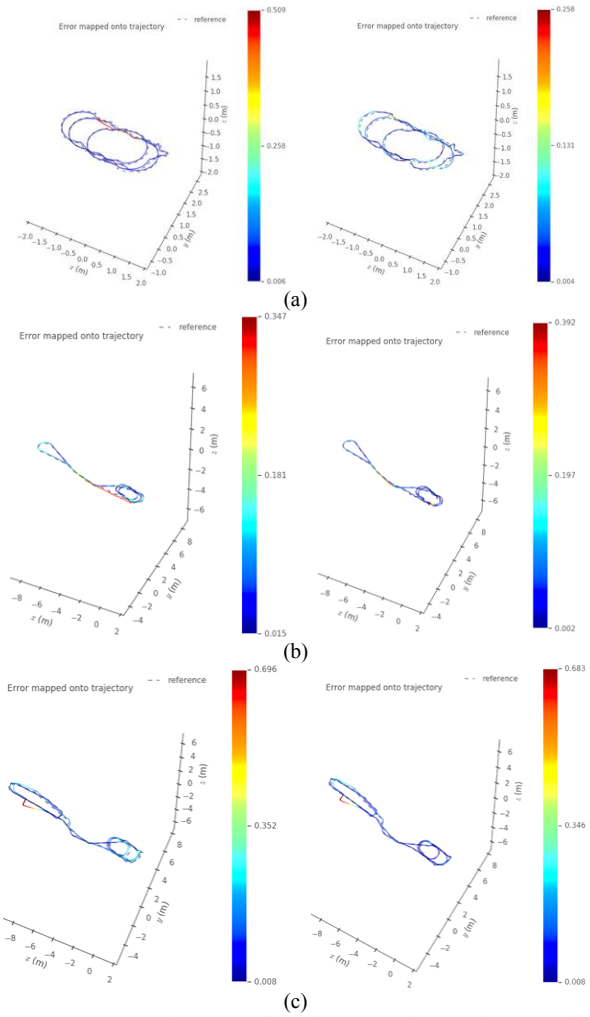

Figure 3. Comparison of UVIP (right) and PL-sVIP (left). In the cases when (a) visual sensor fails. (b) UWB fails. (c) both visual sensor and UWB fail.

VI. CONCLUSION

In this paper, we propose an accurate and robust positioning system based on UWB, visual and inertial sensors for complex indoor environment which includes unavailable GPS signals, structured and repetitive spaces and large illumination changes, low-textured scenarios. In contrast to the classical point-only visual-inertial systems, our method exploits multiple features (point, line and image patch) for the visual-inertial positioning system. In addition, novel loop closure and relocalization approaches are also presented to improve the precision and robustness of the system in complex indoor scenes. Furthermore, the positioning system benefits from the UWB sensor to further suppress the positioning drifts and provides a fix reference frame for more user-friendly applications. Due to these improvements, the proposed UVIP achieves desirable positioning results in complex indoor environments, even in the case when sensors fail, which is unavailable for the existing visual-inertial systems. We demonstrate the clear advantage of our proposed UVIP system experimentally on the challenging dataset and real-world indoor scenarios.

REFERENCES

[1] Zhu X, Yi J, Cheng J, et al. “Adapted Error Map Based Mobile Robot UWB Indoor Positioning,” *IEEE Trans. Instrum Meas*, vol.69, Sep. 2020, pp. 6336-6350.

[2] Xiaoming D, Liefu A and Rong J. “Motion estimation of indoor robot based on image sequences and improved particle filter,” *Multimed Tools Appl*, vol.98, Nov. 2020, pp.29747-29763.

[3] Liu J, Pu J, Sun L, et al. “An Approach to Robust INS/UWB Integrated Positioning for Autonomous Indoor Mobile Robots,” *Sensors*, vol.19, Feb. 2020, pp.950.

[4] Zhang H , Hu B , Xu S , et al. “Feature fusion using stacked denoising auto-encoder and GBDT for Wi-Fi fingerprint based indoor positioning,” *IEEE Access*, vol.8, 2020, pp. 114741-114751.

[5] Liu S , Atia M M , Karamat T B , et al. “A LiDAR-aided indoor navigation system for UGVs,” Journal of Navigation, *J Navigation*, vol.68, MAR. 2015, pp.253-273.

[6] Qin T, Li P and Shen S. “VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator,” *IEEE Trans. Robot*, vol.34, 2018, pp. 1004-1020.

[7] Feng D, Wang C, He C, et al. “Kalman-Filter-Based Integration of IMU and UWB for High-Accuracy Indoor Positioning and Navigation,” *IEEE Internet Things*, vol.7, 2020, pp. 3133-3146.

[8] Wen K, Yu K, Li Y, et al. “A New Quaternion Kalman Filter Based Foot-Mounted IMU and UWB Tightly-Coupled Method for Indoor Pedestrian Navigation,” *IEEE Trans. Veh technol*, vol.69, 2020, pp. 4340-4352.

[9] Li M, Mourikis A. “High-precision, consistent EKF-based visual-inertial odometry,” *Int J Robot Res*, vol.32, 2013, pp. 690-711.

[10] Leutenegger S, Lynen S, Bosse M, et al. “Keyframe-based visual–inertial odometry using nonlinear optimization,” *Int J Robot Res*, vol.34, 2015, pp.314–334.

[11] Mur-Artal R, Tardós J D. “Visual-Inertial Monocular SLAM with Map Reuse,” *IEEE Robot Autom Let*, 2017, 2(2): 796-803.

[12] Qin T, Li P and Shen S. “A General Optimization-based Framework for Local Odometry Estimation with Multiple Sensors”, 2019, *arXiv: 1901.03638*.

[13] He Y, Zhao J, Guo Y. et al. “PL-VIO: Tightly-Coupled Monocular Visual–Inertial Odometry Using Point and Line Features,” *Sensors* vol.18, 2018, pp. 1159.

[14] Zheng F, Tsai G, Zhang Z, et al, "Trifo-VIO: Robust and Efficient Stereo Visual Inertial Odometry Using Points and Lines," *in Conf. Rec. 2018 IEEE Int. Conf. Intelligent Robots and Systems*, pp.3686-3693.

[15] Mendoza-Silva GM, Torres-Sospedra J, Huerta J. “A Meta-Review of Indoor Positioning Systems,” *Sensors*. vol.19, 2019, pp.4507.

[16] Galvez-López D and Tardos J D, "Bags of Binary Words for Fast Place Recognition in Image Sequences," *IEEE Trans. Robot*, vol. 28, Oct. 2012, pp. 1188-1197.

[17] Sünderhauf N, Shirazi S, Dayoub F, et al, "On the performance of ConvNet features for place recognition," *in Conf. Rec. 2015 IEEE Int. Conf. Intelligent Robots and Systems (IROS)*, pp. 4297-4304.

[18] Sünderhauf N, Shirazi S, Jacobson, et al. “Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free,” 2015 *Robotics: Science and Systems*, 2015.

[19] Hou Y. Zhang H, Zhou S. “Evaluation of Object Proposals and ConvNet Features for Landmark-based Visual Place Recognition,” *J Intell Robot Syst*, vol.92, 2018, pp.505–520.

[20] Song Y, Guan M, Tay W P, et al., "UWB/LiDAR Fusion For Cooperative Range-Only SLAM," i*n Conf. Rec. 2019 IEEE Int. Conf. Robotics and Automation(ICRA)*, pp. 6568-6574.

[21] Li M, Zhu H, You S, Tang C. “UWB-based Localization System Aided with Inertial Sensor for Underground Coal Mine Applications,” *IEEE Sensors Journal*, vol.20, Jun. 2020. pp. 6652-6669.

[22] Li J, Bi Y, Li K, et al., "Accurate 3D Localization for MAV Swarms by UWB and IMU Fusion," in *Proc. 14th Int. Conf. Control and Automation*, Anchorage, 2018, pp. 100–105.

[23] Vakhitov A, Funke J and Moreno-Noguer F. “Accurate and Linear Time Pose Estimation from Points and Line,” *European Conference on Computer Vision (ECCV)*. Cham, 2016. pp. 583-599.

[24] Rublee E, Rabaud V, Konolige K, Bradski G. “ORB: an efficient alternative to SIFT or SURF,” in *Proc. Int. Conf. Computer Vision (ICCV)*, Spain, 2011, pp.2564–2571.

[25] von Gioi R G, Jakubowicz J, Morel J M, et al. “LSD: a fast line segment detector with a false detection control,” *IEEE Trans. Pattern Anal*, vol. 32, 2010, pp.722–732.

[26] Zhang L, Koch R. “An efficient and robust line segment matching approach based on LBD descriptor and pairwise geometric consistency,” *J Vis Commun Image R*. vol. 24, 2013, pp. 794–805.

[27] Krizhevsky A, Sutskever I, Hinton G E. “ImageNet Classification with Deep Convolutional Neural Networks,” *Advances in Neural Information Processing Systems (NIPS)*. USA, 2012. pp. 1097–1105.

[28] Sanjoy D. “Experiments with random projection,” *Conference on Uncertainty in Artificial Intelligence*, 2000. Pp. 143–151.

[29] Yang B, Xu X, Li J, et al. “Landmark Generation in Visual Place Recognition Using Multi-Scale Sliding Window for Robotics,” *Appl Sci-Basel*, vol.9, 2019, pp. 3146.

[30] Bartoli A, Sturm P, “Structure from motion using lines: Representation, triangulation and bundle adjustment,” *Comput. Vis. Image. Und.* , vol. 100, Dec. 2005, pp.416–441.

[31] Forster C, Carlone L, Dellaert F, Scaramuzza D. “On-Manifold Preintegration for Real-Time Visual--Inertial Odometry,” *IEEE Trans. Robot*, vol.33, 2017. pp. 1-21.

[32] Yang B, Xu X, Li J, "Keyframe-Based Camera Relocalization Method Using Landmark and Keypoint Matching," *IEEE Access*, vol. 7, 2019. pp. 86854-86862.

[33] Fang X , Wang C , Nguyen T M , et al. “Graph Optimization Approach to Localization with Range Measurements”. 2018. *arXiv:1802.10276*

[34] Burri M, Nikolic J, Gohl P, et al. “The EuRoC micro aerial vehicle datasets,” *Int J Robot Res,* vol. 35, 2016 pp. 1157–1163.

[35] Qin C, Ye H, Pranata C E, et al., "LINS: A Lidar-Inertial State Estimator for Robust and Efficient Navigation," *in Proc. Int. Conf. Robotics and Automation(ICRA)*, Paris, 2020, pp. 8899-8906.