Tightly Coupled Integration of GNSS and Vision SLAM Using 10-DoF Optimization on Manifold

Zheng Gong, Rendong Ying, Wen Fei, *Member, IEEE*, Jiuchao Qian, and Peilin Liu, *Senior Member, IEEE*

***Abstract*—Vision navigation technique, especially the vision-based simultaneous localization and mapping (V-SLAM), plays a critical role in robotic navigation. As a relative positioning technique, V-SLAM often suffers from drift and scale uncertainty problems which incur bias increasing over time. To overcome these drawbacks and to improve the robustness and accuracy of localization, an effective way is to fuse global navigation satellite system (GNSS) with V-SLAM. In this paper, we propose a novel GNSS and SLAM fusion algorithm, which provides ego-motion estimation through tightly coupling GNSS pseudo-range measurements and camera feature points. It first decomposes the pose state into basic motion vectors, based on which an asynchronous tracking is performed. Then, a 10-DoF joint-optimization formulation on manifold is proposed to achieve tight fusion of the raw measurement from camera and GNSS. Finally, this formulation is solved to calculate the ego-motion state. The proposed algorithm is verified on an autonomous ground vehicle in two typical environments. The results demonstrated that, the new algorithm can amend the bias in vision SLAM and constrain the GNSS solution, which achieves a better localization result than the traditional methods.**

***Index Terms*—Sensor data fusion, robotic navigation, GNSS, vision SLAM, autonomous vehicle.**

I. Introduction

MORDEN navigation systems used in autonomous vehicles and robots rely on various kinds of systems such as global navigation satellite systems (GNSS), inertial navigation systems (INS), visual positioning systems (VPS) and wireless radio positioning systems [1] etc.. To further overcome drawbacks of individual sensors under challenging environments, studies on sensor fusion techniques draw attentions from researchers [2], [3]. Among the most commonly used sensors, GNSS has a significant complementarity with relative positioning methods, such as INS or VPS. On one hand, the fusion of INS with GNSS has been studied for years and has mature solutions [4], [5]. On the other hand, the fusion of vision positioning system with GNSS experiences less exploration due to their heterogeneous data types, which is the essential motivation of this paper.

Generally, there exist two classes of approaches for GNSS and VPS fusion, namely, the loosely-coupled approaches and the tightly-coupled approaches. Most of existing researches have focused on the loose-coupling methods [6]. Such methods couple the localization results separately obtained from GNSS receiver and on-board vision-processing unit with the use of filter and dead-reckoning [3], [7], [8]. Though efficient, loose-coupling cannot make full use of the information from multi-sensors [9], [10].

In comparison, tightly-coupled methods can make better use of the information from the sensors to further improve the performance. As vision is a backbone component in such fusion methods, we first introduce a major technique for VPS, vision-based simultaneous localization and mapping (V-SLAM). Developed from the structure from motion (SFM), V-SLAM has significant progress in the last decade, and the mature solutions [11], [12] make it a widely used component in robotic and autonomous vehicle navigation.

From the filter-based EKF-SLAM [13], [11] proposed a decade ago, recent progresses in V-SLAM have benefited a lot from the pose-graph based optimization methods [14], [12], which not only improves the robustness and accuracy but also reduces the computation load. However, this evolution poses a challenge to the existing fusion approaches [15], [16], [7], especially the tightly-coupled ones [9]. Specifically, the challenge is twofold: First, due to the heterogeneous data composition from different sensors, the information update and synchronization affect more than one coordinate at the same time. Second, a tightly-coupled method involves not only the location of the vehicle, but also its posture. As the global positioning result of GNSS is not independent from the local motion result of a vision system, the involved vision-GNSS joint processing has to take the displacement of GNSS into consideration under a local tangent space.

To overcome these difficulties, we propose a tightly coupled fusion scheme based on the pose-graph SLAM framework in this paper. By treating navigation satellites as known "feature points" using their pseudo-range and position information, we resolve the vehicle pose and the map points simultaneously from GNSS pseudo-ranges, vision feature points. Compared with existing methods, this unification resolving is tight on both GNSS and vision sides. That is, it can make better use of the full raw information on both sensors, while maintaining the maximum compatibility with existing systems. Further, the complementarity of these sensors information can promise a synergistic convergence effect to boost the system accuracy.

The first contribution of this paper is a motion-decomposition-based asynchronous tracking strategy which accommodates the heterogeneous data properly. This method first decomposes the vehicle motion into 3 sets of basic pose vectors corresponding to different reference coordinate systems. Then, ego-motion tracking is operated individually on each effective degree-of-freedom (DoF). The motivation for this decomposition comes from the fact that, measurements

The authors are with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China (e-mail: {gongzheng, rdying, wenfei, jcqian, liupeilin}@sjtu.edu.cn).

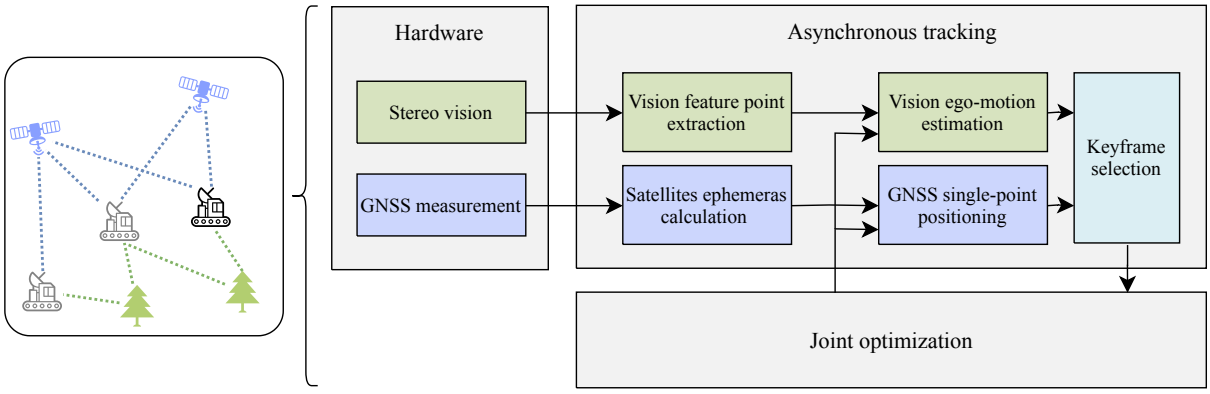


Fig. 1. System overview, showing the steps performed by asynchronous tracking and joint optimization.

provided by each sensor, although normally under different coordinates, exist overlaps. Meanwhile, different sensors have distinct observation significance on the overlapped parts. The decomposition not only maximizes the contribution of each sensor on its most significant observation, but also make the information asynchronous update feasible.

As a second contribution, this paper presents a novel joint graph-optimization formulation on manifold to solve both of the GNSS and vision states. The first step towards this goal is to collaborate a graph-based GNSS pseudo-range processing algorithm with the vision measurements. This is contrast to [9], which adopted a filter-based fusion algorithm. It has been shown in [17], [18], the graph-based pseudo-range processing algorithm is more robust and accurate, and has a good compatibility with V-SLAM algorithms. Then, in this proposed formulation, the dependency between the relative motion from vision estimation and the absolute localization result from GNSS is also taken into consideration. We properly addresses the manifold structure of these relative and global motions, and analytically derive all Jacobians with respected to GNSS pseudo-range measurement. This is a fundamental evolution compared with existing tightly coupled GNSS and vision fusion algorithms [9], [19], in which the state is represented in quaternion or Euler angles, and the state propagation between GNSS and vision has not been considered.

This rest of this paper is organized as follows. Section II introduces the overall system structure. In Section III, the motion decomposition strategy and the graph-based GNSS processing method are presented, respectively. Then, the asynchronous tracking algorithm along with the key frame selection procedure is presented. Section IV presents the detailed joint-optimization model with its derivative on tangent space. In Section V, the experiment platform implementation is introduced. The results with analysis are provided in Section VI. Finally, in Section VII, concluding remarks are given.

II. System Overview

The overall structure of the proposed fusion scheme is shown in Fig. 1. As a V-SLAM originated fusion system, the structure contains the typical dual-pipeline V-SLAM hierarchy [11], including tracking and mapping (bundle adjustment). The tracking stage, which we name it as asynchronous tracking in our system, takes an initial guess on the ego-motion estimation based on individual sensors and builds up the keyframes to be further optimized. Then, the joint optimization stage iteratively finds the global optimum considering both the GNSS and vision sensors using the initial guess from the tracking stage. The entire system is developed based on the state-of-the-art ORB-SLAM2 system [20], in which GNSS pseudo-range measurement is fused. ORB-SLAM2 is based on the ORB feature [21].

From the perspective of heterogeneous sensors fusion, we present a novel method to fuse the heterogeneous observations from the GNSS and vision sensors. On the vision side, we extract the ORB features following the setting in ORB-SLAM2 [20], which has been shown to have an outstanding overall performance. On the GNSS side, satellites are considered as landmarks, since each satellite can be uniquely identified by the transmitted pseudo random noise (PRN) code. There are two ways to model the satellite measurement, including the pseudo-range model and the carrier-phase model. Although nowadays these two models commonly work complementarily together [22], we choose the pseudo-range model here, since this work mainly concentrates on the concept verification of graph-based fusion rather than on GNSS position calculation.

III. Asynchronous Tracking on Decomposed Pose States

To achieve asynchronized tracking of the information of heterogeneous sensors under diverse coordinates, this section first proposes a motion decomposition analysis of ego-motion pose state. Furthermore, in order to unify the optimization framework, we propose a graph-based GNSS pseudo-range processing method and fit it under the decomposed coordinates. Then, based on these two prerequisites, we present the vision-GNSS joint asynchronous tracking. Finally the keyframe selection strategy is introduced based on the asynchronous tracking.

A. Ego-Motion Pose Decomposition

This subsection analyzes the coordinate systems involved in the considered multi-sensor system and presents a decomposition of them into elemental motions. Considering a vehicle

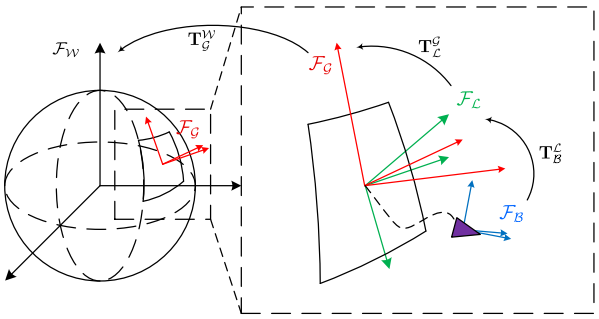


Fig. 2. Coordinates involved in vehicle motion with the transformation matrices between each other.

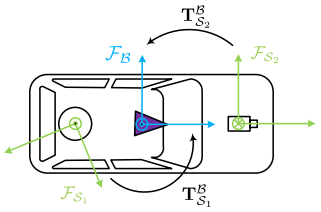


Fig. 3. Different sensor local coordinates on the same carrier. In this example, $\mathcal{F}_{\mathcal{S}1}$ stands for the GNSS frame, whilst the $\mathcal{F}_{\mathcal{S}2}$ for the vision frame.

equipped with a GNSS receiver, a 9-axis inertial measurement unit (IMU, including an accelerometer, a gyroscope and a magnetometer) and a vision system navigating in 3D space, four correlative coordinates are involved: body frame $\mathcal{F}_{\mathcal{B}}$, local ground world frame $\mathcal{F}_{\mathcal{G}}$, local world frame $\mathcal{F}_{\mathcal{L}}$, and global earth-centered, earth-fixed (ECEF) frame $\mathcal{F}_{\mathcal{W}}$, as shown in Fig. 2.

The frame $\mathcal{F}_{\mathcal{W}}$ is the ECEF conventional terrestrial system, which is the global inertial reference frame used by GNSS, and can be easily transformed with the geodetic LLH (Latitude-Longitude-Height) coordinate. The local ground tangent plane world frame $\mathcal{F}_{\mathcal{G}}$, which is denoted by red coordinate in Fig. 2, is the local world frame with geographical orientation information, such as east-north-up (ENU) coordinate in our choice. Then the local world frame $\mathcal{F}_{\mathcal{L}}$ (the green coordinate shown in Fig. 2) is the world frame used by the relative positioning sensors without the information of geographical orientation. Notice that, different sensors, especially the relative positioning sensors such as cameras and acceleration meters, can have different local world frames due to their separated mount points on the vehicle, as shown in Fig. 3. However, these frames can be unified into a single main local world frame through calibration or supposed to be negligible under large scale assumption [23]. For this reason, in this paper we assume that all transform matrices between these coordinates, including translation and rotation, can be expressed form as follows under homogeneous coordinate:

$$\begin{aligned}\mathbf{T}_{\mathcal{B}}^{\mathcal{L}} &= \begin{bmatrix}\mathbf{R}_{\mathcal{B}}^{\mathcal{L}} & \boldsymbol{t}_{\mathcal{B}}^{\mathcal{L}}\\ 0 & 1\end{bmatrix}\\ \mathbf{T}_{\mathcal{L}}^{\mathcal{G}} &= \begin{bmatrix}\mathbf{R}_{\mathcal{L}}^{\mathcal{G}} & \mathbf{0}\\ 0 & 1\end{bmatrix}\\ \mathbf{T}_{\mathcal{G}}^{\mathcal{W}} &= \begin{bmatrix}\mathbf{R}_{\mathcal{G}}^{\mathcal{W}}(\boldsymbol{t}_{\mathcal{G}}^{\mathcal{W}}) & \boldsymbol{t}_{\mathcal{G}}^{\mathcal{W}}\\ 0 & 1\end{bmatrix},\end{aligned} \tag{1}$$

where $\mathbf{R}_{\mathcal{B}}^{\mathcal{L}}$, $\mathbf{R}_{\mathcal{L}}^{\mathcal{G}}$, $\mathbf{R}_{\mathcal{G}}^{\mathcal{W}}(\boldsymbol{t}_{\mathcal{G}}^{\mathcal{W}})$ are 3×3 rotation matrices with 3 DoF and $\boldsymbol{t}_{\mathcal{B}}^{\mathcal{L}}$, $\boldsymbol{t}_{\mathcal{G}}^{\mathcal{W}}$ are 3×1 translation vectors with 3 DoF. It is worth noting that, the translation for $\mathbf{T}_{\mathcal{L}}^{\mathcal{G}}$, i.e. $\boldsymbol{t}_{\mathcal{L}}^{\mathcal{G}}$, is a zero-vector, for the reason that the local world frame, which is commonly initialized by relative positioning sensors such as V-SLAM, usually shares the same origin point with the geographical world frame $\mathcal{F}_{\mathcal{G}}$. Nevertheless, the rotation transform from the local ground world frame $\mathcal{F}_{\mathcal{G}}$ to the ECEF world frame $\mathcal{F}_{\mathcal{W}}$, $\mathbf{R}_{\mathcal{G}}^{\mathcal{W}}$, is not independent. It is determined by the given reference original point $\boldsymbol{t}_{\mathcal{G}}^{\mathcal{W}}$, which is the initial global ECEF location of the system. Here we represent $\boldsymbol{t}_{\mathcal{G}}^{\mathcal{W}}$, the translation from $\mathcal{F}_{\mathcal{G}}$ to $\mathcal{F}_{\mathcal{W}}$, as $\boldsymbol{t}_{\mathrm{o}}$,to emphasize that it is the original point of $\mathcal{F}_{\mathcal{G}}$ under $\mathcal{F}_{\mathcal{W}}$. Further, while $\boldsymbol{t}_{\mathrm{o}} \triangleq [t_{\mathrm{o}x}\ t_{\mathrm{o}y}\ t_{\mathrm{o}z}]^{\mathsf{T}}$ is under ECEF Cartesian coordinate system, we define $\boldsymbol{t}_{\mathrm{o}_{\mathrm{LLH}}} \triangleq [t_{\mathrm{o}\lambda}\ t_{\mathrm{o}\phi}\ t_{\mathrm{o}h}]^{\mathsf{T}}$, which is the curvilinear coordinate system representation of $\boldsymbol{t}_{\mathrm{o}}$. The transformation between these two variables is derived in [24]. Then, $\mathbf{R}_{\mathcal{G}}^{\mathcal{W}}$ is given by

$$\begin{aligned}&\mathbf{R}_{\mathcal{G}}^{\mathcal{W}}(\boldsymbol{t}_{\mathrm{o}_{\mathrm{LLH}}}) =\\ &\begin{bmatrix}-\sin t_{\mathrm{o}\lambda} & \cos t_{\mathrm{o}\lambda} & 0\\ -\sin t_{\mathrm{o}\phi}\cos t_{\mathrm{o}\lambda} & -\sin t_{\mathrm{o}\phi}\sin t_{\mathrm{o}\lambda} & \cos t_{\mathrm{o}\phi}\\ \cos t_{\mathrm{o}\phi}\cos t_{\mathrm{o}\lambda} & \cos t_{\mathrm{o}\phi}\sin t_{\mathrm{o}\lambda} & \sin t_{\mathrm{o}\phi}\end{bmatrix}.\end{aligned} \tag{2}$$

In the following, we denote $\mathbf{R}_{\mathcal{G}}^{\mathcal{W}}(\boldsymbol{t}_{\mathrm{o}_{\mathrm{LLH}}})$ by $\mathbf{R}_{\mathcal{G}}^{\mathcal{W}}$ for notation succinctness. Then, a point $\boldsymbol{p}^{\mathcal{B}} = [p_x\ p_y\ p_z]^{\mathsf{T}}$ under the body frame $\mathcal{F}_{\mathcal{B}}$, can be transferred into the ECEF frame as

$$\begin{aligned}\tilde{\boldsymbol{p}}^{\mathcal{W}} &= \mathbf{T}_{\mathcal{G}}^{\mathcal{W}}\mathbf{T}_{\mathcal{L}}^{\mathcal{G}}\mathbf{T}_{\mathcal{B}}^{\mathcal{L}}\tilde{\boldsymbol{p}}^{\mathcal{B}}\\ &= \begin{bmatrix}\mathbf{R}_{\mathcal{G}}^{\mathcal{W}} & \boldsymbol{t}_{\mathrm{o}}\\ 0 & 1\end{bmatrix}\begin{bmatrix}\mathbf{R}_{\mathcal{L}}^{\mathcal{G}} & \mathbf{0}\\ 0 & 1\end{bmatrix}\begin{bmatrix}\mathbf{R}_{\mathcal{B}}^{\mathcal{L}} & \boldsymbol{t}_{\mathcal{B}}^{\mathcal{L}}\\ 0 & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{p}^{\mathcal{B}}\\ 1\end{bmatrix},\end{aligned} \tag{3}$$

where $\tilde{(\cdot)}$ is the homogeneous representation of points as introduced in [25].

In practical autonomous vehicle navigation applications, for the convenience in aligning geographical map, the most important states in ego-motion estimation are $\mathbf{R}_{\mathcal{B}}^{\mathcal{G}}$ the attitude under a local ENU coordinate system, and $\boldsymbol{t}_{\mathcal{B}}^{\mathcal{W}}$ the position under an ECEF coordinate system. Specifically, we mainly concern the states of

$$\mathbf{R}_{\mathcal{B}}^{\mathcal{G}} = \mathbf{R}_{\mathcal{L}}^{\mathcal{G}}\mathbf{R}_{\mathcal{B}}^{\mathcal{L}} \tag{4}$$

and

$$\boldsymbol{t}_{\mathcal{B}}^{\mathcal{W}} = \mathbf{R}_{\mathcal{G}}^{\mathcal{W}}\mathbf{R}_{\mathcal{L}}^{\mathcal{G}}\boldsymbol{t}_{\mathcal{B}}^{\mathcal{L}} + \boldsymbol{t}_{\mathrm{o}}\ , \tag{5}$$

which are both composition of several basic states under different coordinates. Upon this, it is easy to see that the states contributing to the final results are $\mathbf{R}_{\mathcal{L}}^{\mathcal{G}}$, $\mathbf{R}_{\mathcal{B}}^{\mathcal{L}}$, $\boldsymbol{t}_{\mathcal{B}}^{\mathcal{L}}$ and $\boldsymbol{t}_{\mathrm{o}}$. Thus, with each state contains three DoF, there are totally 12 DoF should be taken into consideration.

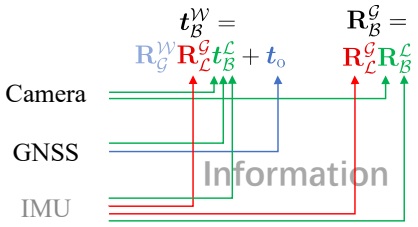


Fig. 4. Information contribution of sensors on the ego-motion estimation under the global world frame $\mathcal{F}_G$.

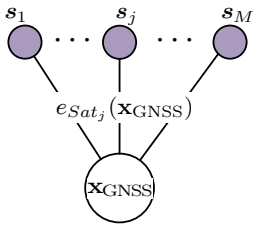


Fig. 5. The general pose graph model for the GNSS pseudo-range positioning problem.

After expressing the concerned states in terms of basic elements, we need to analyze the information contribution provided by each sensor. Considering a ground vehicle equipped with a stereo camera, a 9-axis IMU and a GNSS receiver, the observation model is shown in Fig. 4. Specifically, $\mathbf{R}_{\mathcal{L}}^{\mathcal{G}}$, the rotation from the local world frame to the local ENU frame, can only be measured by the magnetometer. In this work, IMU is only used for determining $\mathbf{R}_{\mathcal{L}}^{\mathcal{G}}$ for simplification. When it comes to $\mathbf{R}_{\mathcal{B}}^{\mathcal{L}}$, the rotation from the body frame to the local world frame, both V-SLAM and the gyroscope inside IMU can provide information. Then, the translation from the body frame to the local world frame, $\boldsymbol{t}_{\mathcal{B}}^{\mathcal{L}}$, can be estimated from accelerometer, V-SLAM and GNSS. Finally, $\boldsymbol{t}_{\text{o}}$, the original translation from the local frame to the global frame, can only be observed by GNSS through satellites. Noticing that $\mathbf{R}_{\mathcal{G}}^{\mathcal{W}}$ is dependent on $\boldsymbol{t}_{\text{o}}$, hence we do not need to establish a standalone observation on $\mathbf{R}_{\mathcal{G}}^{\mathcal{W}}$. Moreover, if $\boldsymbol{p}^{\mathcal{B}}$ is a map point under the body frame, only camera can give out its information.

This decomposition reveals one of the core cogitations of our work: Instead of fusing sensors directly on their final results, we resort to tighter fusion via assembling the decomposed observation vectors. This approach allows an asynchronous information update on the correspondence motion observation. To verity this idea, we simplify the condition and consider a system which contains only GNSS and stereo camera. In this condition, only $\mathbf{R}_{\mathcal{B}}^{\mathcal{L}}$, $\boldsymbol{t}_{\mathcal{B}}^{\mathcal{L}}$ and $\boldsymbol{t}_{\text{o}}$ are involved, while $\mathbf{R}_{\mathcal{L}}^{\mathcal{G}}$ is fixed through calibration using IMU.

B. Graph-based GNSS Pseudo-range Positioning

In this subsection, we discuss the pose-graph based pseudo-range optimization method mentioned in [17] and [18]. Using the concept from V-SLAM, we consider GNSS-based localization as a 3D localization problem with range-only observations to distant landmarks. In this manner, the localization problem is formulated into a pose-graph model as shown in Fig. 5, where $\mathbf{x}_{\text{GNSS}}$ is the GNSS state vector. $\boldsymbol{s}_j \triangleq [s_{j_x}\ s_{j_y}\ s_{j_z}]^{\mathsf{T}}$ represents the j-th satellite state under the world frame $\mathcal{F}_{\mathcal{W}}$ with related unary edge containing the error function $e_{Sat_j}(\mathbf{x}_{\text{GNSS}})$. As defined in [26], the edge is a concept representing the estimation error constrains. The landmarks in this scenario are the satellites which are uniquely identifiable via their transmitted PRN codes. The positions of the observed satellites/landmarks in space are known since each satellite transmits ephemeris parameters which describe its orbit. From the literatures such as [27], each landmark distance is known as pseudo-range ρ, which is the measurement from C/A code phase and ephemeris. Such pseudo-range measurements are subjected to a series of possible error sources, including but not limited to:

- receiver clock error δ_{Clock},
- ionospheric and tropospheric propagation error $\delta_{Atmosphere}$,
- earth rotation error $\delta_{EarthRotation}$,
- satellite clock and ephemeris error δ_{Sat}.

Among these errors, the receiver clock error has the largest effect as it is of rapidly-varying over time, compared with the slowly-varying bias-like $\delta_{Atmosphere}$, δ_{Sat} and $\delta_{EarthRotation}$. We take the unknown receiver clock error into consideration in the 3 DoF global ECEF position state of the vehicle, leading to a 4-dimensional state space:

$$\mathbf{x}_{\text{GNSS}} \triangleq [\boldsymbol{t}_{\mathcal{B}}^{\mathcal{W}\mathsf{T}}\ \delta_{Clock}]^{\mathsf{T}} \in \mathbb{R}^4\ , \tag{6}$$

where $\boldsymbol{t}_{\mathcal{B}}^{\mathcal{W}}$ is the vehicle position under the global world frame $\mathcal{F}_{\mathcal{W}}$ as defined in Eq. (5).

To achieve fused tacking based on heterogeneous sensors under different coordinate systems, we need to take the pose states decomposition mentioned in the previous subsection into consideration. GNSS measurement provides the information of global location, which contains information on both $\boldsymbol{t}_{\mathcal{B}}^{\mathcal{L}}$ and $\boldsymbol{t}_{\text{o}}$, the local relative translation and the global initial location. However, compared with vision's centimeter-level relative accuracy [19], [12], [11], the meter-level single-point localization result from GNSS is unreliable at the tracking stage. On the other hand, GNSS is the only sensor can provide the global information. With this in mind, the tracking strategy of GNSS here is to only refine the global initial location $\boldsymbol{t}_{\text{o}}$. Intuitively, given a reliable relative translation $\boldsymbol{t}_{\mathcal{B}}^{\mathcal{L}}$, refining a more accurate original point will improve the accuracy of the desired global position. Hence, we rewrite the state of GNSS $\mathbf{x}_{\text{GNSS}}$ as

$$\mathbf{x}_{\text{GNSS}} = [(\boldsymbol{t}_{rel} + \boldsymbol{t}_{\text{o}})^{\mathsf{T}}\ \delta]^{\mathsf{T}}\ , \tag{7}$$

where $\boldsymbol{t}_{rel} = \mathbf{R}_{\mathcal{G}}^{\mathcal{W}} \mathbf{R}_{\mathcal{L}}^{\mathcal{G}} \boldsymbol{t}_{\mathcal{B}}^{\mathcal{L}}$ is the constant relative motion prior from vision sensor and $\delta = \delta_{Clock}$ for short.

Once the state space and the measurement errors are given, we can build up the pose graph model for GNSS measurement.

Assuming n satellites are observed from the vehicle state $\mathbf{x}_{\mathrm{GNSS}}$ at a certain time frame, with each satellite providing a pseudo-range observation, e.g., ρ_j of the j-th satellite. Given the position of the observed j-th satellite $\boldsymbol{s}_j$, the pseudo-range measurement model is given by the function

$$\begin{aligned} h_{GNSS}(\mathbf{x}_{\mathrm{GNSS}}, \boldsymbol{s}_j) =& h_{GNSS}(\boldsymbol{t}_{\mathrm{o}}, \delta, \boldsymbol{s}_j) \\ =& ||\boldsymbol{s}_j - (\boldsymbol{t}_{rel} + \boldsymbol{t}_{\mathrm{o}})||_2 + \delta + \delta_{Sat} \\ &+ \delta_{Atmosphere} + \delta_{EarthRotation} \ . \end{aligned} \tag{8}$$

Then the pseudo-range observation ρ_j is modeled by the measurement function $h_{GNSS}(\mathbf{x}_{\mathrm{GNSS}}, \boldsymbol{s}_j)$ plus a zero-mean Gaussian noise term

$$\rho_j = h_{GNSS}(\mathbf{x}_{\mathrm{GNSS}}, \boldsymbol{s}_j) + \mathcal{N}(0, \sigma_{Sat_j}) \ , \tag{9}$$

from which we can derive the cost function for each individual pseudo-range observation:

$$g_{Sat_j}(\mathbf{x}_{\mathrm{GNSS}}) = (e_{Sat_j}(\mathbf{x}_{\mathrm{GNSS}})\sigma_{Sat_j}^{-1})^2 \ , \tag{10}$$

where

$$e_{Sat_j}(\mathbf{x}_{\mathrm{GNSS}}) = \rho_j - h_{GNSS}(\mathbf{x}_{\mathrm{GNSS}}, \boldsymbol{s}_j) \ , \tag{11}$$

with σ_{Sat_j} being the variance associated to the pseudo-range measurement ρ_j. Noticing that in this stage, although the cost is related to $\boldsymbol{t}_{rel}$, $\boldsymbol{t}_{\mathrm{o}}$, δ and $\boldsymbol{s}_j$, only $\boldsymbol{t}_{\mathrm{o}}$ and δ is considered to be the optimization variables here. Then, consequently, the Jacobian of e_{Sat_j} is given by:

$$\begin{aligned} \mathbf{J}(e_{Sat_j})|_{\mathbf{x}_{\mathrm{GNSS}}} &= \frac{\partial e_{Sat_j}}{\partial \mathbf{x}_{\mathrm{GNSS}}^{\mathsf{T}}} \\ &= \begin{bmatrix} \frac{s_{j_x} - t_{\mathrm{o}_x}}{||\boldsymbol{s}_j - \boldsymbol{t}_{rel} - \boldsymbol{t}_{\mathrm{o}}||_2} \\ \frac{s_{j_y} - t_{\mathrm{o}_y}}{||\boldsymbol{s}_j - \boldsymbol{t}_{rel} - \boldsymbol{t}_{\mathrm{o}}||_2} \\ \frac{s_{j_z} - t_{\mathrm{o}_z}}{||\boldsymbol{s}_j - \boldsymbol{t}_{rel} - \boldsymbol{t}_{\mathrm{o}}||_2} \\ 1 \end{bmatrix}^{\mathsf{T}} , \end{aligned} \tag{12}$$

which will be used in linearization. Given this cost function of GNSS measurement, the maximum a posterior solution for a single GNSS state $\mathbf{x}_{\mathrm{GNSS}}$ during the tracking stage is given by solving the following least squares problem:

$$\hat{\mathbf{x}}_{\mathrm{GNSS}} = \arg\min_{\boldsymbol{t}_{\mathrm{o}}, \delta} \sum_j g_{Sat_j}(\mathbf{x}_{\mathrm{GNSS}}) \ . \tag{13}$$

Noticing that Eq. (13) is a single-point positioning formulation and will only be preformed at the tracking stage, as will be detailed in the next subsection.

C. Vision-GNSS Joint Asynchronous Tracking

Before proceeding to the proposed asynchronous tracking, we first briefly review the graph-based ego-motion estimation in vision SLAM. Following the classic pinhole camera projection model [25], [28] , the measurement function of graph-based ego-motion estimation can be expressed as

$$\begin{aligned} \boldsymbol{p}_i &= h_{Cam}(\boldsymbol{l}_i^{\mathcal{B}}) + \mathcal{N}(0, \boldsymbol{\Sigma}_{Cam_i}) \\ &= \begin{bmatrix} f_x \frac{l_{i_x}^{\mathcal{B}}}{l_{i_z}^{\mathcal{B}}} + c_x \\ f_y \frac{l_{i_y}^{\mathcal{B}}}{l_{i_z}^{\mathcal{B}}} + c_y \end{bmatrix} + \mathcal{N}(0, \boldsymbol{\Sigma}_{Cam_i}) \ , \end{aligned} \tag{14}$$

where f and c are intrinsic camera parameters, $\boldsymbol{p}_i = [u \ v]^{\mathsf{T}}$ is the feature point measurement under the camera image plane frame, and $\boldsymbol{l}_i^{\mathcal{B}} \triangleq [l_{i_x}^{\mathcal{B}} \ l_{i_y}^{\mathcal{B}} \ l_{i_z}^{\mathcal{B}}]^{\mathsf{T}}$ is the 3-D position under the body frame of the i-th landmark $\boldsymbol{l}_i^{\mathcal{L}}$ projected by the 6-DoF camera pose $\mathbf{T}_{\mathcal{B}}^{\mathcal{L}}$. Then, we express the camera ego-motion state under the Lie Algebra tangent space $\mathfrak{se}3$, as specified in [29], [30], which is

$$\begin{aligned} \mathbf{x}_{\mathrm{Cam}} \triangleq \boldsymbol{\xi}_{\mathcal{B}}^{\mathcal{L}} &= \ln(\mathbf{T}_{\mathcal{B}}^{\mathcal{L}})^{\vee} \\ &= \ln\left(\begin{bmatrix} \mathbf{R}_{\mathcal{B}}^{\mathcal{L}} & \boldsymbol{t}_{\mathcal{B}}^{\mathcal{L}} \\ 0 & 1 \end{bmatrix}\right)^{\vee} \ , \end{aligned} \tag{15}$$

and the relations between $X_{l_i}^{Cam}$, X_{l_i} and $\mathbf{x}_{\mathrm{Cam}}$ are

$$\begin{aligned} \mathbf{T}_{\mathcal{B}}^{\mathcal{L}} &= \exp(\mathbf{x}_{\mathrm{Cam}}{}^{\wedge}) \\ \boldsymbol{l}_i^{\mathcal{B}} &= \mathbf{T}_{\mathcal{B}}^{\mathcal{L}\,-1} \boldsymbol{l}_i^{\mathcal{L}} \ . \end{aligned} \tag{16}$$

Here $(\cdot)^{\wedge}$ and $(\cdot)^{\vee}$, respectively, represent the skew-symmetric hat operator and its inverse, as defined in [29]. $\boldsymbol{l}_i^{\mathcal{L}}$ is the 3-D vision point under the local world frame $\mathcal{F}_{\mathcal{L}}$, which is updated in the mapping procedure as will be introduced in Section IV. The cost function on the vision edge is given by

$$g_{Cam_i}(\mathbf{x}_{\mathrm{Cam}}) = ||\boldsymbol{e}_{Cam_i}(\mathbf{x}_{\mathrm{Cam}})||^2_{\boldsymbol{\Sigma}_{Cam_i}} \ , \tag{17}$$

and

$$\boldsymbol{e}_{Cam_i}(\mathbf{x}_{\mathrm{Cam}}) = \boldsymbol{p}_i - h_{Cam}(\mathbf{x}_{\mathrm{Cam}}, \boldsymbol{l}_i^{\mathcal{L}}) \ , \tag{18}$$

where $||\boldsymbol{e}||_{\boldsymbol{\Sigma}} \triangleq \boldsymbol{e}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{e}$ represents the squared Mahalanobis distance with covariance matrix Σ. Also, we should noticing that the $\boldsymbol{l}_i^{\mathcal{L}}$ is not included in the optimization variables here. Then the optimization formulation on the state vector $\mathbf{x}_{\mathrm{Cam}}$ is

$$\hat{\mathbf{x}}_{\mathrm{Cam}} = \arg\min_{\mathbf{x}_{\mathrm{Cam}}} \sum_i g_{Cam_i}(\mathbf{x}_{\mathrm{Cam}}) \ . \tag{19}$$

Solving the optimization problem involves computing the derivative with respect to the 3-D rotation parameters. The corresponding Jacobian used for the disturbance increment is

$$\begin{aligned} &\mathbf{J}(\boldsymbol{e}_{Cam_i})|_{\mathbf{x}_{\mathrm{Cam}}} = \\ &-\begin{bmatrix} \frac{f_x}{z'} & 0 & -\frac{f_x x'}{z'^2} & -\frac{f_x x' y'}{z'^2} & f_x + \frac{f_x x'^2}{z'^2} & -\frac{f_x y'}{z'^2} \\ 0 & \frac{f_y}{z'} & -\frac{f_y y'}{z'^2} & -f_y - \frac{f_y y'^2}{z'^2} & -\frac{f_y x' y'}{z'^2} & -\frac{f_y x'}{z'^2} \end{bmatrix} . \end{aligned} \tag{20}$$

Here, $l_{i_x}^{\mathcal{B}}$, $l_{i_y}^{\mathcal{B}}$ and $l_{i_z}^{\mathcal{B}}$ are replaced by x', y' and z' for concision.

Next, we present the proposed joint asynchronous tracking method. Although the synchronization in a multi-sensor system can be achieved by using dedicated hardware with external-trigger support, it would be costly. A low-cost strategy for synchronization is to stamp all the measurements with a high-precision timing device and process the stamped data asynchronously. To avoid accuracy degradation, an asynchronous processing pipeline needs to be designed with consideration of each sensor' trait.

In our implementation, both camera and GNSS signals will trigger the tracking procedure to build up the related single point pose graph on their own state components. In this stage, the GNSS part will trust the relative motion under the local world frame and only update the local-to-global origin $\boldsymbol{t}_{\mathrm{o}}$. On the vision part, tracking will utilize the change of $\boldsymbol{t}_{\mathrm{o}}$ to rectify the relative motion. This procedure is shown in Fig. 6.

Specifically, we first define the ego-motion state of the vehicle as the combination states of all the sensors, $\mathbf{x}_{\mathrm{V}} = [\boldsymbol{\xi}_{\mathcal{B}}^{\mathcal{L}\mathsf{T}}\ \boldsymbol{t}_{\mathrm{o}}^{\mathsf{T}}\ \delta]^{\mathsf{T}}$. Then, the initial value for tracking is calculated from two parts, the reference keyframe state $\mathbf{x}_{\mathrm{V_r}}$ and the inter-keyframe motion accumulation $\Delta\mathbf{x}_{\mathrm{V}}$, where

$$\mathbf{x}_{\mathrm{V_r}} = \begin{bmatrix} \boldsymbol{\xi}_{\mathcal{B}_{\mathrm{r}}}^{\mathcal{L}} \\ \boldsymbol{t}_{\mathrm{or}} \\ \delta_r \end{bmatrix}, \tag{21}$$

and

$$\Delta\mathbf{x}_{\mathrm{V}} = \begin{bmatrix} \boldsymbol{\xi}_{\mathcal{B}_{\mathrm{c}}}^{\mathcal{B}_{\mathrm{r}}} \\ \boldsymbol{t}_{\mathrm{oc}} \\ \delta_c \end{bmatrix}. \tag{22}$$

Here the footnotes, $(\cdot)_r$ for reference frame and $(\cdot)_c$ for current frame, indicate the frame type of the state being subjected to.

Then, if the current incoming signal is from GNSS, the initial value (represented by $\check{(\cdot)}$) is composed by the following local relative parts

$$\begin{aligned} \begin{bmatrix} \check{\boldsymbol{R}}_{\mathcal{B}}^{\mathcal{L}} & \check{\boldsymbol{t}}_{\mathcal{B}}^{\mathcal{L}} \\ 0 & 1 \end{bmatrix} &= \exp((\boldsymbol{\xi}_{\mathcal{B}_{\mathrm{r}}}^{\mathcal{L}} \boxplus \boldsymbol{\xi}_{\mathcal{B}_{\mathrm{c}}}^{\mathcal{B}_{\mathrm{r}}})^{\wedge}) \\ \check{\boldsymbol{t}}_{rel} &= \mathbf{R}_{\mathcal{G}}^{\mathcal{W}} \mathbf{R}_{\mathcal{L}}^{\mathcal{G}} \check{\boldsymbol{t}}_{\mathcal{B}}^{\mathcal{L}}, \end{aligned} \tag{23}$$

where $\boxplus$ represents the plus oration on $\mathfrak{se}3$ as introduced in [29]. Then we neglect the change in the absolute global origin $\boldsymbol{t}_{\mathrm{o}}$ and the clock bias δ, take $\check{\boldsymbol{t}}_{\mathrm{o}} = \boldsymbol{t}_{\mathrm{or}}$ and $\check{\delta} = \delta_r$, to form the GNSS initial state of $\check{\mathbf{x}}_{\mathrm{GNSS}} = [(\check{\boldsymbol{t}}_{rel} + \check{\boldsymbol{t}}_{\mathrm{o}})^{\mathsf{T}}, \check{\delta}]^{\mathsf{T}}$. Correspondingly, if the incoming data is vision image, then the initial value is the compensated relative motion

$$\check{\mathbf{x}}_{\mathrm{Cam}} = \boldsymbol{\xi}_{\mathcal{B}_{\mathrm{r}}}^{\mathcal{L}} \boxplus \boldsymbol{\xi}_{\mathcal{B}_{\mathrm{c}}}^{\mathcal{B}_{\mathrm{r}}} \boxplus \ln\left(\begin{bmatrix} \mathbf{I} & \boldsymbol{t}_{\mathrm{oc}} - \boldsymbol{t}_{\mathrm{or}} \\ 0 & 1 \end{bmatrix}\right)^{\vee}. \tag{24}$$

D. Keyframe Selection

The goal of the tracking procedure mentioned in the last subsection is to provide a preliminary estimation of the vehicle state for each individual sensor frame. Then, for keyframe selection in the case of mixed heterogeneous sensors, we take the vision camera as the prime sensor, and use the strategy shown in Fig. 7 to select keyframes. As long as the vision camera remain tracking, the criterion for vision keyframe selection is the same as that in [12]. Specifically, when a vision-predominant keyframe is created, the latest GNSS observation is checked to be included or not as follows. If the last GNSS observation time is close enough, e.g., $t_{CurrentCamera} - t_{LatestGNSS} \triangleq \delta_f < \delta_{Thershold}$, the GNSS observation is still trustable and, hence, is counted in as a persistence measurement, which is the Case (a) in Fig. 7. Otherwise, it is Case (b) in Fig. 7: we create a stand-alone normal vision keyframe. $\delta_{Thershold}$ here is determined by the prior on the motion speed of the vehicle and the sensor accuracy. Here we set $\delta_{Thershold} = 0.3$ s, as the linear speed of our vehicle is assumed to be less than 1 m/s. In this condition, the error is ensured to be smaller than 0.3 m, which is acceptable compared with the 5 m level accuracy of a typical GNSS receiver.

On the other hand, at receiving a GNSS measurement, a keyframe is generated immediately along with the latest vision frame as long as it is not already a keyframe, which is the Case (c) in Fig. 7. Otherwise, if that vision frame is already a keyframe or the vision is lost of tracking, then, a keyframe only contains GNSS observation will be created, which is the Case (d) in Fig. 7.

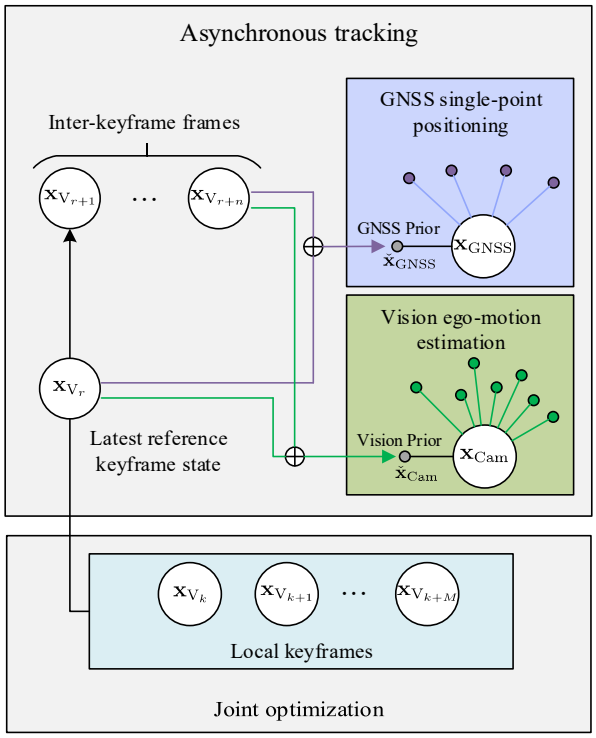


Fig. 6. Graph model based asynchronous tracking. The ego-motion estimation is reckoned based on the latest reference keyframe state $\mathbf{x}_{\mathrm{V_r}}$, whose motion state is optimized by joint bundle adjustment. Then, before next keyframe is created, several data frames (e.g., n frames) of the sensors will be processed.

IV. Joint 10-DoF Graph-Optimization on Manifold

The back-end optimization in the mapping stage is a joint optimization procedure, which is also known as the bundle adjustment (BA) in the V-SLAM community. In this section, we first propose a joint optimization model, then, the derivative of this model on manifold is explicitly derived.

As introduced in the last section, we know that the tracking is operated on the state space of each individual sensor: $\mathbf{x}_{\mathrm{Cam}} = \boldsymbol{\xi}_{\mathcal{B}}^{\mathcal{L}}$ and $\mathbf{x}_{\mathrm{GNSS}} = [(\boldsymbol{t}_{rel} + \boldsymbol{t}_{\mathrm{o}})^{\mathsf{T}}\ \delta]^{\mathsf{T}}$ as defined in Eq. (7). Accordingly, the whole vehicle ego-motion state space $\mathbf{x}_{\mathrm{V}}$ at the k-th frame is

$$\mathbf{x}_{\mathrm{V}_k} \triangleq [\boldsymbol{\xi}_{\mathcal{B}_k}^{\mathcal{L}\ \mathsf{T}}\ \boldsymbol{t}_{ok}^{\mathsf{T}}\ \delta_k]^{\mathsf{T}} \in \mathbb{R}^{10}, \tag{25}$$

where $\boldsymbol{\xi}_{\mathcal{B}_k}^{\mathcal{L}} = \ln(\mathbf{T}_{\mathcal{B}_k}^{\mathcal{L}})^{\vee}$ is the 6-DOF ego-motion state under the local world frame $\mathcal{F}_{\mathcal{L}}$, and $\boldsymbol{t}_{rel_k}$ used in $\mathbf{x}_{\mathrm{GNSS}}$ is also included in $\boldsymbol{\xi}_{\mathcal{B}_k}^{\mathcal{L}}$. As the IMU is not included, $\mathbf{T}_{\mathcal{L}}^{\mathcal{G}}$ is fixed. Hence, the free variables, including $\mathbf{R}_{\mathcal{B}_k}^{\mathcal{L}}$, $\boldsymbol{t}_{\mathcal{B}_k}^{\mathcal{L}}$, $\boldsymbol{t}_{\mathrm{o}}$ and δ_k, generate a state space of 10 DoF. Additionally, in the BA

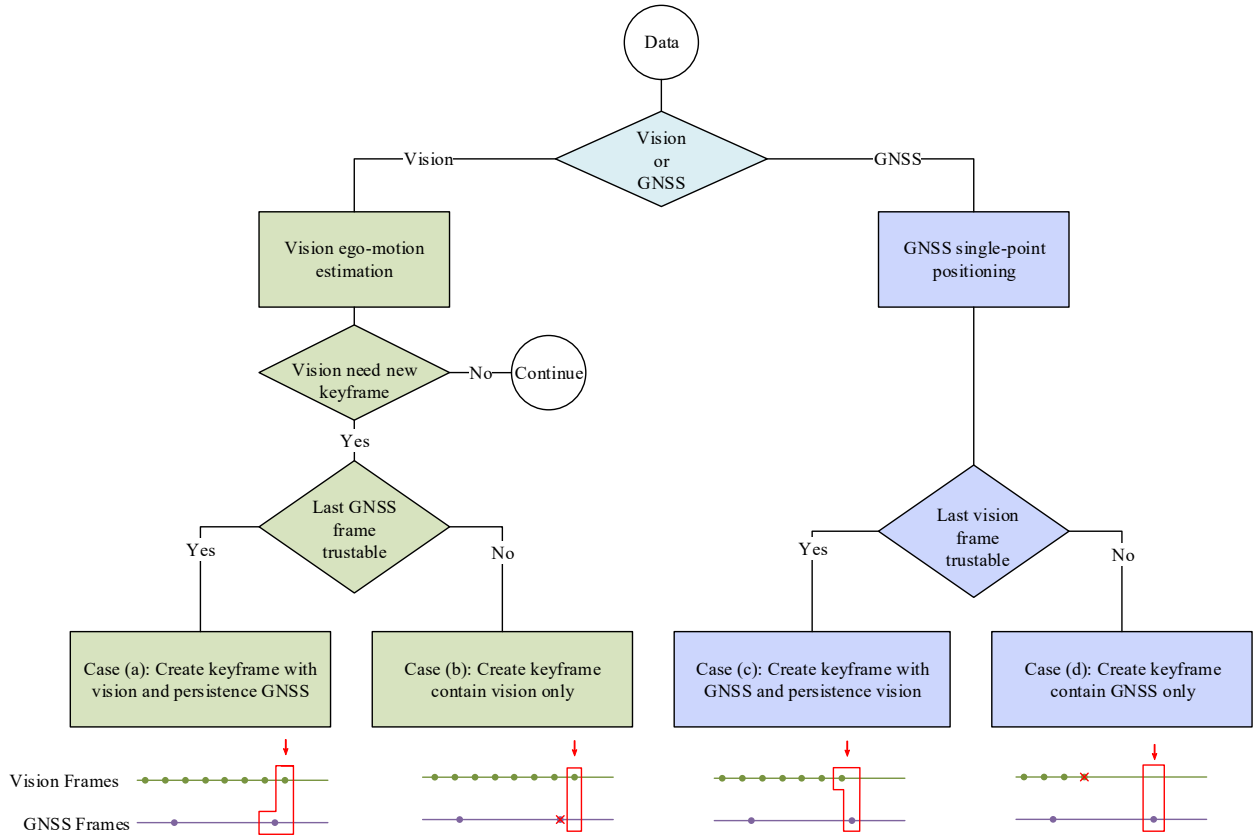


Fig. 7. Keyframe selection strategy.

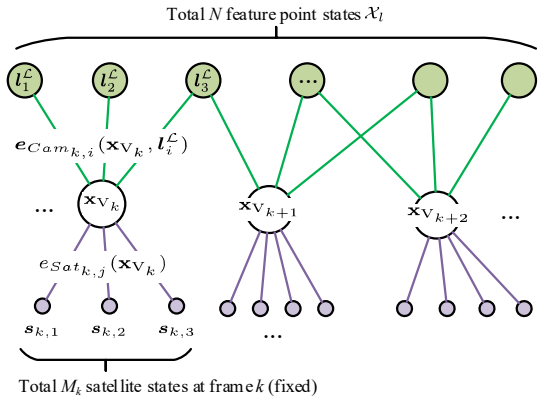


Fig. 8. Joint optimization graph model. The measurement error in each frame V_k is composed by two parts: the vision part that involves feature point states $\boldsymbol{l}_i^{\mathcal{L}}$ and $\mathbf{x}_{V_k}$ and the GNSS satellite part that only involves $\mathbf{x}_{V_k}$. Noticing that the satellite states here is fixed and will not optimized.

procedure, we also need to optimize the positions of all the vision feature points under the local world frame. Assuming there are totally N points in the local map, then, for the i-th landmark point, the position under the local world frame $\mathcal{F}_L$ is

$$\boldsymbol{l}_i^{\mathcal{L}} = [l_{i_x}^{\mathcal{L}} \; l_{i_y}^{\mathcal{L}} \; l_{i_z}^{\mathcal{L}}]^{\mathsf{T}} \; . \tag{26}$$

This results in the entire state space for the BA as

$$\mathbf{x}_{\mathrm{BA}} = \begin{bmatrix} \mathcal{X}_V \\ \mathcal{X}_l \end{bmatrix}, \tag{27}$$

where $\mathcal{X}_V = [\mathbf{x}_{V_1}^{\mathsf{T}} \; \cdots \; \mathbf{x}_{V_k}^{\mathsf{T}} \; \cdots \; \mathbf{x}_{V_K}^{\mathsf{T}}]^{\mathsf{T}}$ is the vector containing all K ego-motion states in the BA process, while $\mathcal{X}_l = [\boldsymbol{l}_1^{\mathcal{L}\mathsf{T}} \; \cdots \; \boldsymbol{l}_i^{\mathcal{L}\mathsf{T}} \; \cdots \; \boldsymbol{l}_N^{\mathcal{L}\mathsf{T}}]^{\mathsf{T}}$ is the set of the feature points.

Then, the next step is to combine the cost functions introduced in Section III to construct a joint optimization formulation on graph. In Section III, we have already introduced the cost function $g_{Sat_j}(\mathbf{x}_{\mathrm{GNSS}})$ and $g_{Cam_i}(\mathbf{x}_{\mathrm{Cam}})$ on the tracking stage. For the joint optimization stage, these cost functions remain the same excepted for the changes on variable states. Specifically, for GNSS cost function in Eq. (10), we only concern about $\boldsymbol{t}_{\mathrm{o}}$ and δ. Then, in this stage, we take $\boldsymbol{t}_{rel}$ (which is a function of $\boldsymbol{\xi}_{\mathcal{B}}^{\mathcal{L}}$ as shown in Eq. (7) and Eq. (15)) into consideration, leading to

$$g_{Satk,j}(\mathbf{x}_{V_k}) = (e_{Sat_{k,j}}(\mathbf{x}_{V_k})\sigma_{Sat_{k,j}}^{-1})^2 \; , \tag{28}$$

where

$$e_{Sat_{k,j}}(\mathbf{x}_{\mathrm{V}_k}) = \rho_{k,j} - h_{GNSS}(\mathbf{x}_{\mathrm{V}_k}, \boldsymbol{s}_{k,j}) \ , \tag{29}$$

and

$$\begin{aligned} h_{GNSS}(\mathbf{x}_{\mathrm{V}_k}, \boldsymbol{s}_{k,j}) =& h_{GNSS}(\boldsymbol{t}^{\mathcal{L}}_{\mathcal{B}_k}, \boldsymbol{t}_{\mathrm{o}_k}, \delta_k, \boldsymbol{s}_{k,j}) \\ =& ||\boldsymbol{s}_{k,j} - (\boldsymbol{t}^{\mathcal{L}}_{\mathcal{B}_k} + \boldsymbol{t}_{\mathrm{o}k})||_2 + \delta_k + \delta_{Sat} \\ &+ \delta_{Atmosphere} + \delta_{EarthRotation} \ . \end{aligned} \tag{30}$$

Then for the of vision measurement, we have

$$g_{Cam_{k,i}}(\mathbf{x}_{\mathrm{V}_k}, \boldsymbol{l}^{\mathcal{L}}_i) = ||\boldsymbol{e}_{Cam_{k,i}}(\mathbf{x}_{\mathrm{V}_k}, \boldsymbol{l}^{\mathcal{L}}_i)||^2_{\boldsymbol{\Sigma}_{Cam_i}} \ , \tag{31}$$

and

$$\boldsymbol{e}_{Cam_{k,i}}(\mathbf{x}_{\mathrm{V}_k}, \boldsymbol{l}^{\mathcal{L}}_i) = \boldsymbol{p}_i - h_{Cam}(\mathbf{x}_{\mathrm{V}_k}, \boldsymbol{l}^{\mathcal{L}}_i) \ . \tag{32}$$

The corresponding graph model is shown in Fig. 8.

Combining the optimization problems Eq. (13) and Eq. (19) and substituting the cost function to Eq. (28) and Eq. (31), we have the optimization formulation for the BA procedure as

$$\begin{aligned} \hat{\mathbf{x}}_{\mathrm{BA}} = \arg\min_{\mathbf{x}_{\mathrm{BA}}} \Bigg(& \sum_{i}^{N} g_{Cam_{k,i}}(\mathbf{x}_{\mathrm{V}_k}, \boldsymbol{l}^{\mathcal{L}}_i) + \\ & \sum_{k}^{K} \sum_{j}^{M_k} g_{Sat_{k,j}}(\mathbf{x}_{\mathrm{V}_k}) \Bigg) \ , \end{aligned} \tag{33}$$

where N is the number of landmark observations, and M_k is the number of satellite measurements in the k-th frame. Then, the corresponding Jacobian is composed by the 2-by-$(10 \times K + 3 \times N)$ matrix $[\mathbf{J}(\boldsymbol{e}_{Cam_{k,i}})|_{\mathcal{X}_V}, \mathbf{J}(e_{Sat_{k,j}})|_{\mathcal{X}_l}]$ and the 1-by-$(10 \times K \times M_k)$ matrix $\mathbf{J}(e_{Sat_{k,j}})|_{\mathcal{X}_V}$, which are given by

$$\begin{aligned} &[\mathbf{J}(\boldsymbol{e}_{Cam_{k,i}})|_{\mathcal{X}_V}, \mathbf{J}(\boldsymbol{e}_{Cam_{k,i}})|_{\mathcal{X}_l}] = \\ &- \Bigg[\mathbf{0}_{2\times 10}, \cdots, \mathbf{0}_{2\times 10}, \left[\frac{\partial \boldsymbol{e}_{Cam_{k,i}}}{\partial \boldsymbol{\xi}^{\mathcal{L}^{\mathsf{T}}}_{\mathcal{B}_k}}, \mathbf{0}_{2\times 4} \right], \cdots, \mathbf{0}_{2\times 10}, \cdots, \\ &\quad \mathbf{0}_{2\times 3}, \cdots, \mathbf{0}_{2\times 3}, \frac{\partial \boldsymbol{e}_{Cam_{k,i}}}{\partial \boldsymbol{l}^{\mathcal{L}^{\mathsf{T}}}_i}, \cdots, \mathbf{0}_{2\times 3} \Bigg] , \end{aligned} \tag{34}$$

and

$$\begin{aligned} &\mathbf{J}(e_{Sat_{k,j}})|_{\mathcal{X}_V} = \\ &- \left[\mathbf{0}_{1\times 10}, \cdots, \mathbf{0}_{1\times 10}, \frac{\partial e_{Sat_{k,j}}}{\partial \mathbf{x}^{\mathsf{T}}_{\mathrm{V}_k}}, \cdots, \mathbf{0}_{1\times 10} \right] , \end{aligned} \tag{35}$$

where $\frac{\partial \boldsymbol{e}_{Cam_{k,i}}}{\partial \boldsymbol{\xi}^{\mathcal{L}^{\mathsf{T}}}_{\mathcal{B}_k}}$ is given in Eq. (20), and

$$\frac{\partial \boldsymbol{e}_{Cam_{k,i}}}{\partial \boldsymbol{l}^{\mathcal{L}^{\mathsf{T}}}_i} = - \begin{bmatrix} \frac{f_x}{z'} & 0 & -\frac{f_x x'}{z'^2} \\ 0 & \frac{f_y}{z'} & -\frac{f_y y'}{z'^2} \end{bmatrix} \mathbf{R}^{\mathcal{L}}_{\mathcal{B}_k} \tag{36}$$

is the derivative with respect to the vision point position state. Then, for the derivative on the satellite measurement $\frac{\partial e_{Sat_{k,j}}}{\partial \mathbf{x}^{\mathsf{T}}_{\mathrm{V}_k}}$, we cannot use Eq. (12) and Eq. (13) directly, since at the BA stage the optimizer will optimize all the 10-DoF vehicle states upon the satellite measurement, instead of just adjusting the global origin $\boldsymbol{t}_{\mathrm{o}}$. Hence, we need to analyze how the relative ego-motion components affect the error function e_{Sat}. The corresponding Jacobian is given as below (derived in Appendix A):

$$\begin{aligned} \frac{\partial e_{Sat_{k,j}}}{\partial \mathbf{x}_{\mathrm{V}_k}} =& [d_x, d_y, d_z, \\ & - t_z d_y + t_y d_z, \\ & t_z d_x - t_x d_z, \\ & - t_y d_x + t_x d_y, \\ & d_x, d_y, d_z, 1] \ , \end{aligned} \tag{37}$$

where

$$[d_x, d_y, d_z] \triangleq \begin{bmatrix} \frac{s_x - t_x}{||\boldsymbol{s}_{k,j} - \boldsymbol{t}^{\mathcal{W}}_{\mathcal{B}_k}||_2} \\ \frac{s_y - t_y}{||\boldsymbol{s}_{k,j} - \boldsymbol{t}^{\mathcal{W}}_{\mathcal{B}_k}||_2} \\ \frac{s_z - t_z}{||\boldsymbol{s}_{k,j} - \boldsymbol{t}^{\mathcal{W}}_{\mathcal{B}_k}||_2} \end{bmatrix}^{\mathsf{T}} , \tag{38}$$

and $\boldsymbol{t}^{\mathcal{W}}_{\mathcal{B}_k} \triangleq [t_x \ t_y \ t_z]^{\mathsf{T}}$, $\boldsymbol{s}_{k,j} \triangleq [s_x \ s_y \ s_z]^{\mathsf{T}}$ for concision.

By combining the vision and GNSS optimization objectives, the goal of the proposed fusion method is to jointly optimize the 10-DoF state parameters based on both the vision measurement and GNSS measurement. This method can improve the accuracy on both the vision and GNSS sides. On the vision side, the accumulate position error can be eliminated with the aid of the GNSS measurement, as the pseudo-range measurement does not suffer from accumulate error. This advantage is particularly considerable in the applications involving long-term movement. On the GNSS side, as the relative motion estimation is done by the more precise vision-based method, a better initial value can be obtained in calculating the global position, which is beneficial for improving the final accuracy.

Additionally, this batch-estimation for global optimization does not introduce excessive computation load compared with the original system. As the computation complexity of a SLAM system is composed by lots of factors [31], here we compare only the main factor, which is the constraints.

From Eq. (33), we notice that there are N constraints created by vision measurement and $K \times M_k$ constraints created by the additional GNSS measurement. Then the overall computation complexity would increase by $1 + (K \times M_k)/N$. For a typical vision SLAM local map of 100 keyframes, the number of map points N has a typical value of 5000, meanwhile each keyframe has a typical number of 10 observations on satellites. As mentioned in Section II, we consider the satellites observations as special map points. Then, for the proposed method, there are about 20% of extra "feature points" being added into each keyframe. This amount of additional constraints, on one hand, can effectively alleviate the bias problem of vision SLAM. One the other hand, it keeps a typical computation complexity increment of around 20%, which maintains a low and constant magnification compared with the original method.

V. Implementation and Experimental Setup

A. Software Implementation

The system is modified based on the ORB-SLAM2 system [20]. The general graphical optimization (g^{2}o) tool [26] is used in the optimization procedure. With GNSS pseudo-range nodes being created and inserted into the optimization graph with a higher dimension of freedom, we are able to impose different sensor constraints into a single graph model. We also implement an extended Kalman filter (EKF) based loosely-coupled method [6] for comparison.

Further, to organize and synchronize multiple sensors in our system, the Robot Operating System (ROS) is employed. ROS is an open-source robotics middleware, which provides the implementation of message-passing between processes as well as the synchronization utilities.

B. System Setup

To evaluate the proposed method, we setup an autonomous ground robot based on the Pioneer® 3-AT ground vehicle, equipped with a stereo camera with dual VGA resolution along with a GNSS receiver from u-blox® M8P running under rover mode. This receiver provides both the raw GNSS observations and the RTK ground truth by connecting to a differential station through Ethernet. Further, a MPU9250-based 9-axis IMU module is attached on the stereo camera to align the vision coordinate with the ENU coordinate of GNSS. All algorithms are performed on an on-board Intel® NUC computer with a Core®-i7 CPU. This robot is shown in Fig. 9.

The calibration of this system includes three parts: the intrinsic and extrinsic parameters estimation of the stereo camera, IMU-camera extrinsic parameters, and the calibration of the camera with GNSS antenna. For the first one, we follow a classical chess-board calibration method [32] to estimate both the external and internal parameters of the stereo camera. For the IMU-camera extrinsics calibration, we use the Kalibr toolbox mentioned in [33]. Then for the antenna-to-camera calibration, we use an external vision-based motion capture system Vicon® and the reflectors on the sensors, as shown in Fig. 9, to estimate the relative translation of these two sensors. After these steps, all the sensors are calibrated to the left camera.

C. Experiment Setup

The proposed algorithm has been tested under two different circumstances, including:

- In the middle of road with open sky: This circumstance gives rise to scale drift of vision SLAM systems because of the relatively small baseline, see Fig. 11 (a). The total travel distance for this case is 204.15 m.
- Beside buildings: This circumstance challenges GNSS due to the existence of signal occlusion and reflection, see Fig. 11 (b). The total travel distance for this case is 537.29 m.

In the first experiment, the differential GNSS result from the ublox® M8P receiver is taken as the ground truth. In the second experiment, as our test setup involve diverse environment in some conditions, the GNSS signal can be obstructed and becomes unreliable. To address this problem, we further setup nine check points by averaging the measurements of a NovAtel® DGPS base station over 10 minutes at each check point, as shown in Fig. 10. Further, to compare the algorithms operating under different coordinate systems, all the localization results are transferred into the local ENU coordinate system with an aligned origin.

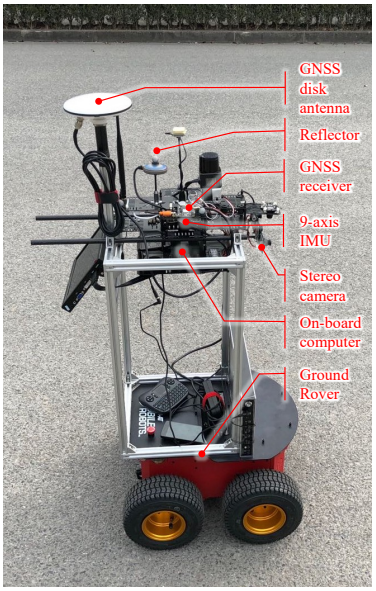


Fig. 9. The ground robot used in the experiment.

Fig. 10. The base station for accurate localization of the check points.

VI. Result and Analysis

A. Test Scenario One

Firstly, we evaluate the test case one, which is the open sky trajectory. In this scenario, the GNSS data is collected under an interference-free environment, hence the ground truth can be trusted through the entire procedure. The localization results are shown in Fig. 12, where we represent the trajectories under

(a) Test scenario 1: In the middle of road with open sky

(b) Test scenario 2: Beside buildings

Fig. 11. Experiment scenarios and trajectories (red lines). The start position is represented by a circle and the end point by a square. The right column is snapshot from the camera's perspective to illustrate the test environments.

both the local world frame $\mathcal{F}_{\mathcal{G}}$ and the global frame $\mathcal{F}_{\mathcal{W}}$. The vision-only trajectory (in green) is generated by ORB-SLAM2, while the GNSS-only trajectory (in cyan) is generated by the single-point pseudo-range optimization method as mentioned in Section III-B.

In the main scene of this experiment, most of the textured objects in sight from the robot are more than 10 meters away, whereas the baseline of the employed stereo camera is only 120 mm. This leads to a large uncertainty in depth estimation of the scene and makes the stereo camera degenerating to a monocular one. Accordingly, it can be seen from Fig. 12 that, the vision-only trajectory has a significant drift, which increases over time. As shown in Table I, the maximal localization error of the vision navigation system is 19.03 m. With the total travel distance being 204.15 m, the maximal drift error is about 9.32% of the total distance. As expected, when the GNSS constraint is applied in the system, the maximal drift error is significantly alleviated to 6.79 m, which is 3.32% of the total distance. Fig. 13 shows the localization error versus time. It can be seen that, as the localization error of the vision-only method increases with time, the localization error of the proposed tight fusion method does not increase with time. This advantage of the proposed method can be more conspicuous as the travel distance increases.

In this scenario, the EKF fusion algorithm presents a typical behavior: When both of the sources provide similar uncertainty (before 150s as shown in Fig. 12), it tends to trust the sources

TABLE I. Localization errors comparison in test case one.

	Vision	GNSS	EKF	Proposed method
RMSE (m)	11.77	7.29	5.71	4.72
Mean Error (m)	10.56	6.86	5.36	4.47
Max Error (m)	19.03	11.50	9.71	6.79

evenly; Then, with vision localization result drifts over time, the fusion result is partial to the GNSS result. Compared with the filter-based fusion, the proposed method also shows advantage, with lower RMSE and mean error, benefited from the global optimal estimation.

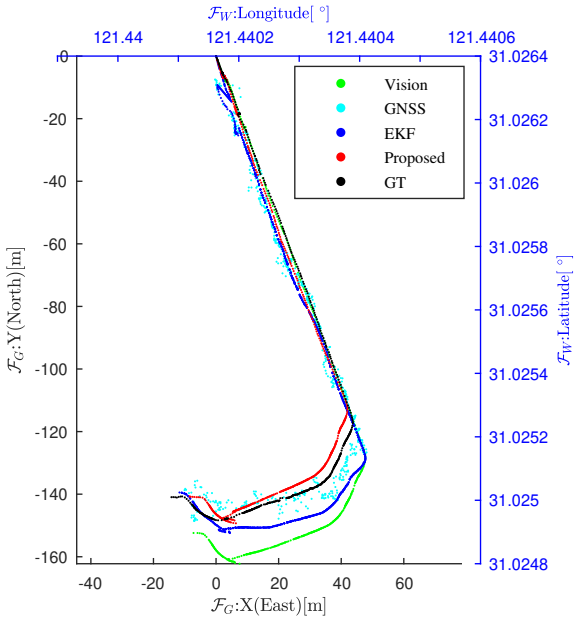


Fig. 12. Localization result comparison in test case one.

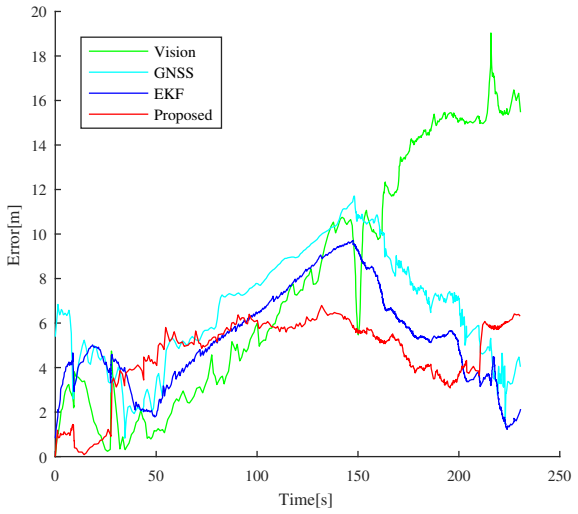


Fig. 13. Localization error comparison in test case one.

B. Test Scenario Two

The circumstance in the second experiment has a different effect on each sensor compared with that in the first experiment. Specifically, in the main scene of this experiment,

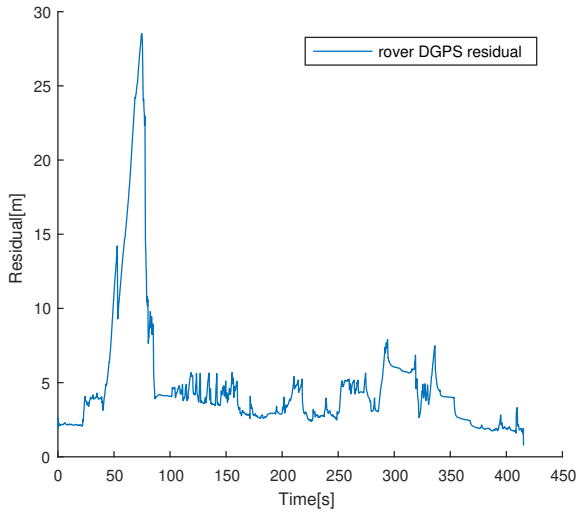


Fig. 14. Rover DGPS localization residual in test case two. As the mean residual reaches 4.854 m and the max residual is 28.525 m, we no longer consider it as a ground truth, but only a reference path to illustrate the expected trajectory.

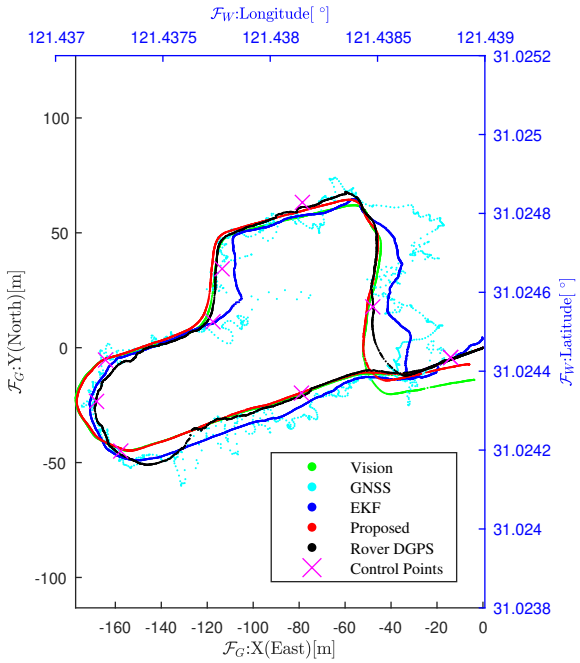


Fig. 15. Test case two localization result with control points.

most of the vision textured features fit in well with the camera baseline setup, as the path is near buildings and landmarks. This condition facilitates reliable depth estimation of the scene in the stereo vision. Meanwhile, the GNSS signal is substantially obstructed and reflected by the surrounding buildings. It significantly degrades the accuracy of the rover DGPS, as shown in Fig. 14. In this case, the rover DGPS is no longer reliable and cannot be taken as the ground truth. The trajectory results of the compared methods are shown in Fig. 15, while Fig. 16 and Table II present detailed error comparison with respect to the control points.

Due to signal obstruction and reflection, the single-point GNSS pseudo-range solution has the largest error, with the maximum error and the root-mean-square error (RMSE) being 37.92 m and 14.73 m, respectively. In this condition, the filter-based loose-coupled method, which directly uses the GNSS positioning result, cannot yield satisfactory performance. Contrast to the results in the first experiment, the vision navigation system yields better performance than GNSS, with the maximum error of 12.84 m and the RMSE of 6.64 m. The maximum relative drift error is 2.38%, which is significantly reduced compared with that in the first experiment. The proposed fusion method gives the best overall performance, with the maximum error of 8.67 m and the mean error of 5.02 m. It achieves a relative error of 1.6%. Moreover, benefited from the vision constraint and the synergistic convergence effect of the tight fusion, the global localization RMSE is reduced to 5.55 m, while those of the GNSS and the EKF fusion methods are 18.39 m and 7.24 m, respectively.

During the entire path, the vision provides the dominant contribution, since the GNSS pseudo-range measurements are significantly biased. However, in some parts of the path with good signal conditions, the constraint from the GNSS measurements can rectify the accumulated drift in vision and improve the overall accuracy. This effect can be seen in the

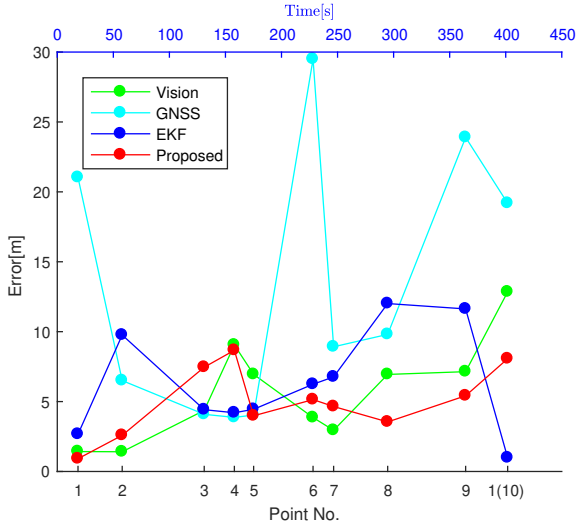


Fig. 16. Localization error comparison in test case two with control points.

TABLE II. Localization error comparison in test case two at the control points.

	Vision	GNSS	EKF	Proposed method
RMSE (m)	6.64	18.39	7.24	5.55
Mean Error (m)	5.70	14.73	6.31	5.02
Max Error (m)	12.84	37.92	12.01	8.67

latest 100 seconds of the test, where the fused result has a smaller drift and is more close to the result of the control points.

Additionally, the average runtime of the proposed method for optimizing a local map is about 1088 ms while the original vision method is about 889 ms, which shows an increment of about 22.4%. This verifies the computation analysis in Section IV.

VII. Conclusion

This paper proposed a SLAM-based fusion method, which fuses GNSS pseudo-range measurements with the vision feature points by utilizing a motion-decomposition-based asynchronous tracking to accommodate the heterogeneous sensor observations. To achieve tight fusion, a joint vision-GNSS 10-DoF optimization formulation on manifold is proposed to refine the localization result at the back-end. The fusion of GNSS into V-SLAM not only provides constraints to rectify the accumulation error in V-SLAM, but also provides the global localization result. Meanwhile, the locally high-accuracy vision measurements can complementarily aid GNSS to mitigate the local disturbance error. The new algorithm has been implemented in a real-time on-board software on a ground robot and verified through two experiments with different scenarios. The results demonstrated that, the proposed fusion method can significantly benefits the localization accuracy of the system, and also presents a better performance than the loosely coupled method. The improvement over the vision only method is especially significant in an environment which is not conducive to vision sensors (e.g., test case one in the experiment).

Appendix
GNSS satellite pseudo-range measurement Jacobian on manifold

In the tracking stage with the GNSS state defined as Eq. (7), the $\boldsymbol{t}_{rel}$ part is determined only. Oppositely, when it comes to BA stage, the GNSS state will take both of the relative motion and absolute origin into consideration, as defined in Eq. (6), which we rewrite here:

$$\mathbf{x}_{\mathrm{GNSS}} \triangleq [\boldsymbol{t}_{\mathcal{B}}^{\mathcal{W}\mathsf{T}} \ \delta]^{\mathsf{T}} \in \mathbb{R}^4 \ , \tag{39}$$

with

$$\boldsymbol{t}_{\mathcal{B}}^{\mathcal{W}} = \mathbf{R}_{\mathcal{G}}^{\mathcal{W}}\mathbf{R}_{\mathcal{L}}^{\mathcal{G}}\boldsymbol{t}_{\mathcal{B}}^{\mathcal{L}} + \boldsymbol{t}_{\mathrm{o}} \ , \tag{40}$$

where $\boldsymbol{t}_{\mathcal{B}}^{\mathcal{L}}$ is related with $\boldsymbol{\xi}_{\mathcal{B}}^{\mathcal{L}}$ in Eq. (25). In this case, the equation Eq. (12) is modified to

$$\begin{aligned}
\mathbf{J}(e_{Sat_{k,j}})|_{\mathbf{x}_{\mathrm{V}_k}} &= \frac{\partial e_{Sat_{k,j}}}{\partial \boldsymbol{x}_{\mathrm{V}_k}^{\mathsf{T}}} \\
&= \frac{\partial e_{Sat_{k,j}}}{\partial \mathbf{x}_{\mathrm{GNSS}_k}^{\mathsf{T}}} \frac{\partial \mathbf{x}_{\mathrm{GNSS}_k}}{\partial \boldsymbol{x}_{\mathrm{V}_k}^{\mathsf{T}}} \\
&= \frac{\partial e_{Sat_{k,j}}}{\partial [\boldsymbol{t}_{\mathcal{B}_k}^{\mathcal{W}\mathsf{T}} \ \delta_k]} \frac{\partial [\boldsymbol{t}_{\mathcal{B}_k}^{\mathcal{W}\mathsf{T}} \ \delta_k]^{\mathsf{T}}}{\partial [\boldsymbol{\xi}_{\mathcal{B}_k}^{\mathcal{L}\mathsf{T}} \ \boldsymbol{t}_{\mathrm{o}_k}^{\mathsf{T}} \ \delta_k]} \\
&= \left[\frac{\partial e_{Sat_{k,j}}}{\partial \boldsymbol{t}_{\mathcal{B}_k}^{\mathcal{W}\mathsf{T}}} \ 1 \right] \begin{bmatrix} \frac{\partial \boldsymbol{t}_{\mathcal{B}_k}^{\mathcal{W}}}{\partial \boldsymbol{\xi}_{\mathcal{B}_k}^{\mathcal{L}\mathsf{T}}} & \mathbf{I}_{3\times 3} & \mathbf{0}_{3\times 1} \\ \mathbf{0}_{1\times 6} & \mathbf{0}_{1\times 3} & 1 \end{bmatrix} ,
\end{aligned} \tag{41}$$

where $\frac{\partial e_{Sat_{k,j}}}{\partial \boldsymbol{t}_{\mathcal{B}_k}^{\mathcal{W}\mathsf{T}}}$ is defined similarly to Eq. (12) as

$$\frac{\partial e_{Sat_{k,j}}}{\partial \boldsymbol{t}_{\mathcal{B}_k}^{\mathcal{W}\mathsf{T}}} = \begin{bmatrix} \frac{s_x - t_x}{||\boldsymbol{s}_{k,j} - \boldsymbol{t}_{\mathcal{B}_k}^{\mathcal{W}}||_2} \\ \frac{s_y - t_y}{||\boldsymbol{s}_{k,j} - \boldsymbol{t}_{\mathcal{B}_k}^{\mathcal{W}}||_2} \\ \frac{s_z - t_z}{||\boldsymbol{s}_{k,j} - \boldsymbol{t}_{\mathcal{B}_k}^{\mathcal{W}}||_2} \end{bmatrix}^{\mathsf{T}} \tag{42}$$

with $\boldsymbol{t}_{\mathcal{B}_k}^{\mathcal{W}} = [t_x \ t_y \ t_z]^{\mathsf{T}}$ and $\boldsymbol{s}_{k,j} = [s_x \ s_y \ s_z]^{\mathsf{T}}$.

Then, the derivative of the translation over $\mathfrak{se}3$ manifold is

$$\frac{\partial \boldsymbol{t}_{\mathcal{B}_k}^{\mathcal{W}}}{\partial \boldsymbol{\xi}_{\mathcal{B}_k}^{\mathcal{L}\mathsf{T}}} = \begin{bmatrix} 1 & 0 & 0 & 0 & t_z & -t_y \\ 0 & 1 & 0 & -t_z & 0 & t_x \\ 0 & 0 & 1 & t_y & -t_x & 0 \end{bmatrix} . \tag{43}$$

Substituting (42) and (43) into (41), we have

$$\begin{aligned}
\mathbf{J}(e_{Sat_{k,j}})|_{\mathbf{x}_{\mathrm{V}_k}} =& [d_x, d_y, d_z, \\
& - t_z d_y + t_y d_z, \\
& t_z d_x - t_x d_z, \\
& - t_y d_x + t_x d_y, \\
& d_x, d_y, d_z, 1] \ ,
\end{aligned} \tag{44}$$

where $[d_x \ d_y \ d_z] \triangleq \frac{\partial e_{Sat_{k,j}}}{\partial \boldsymbol{t}_{\mathcal{B}_k}^{\mathcal{W}\mathsf{T}}}$ is given by (42).

References

[1] Z. He, Y. Li, L. Pei, and K. O'Keefe, "Enhanced gaussian process-based localization using a low power wide area network," *IEEE Communications Letters*, vol. 23, no. 1, pp. 164–167, Jan 2019.

[2] Y. Li, Z. He, Z. Gao, Y. Zhuang, C. Shi, and N. El-Sheimy, "Toward robust crowdsourcing-based localization: A fingerprinting accuracy indicator enhanced wireless/magnetic/inertial integration approach," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3585–3600, April 2019.

[3] Y. Li, S. Zahran, Y. Zhuang, Z. Gao, Y. Luo, Z. He, L. Pei, R. Chen, and N. El-Sheimy, "Imu/magnetometer/barometer/mass-flow sensor integrated indoor quadrotor uav localization with robust velocity updates," *Remote Sensing*, vol. 11, no. 7, 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/7/838

[4] W. Jiang, Y. Li, and C. Rizos, "Optimal data fusion algorithm for navigation using triple integration of PPP-GNSS, INS, and terrestrial ranging system," *IEEE Sens. J.*, vol. 15, no. 10, pp. 5634–5644, 2015.

[5] Y. Zhao, "Cubature + extended hybrid Kalman filtering method and its application in PPP/IMU tightly coupled navigation systems," *IEEE Sens. J.*, vol. 15, no. 12, pp. 6973–6985, 2015.

[6] S. Lynen, M. Achtelik, S. Weiss, M. Chli, and R. Siegwart, "A robust and modular multi-sensor fusion approach applied to mav navigation," in *Proc. of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, 2013.

[7] A. B. E. N. Afia, A.-c. Escher, and C. Macabiau, "A low-cost GNSS / IMU / Visual monoSLAM / WSS integration based on kalman filtering for navigation in urban environments," in *28th Int. Tech. Meet. Satell. Div. Inst. Navig. (ION GNSS+ 2013)*, Tampa, Florida, 2013, pp. 618–628.

[8] S.-B. Kim, J.-C. Bazin, H.-K. Lee, K.-H. Choi, and S.-Y. Park, "Ground vehicle navigation in harsh urban conditions by integrating inertial navigation system, global positioning system, odometer and vision data," *IET Radar, Sonar Navig.*, vol. 5, no. 8, p. 814, 2011.

[9] B. M. Aumayer, M. G. Petovello, and G. Lachapelle, "Development of a tightly coupled vision/GNSS system," in *Proc. 27th Int. Tech. Meet. Satell. Div. Inst. Navig. (ION GNSS+ 2014)*, Tampa, FL, 2014, pp. 2202–2211.

[10] P. Roberts, R. Walker, and P. O'Shea, "Tightly coupled GNSS and vision navigation for unmanned air vehicle applications," in *Proc. Aust. Int. Aerosp. Congr.*, 2005.

[11] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *2007 6th IEEE ACM Int. Symp. Mix. Augment. Reality, ISMAR*, 2007.

[12] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, 2015.

[13] S. Thrun, "Simultaneous localization and mapping with sparse extended information filters," *Int. J. Rob. Res.*, vol. 23, no. 7-8, pp. 693–716, 2004.

[14] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. J. Leonard, and F. Dellaert, "iSAM2: Incremental smoothing and mapping using the Bayes tree," *Int. J. Rob. Res.*, vol. 31, no. 2, pp. 216–235, 2012.

[15] B. M. Aumayer, M. G. Petovello, and G. Lachapelle, "Stereo-vision aided GNSS for automotive navigation in challenging environments," in *Proc. 26th Int. Tech. Meet. Satell. Div. Inst. Navig. (ION GNSS+ 2013)*, Nashville, TN, 2013, pp. 511–520.

[16] S. Shen, Y. Mulgaonkar, N. Michael, and V. Kumar, "Multi-sensor fusion for robust autonomous flight in indoor and outdoor environments with a rotorcraft MAV," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2014, pp. 4974–4981.

[17] N. Sünderhauf and P. Protzel, "Towards robust graphical models for GNSS-based localization in urban environments," in *Int. Multi-Conference Syst. Signals Devices, SSD 2012 - Summ. Proc.*, 2012.

[18] R. Watson and J. Gross, "Robust navigation in gnss degraded environment using graph optimization," in *Proc. 26th Int. Tech. Meet. Satell. Div. Inst. Navig. (ION GNSS+ 2017)*, 09 2017.

[19] K. M. Pesyna, "Advanced techniques for centimeter-accurate GNSS positioning on low-cost mobile platforms," Ph.D. dissertation, The University of Texas at Austin, 2015.

[20] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, 2017.

[21] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov 2011, pp. 2564–2571.

[22] M. L. Psiaki and S. Mohiuddin, "Modeling, analysis, and simulation of GPS carrier phase for spacecraft relative navigation," *J. Guid. Control. Dyn.*, vol. 30, no. 6, pp. 1628–1639, 2007.

[23] J. Rehder, J. Nikolic, T. Schneider, T. Hinzmann, and R. Siegwart, "Extending kalibr: Calibrating the extrinsics of multiple IMUs and of individual axes," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2016, pp. 4304–4311.

[24] J. Zhu, "Conversion of Earth-centered Earth-fixed coordinates to geodetic coordinates," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 30, no. 3, pp. 957–961, 1994.

[25] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.

[26] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "G2Oj: A general framework for graph optimization," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 3607–3613.

[27] M. S. Grewal, L. R. Wrill, and A. P. Andrews, *Global Positioning Systems, Inertial Navigation, and Integration*, 2007.

[28] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005.

[29] T. D. Barfoot, *State Estimation for Robotics*, 1st ed. Cambridge University Press, 2017, ch. 7.

[30] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation," *Robot. Sci. Syst. XI*, 2015.

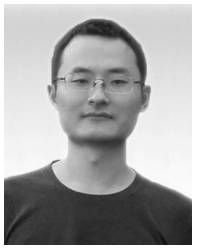
[31] K. M. Frey, T. J. Steiner, and J. P. How, "Complexity analysis and efficient measurement selection primitives for high-rate graph slam," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018. [Online]. Available: http://dx.doi.org/10.1109/icra.2018.8460708

[32] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, 2000.

[33] J. Rehder, J. Nikolic, T. Schneider, T. Hinzmann, and R. Siegwart, "Extending kalibr: Calibrating the extrinsics of multiple imus and of individual axes," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 4304–4311.

Zheng Gong received the B.Eng. degree (with honors) in electronic and computer engineering from the Shanghai Jiao Tong University, Shanghai, China in 2014. Since 2015, he has been with the institute of sensing and navigation, SJTU, where he is currently a Ph.D. student. His main areas of research interest are all source positioning and navigation (ASPN) and UAV.

Rendong Ying received his B.S. in 1994 from East China Norm University, Shanghai, China, Master and Ph.D. degree in 2001 and 2007 respectively from Shanghai Jiao Tong University, all in electronic engineering. He is now associate professor at the Dept. of EE at Shanghai Jiaotong Univ. He is the author of "Embedded System - Principle and Design," Publishing House of Electronics Industry 2011. His research area includes Digital signal processing, SoC architecture and Machine Thinking.



Fei Wen received the B.S. degree from the University of Electronic Science and Technology of China (UESTC) in 2006, and the Ph.D. degree in communications and information engineering from UESTC in 2013. Now he is a research fellow in Shanghai Jiao Tong University. His main research interests are nonconvex optimization, large-scale numerical optimization, sparse and statistical signal processing.

Jiuchao Qian received his B.S. degree from Shan Dong University, China and M.E. degree from Shanghai Maritime University, China. He is currently pursuing his Ph.D. degree in the Department of Information and Communication Engineering at Shanghai Jiao Tong University, Shanghai, China. His current research focuses on indoor localization technologies, signal processing, ubiquitous computing and location-based services.

Peilin Liu received the Doctor of Engineeringdegree from the University of Tokyo, Japan, in 1998, where she worked as a researcher in 1999. From 1999 to 2003, she was a senior researcher at the Central Research Institute of Fujitsu, Tokyo, Japan, where her research topics included multimedia processing, IC design and high-performance processor architecture. She joined Shanghai Jiao Tong University, China in 2003, and is currently a Professor in the Department of Electronic Engineering.