Optimization-Based Visual-Inertial SLAM Tightly Coupled with Raw GNSS Measurements

Jinxu Liu, Wei Gao* and Zhanyi Hu

Abstract— Unlike loose coupling approaches and the EKF-based approaches in the literature, we propose an optimization-based visual-inertial SLAM tightly coupled with raw Global Navigation Satellite System (GNSS) measurements, a first attempt of this kind in the literature to our knowledge. More specifically, reprojection error, IMU pre-integration error and raw GNSS measurement error are jointly minimized within a sliding window, in which the asynchronism between images and raw GNSS measurements is accounted for. In addition, issues such as marginalization, noisy measurements removal, as well as tackling vulnerable situations are also addressed. Experimental results on public dataset in complex urban scenes show that our proposed approach outperforms state-of-the-art visual-inertial SLAM, GNSS single point positioning, as well as a loose coupling approach, including scenes mainly containing low-rise buildings and those containing urban canyons.

I. INTRODUCTION

Positioning in outdoor scenes has long been a concerned task in the realms such as autonomous driving. Global Navigation Satellite System (GNSS) has been widely exploited for decades, due to its ability to acquire geodetic coordinates and its stable performance in open area. However, in *urban canyons* where there are lots of skyscrapers in the surroundings, GNSS positioning results dramatically deteriorate. On the other hand, V-SLAM or VI-SLAM approaches do not suffer from *urban canyons*, and their local accuracy is much higher than that of single point GNSS positioning.

However, V-SLAM and VI-SLAM also have their inherent shortcomings. Besides the deterioration of accuracy in low-textured scenes and the scenes with many moving objects, they typically suffer from drift. And more importantly, since global position is unobservable for SLAM approaches, what can be obtained from SLAM approaches are the poses in a local world frame instead of a geodetic coordinate frame. Hence GNSS positioning and V-SLAM or VI-SLAM approaches are complementary, due to their different global and local accuracy, as well as their different vulnerable scenes. Hence fusing vision, IMU and GNSS is promising for accurate global positioning in complex urban scenes.

II. RELATED WORK

A. Loose Coupling Approaches

Loose coupling approaches firstly derive poses from feature points, or the positions and velocities from raw GNSS measurements, or both of the above two quantities. Then the derived quantities rather than raw measurements are involved in the fusion stage. Some approaches [1], [2] derive poses from feature points, and fuse them with raw GNSS measurements such as pseudorange and Doppler shift. Conversely, some approaches [3], [4], [26] derive positions from raw GNSS measurements, and fuse them with feature points from the images. Other approaches [5], [6], [7], [8] apply loose coupling on both the vision side and the GNSS side. Some approaches [2], [9] use visual information to aid GNSS positioning by judging which satellites are visible. The above loose coupling approaches are not optimal in the sense that the available information is not fully exploited.

B. Tight Coupling Approaches

This category of approaches employs tight coupling on both the vision side and the GNSS side, i.e., both the feature points from the image and the raw GNSS measurements serve as measurements in either EKF-based or optimization-based framework. The EKF-based approaches in [10], [11] tightly couple vision, IMU and single point GNSS. They perform state prediction using the IMU measurements, and update the states using the feature points as well as the pseudorange and Doppler shift measurements. Similar EKF-based framework is adopted in [12], while [12] utilizes double-differenced GNSS instead of single point GNSS. Although double-differenced GNSS provides measurements for more accurate positioning, it requires an additional base station in comparison with single point GNSS.

Besides EKF-based approaches, optimization-based approaches are emerging as well. Compared with EKF, batch optimization allows for the reduction of error through re-linearization in visual-inertial navigation systems [13] and proves to have better performance in GNSS-IMU fusion task [14]. Therefore, optimization is a promising solution to fuse vision, IMU and GNSS measurements. The optimization-based approach [15] tightly couples feature points with pseudorange from single point GNSS in bundle adjustment. However, Doppler shift is not utilized, and IMU measurements are not tightly coupled in [15], and only the magnetometer in IMU is exploited, to determine the direction of the local world frame. Nevertheless, tightly coupling IMU measurements can eliminate the effect of the asynchronism between vision and GNSS measurements, thus the low-speed motion assumption applied in [15] is not needed in this case.

In this paper we propose an optimization-based approach that tightly couples vision, IMU and single point raw GNSS

measurements including pseudoranges and Doppler shift, a first attempt in literature to our knowledge.[1]

III. FRAMES AND NOTATIONS

In this paper, the world frames include the earth-centered, earth-fixed (ECEF) frame, the ground east-north-up (ENU) frame, and the local world frame for SLAM. The sensor frames include the camera frame, the IMU frame and the GNSS receiver frame. The above world frames and sensor frames are illustrated in Fig.1. We use $(\cdot)^{WE}$, $(\cdot)^{WG}$ and $(\cdot)^{WL}$ to denote the ECEF frame, the ground ENU frame and the local world frame for SLAM respectively, and use $(\cdot)^c$, $(\cdot)^b$ and $(\cdot)^g$ to denote the camera frame, IMU frame and GNSS receiver frame respectively. Specifically, we use $(\cdot)^{b_k}$ and $(\cdot)^{g_k}$ to denote the IMU frame and GNSS receiver frame corresponding to the kth image, and slightly abuse the symbols using $(\cdot)^{b_t}$ and $(\cdot)^{g_t}$ to denote the IMU frame and GNSS receiver frame at some moment t.
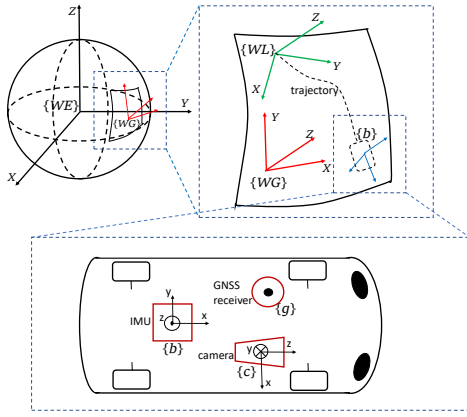

Fig. 1. Illustration of world frames and sensor frames. The globe in the upper left of the figure is the earth globe. The origin of the ECEF frame $\{WE\}$ is on the center of the earth. The X axis of the ECEF frame points to the point of intersection between the prime meridian and equator, and the Z axis points to the North Pole. The ground ENU frame $\{WG\}$ and the local world frame for SLAM $\{WL\}$ both locate on the ground, with their Z axes pointing upward.

Let $\mathbf{R}_A^B$ denote the rotation matrix that takes a vector in frame $\{A\}$ to frame $\{B\}$, and $\mathbf{q}_A^B$ is its quaternion form. $\mathbf{p}_A^B$ is the coordinate of the origin point of frame $\{A\}$ in frame $\{B\}$, and $\mathbf{v}_A^B$ is the velocity of the origin point of frame $\{A\}$ measured in frame $\{B\}$. Note that for GNSS receiver frame, only the coordinate of its origin point in frame $\{WL\}$ matters in our approach, and the orientations of its axes can be regarded as arbitrary. Let $\mathbf{b}_{a_k}$ and $\mathbf{b}_{\omega_k}$ denote the accelerometer bias and gyroscope bias corresponding to image k respectively. Moreover, we use $[\cdot]_\times$ to denote the skew symmetric matrix corresponding to a vector.

IV. OUR APPROACH

A. GNSS-SLAM Initialization

After our system starts, the initialization for VI-SLAM is performed, and the nonlinear optimization for VI-SLAM

starts performing temporarily following the routine of [16]. Meanwhile, the GNSS single point positioning is performed. Hence we have obtained the VI-SLAM trajectory and the GNSS single point positioning trajectory respectively. At GNSS-SLAM initialization stage we align the above two trajectories, in order to compute the value of $\mathbf{R}_{WG}^{WE}$ and $\mathbf{p}_{WG}^{WE}$, as well as the initial value of $\mathbf{R}_{WL}^{WG}$ and $\mathbf{p}_{WL}^{WG}$.

Firstly, we select one position calculated by GNSS single point positioning as the reference point $\mathbf{p}_{ref}^{WE}$, which serves as the origin point of the ground ENU frame. In our implementation, the third GNSS single point positioning result since the system starts is selected as the reference point. Given the reference point, $\mathbf{R}_{WG}^{WE}$ and $\mathbf{p}_{WG}^{WE}$ can be computed straightforwardly [17]. Then the GNSS single point positioning trajectory is converted from the ECEF frame to the ground ENU frame as

$$\mathbf{p}_g^{WG} = \mathbf{R}_{WG}^{WE\,T}(\mathbf{p}_g^{WE} - \mathbf{p}_{WG}^{WE}). \tag{1}$$

Hence we have two sets of positions $\{\mathbf{p}_{g_l}^{WG}|l = 0 \dots L-1\}$ and $\{\mathbf{p}_{g_l}^{WL}|l = 0 \dots L-1\}$, of which the former includes the positions from GNSS single point positioning in frame $\{WG\}$, and the latter includes the positions estimated from VI-SLAM in frame $\{WL\}$. Note that the positions from VI-SLAM are in fact $\mathbf{p}_{b_l}^{WL}$, but at the initialization stage we neglect the translation from IMU frame to GNSS receiver frame, thus we regard $\mathbf{p}_{b_l}^{WL}$ as $\mathbf{p}_{g_l}^{WL}$. At this stage interpolation is performed to align the timestamps of the above two sets of positions. At last a 5-Degrees of Freedom(DoF) alignment between the above two sets is performed to compute the initial value of $\mathbf{R}_{WL}^{WG}$ and $\mathbf{p}_{WL}^{WG}$, by minimizing the following cost:

$$\min_{s,\mathbf{R}_{WL}^{WG},\mathbf{p}_{WL}^{WG}} \sum_{l=0}^{L-1} \| \mathbf{p}_{g_l}^{WG} - s\mathbf{R}_{WL}^{WG}\mathbf{p}_{g_l}^{WL} - \mathbf{p}_{WL}^{WG} \|^2. \tag{2}$$

The scale parameter s is estimated because the VI-SLAM is sensitive to scale drift in degenerate cases [18]. Note that $\mathbf{R}_{WL}^{WG}$ only contains one DoF, because the Z axes of both frame $\{WG\}$ and frame $\{WL\}$ point upward.

B. Tightly Coupled Optimization

After GNSS-SLAM initialization, raw GNSS measurements, i.e. pseudoranges and Doppler shift, are integrated into the optimization. The states to be estimated include

$$\mathscr{X} = [\mathbf{x}_0 \dots \mathbf{x}_{K-1}, f_0 \dots f_{M-1}, \mathbf{T}_c^b, \mathbf{T}_{WL}^{WG}, \boldsymbol{\delta t}_0^r \dots \boldsymbol{\delta t}_{E-1}^r, \dot{\boldsymbol{\delta t}}_0^r \dots \dot{\boldsymbol{\delta t}}_{E-1}^r], \tag{3}$$

where K is the number of images in the sliding window, M is the number of landmarks, and E is the number of GNSS epochs in the sliding window. f is the inverse depth of one landmark in camera frame, and

$$\begin{aligned}
\mathbf{x}_k &= \left[\mathbf{p}_{b_k}^{WL}, \mathbf{v}_{b_k}^{WL}, \mathbf{q}_{b_k}^{WL}, \mathbf{b}_{a_k}, \mathbf{b}_{\omega_k}\right], k = 0 \dots K-1, \\
\mathbf{T}_c^b &= \left[\mathbf{R}_c^b, \mathbf{p}_c^b\right], \mathbf{T}_{WL}^{WG} = \left[\mathbf{R}_{WL}^{WG}, \mathbf{p}_{WL}^{WG}\right], \\
\boldsymbol{\delta t}_e^r &= \left[\{\delta t_{j,e}^r | \forall j \in \mathscr{O}(e)\}\right], e = 0 \dots E-1, \\
\dot{\boldsymbol{\delta t}}_e^r &= \left[\{\dot{\delta t}_{j,e}^r | \forall j \in \mathscr{O}(e)\}\right], e = 0 \dots E-1,
\end{aligned} \tag{4}$$

where $\mathscr{O}(e)$ is the set of observable GNSS constellations such as GPS, GLONASS, etc. at epoch e, $\delta t_{j,e}^r$ and $\dot{\delta t}_{j,e}^r$

are the receiver clock bias and clock drift w.r.t GNSS constellation j at epoch e respectively, and $\boldsymbol{\delta t}_e^r$ and $\dot{\boldsymbol{\delta t}}_e^r$ are the vectors containing the receiver clock biases and clock drifts for all the observable constellations at epoch e respectively. Compared with VINS-mono [16], the introduced additional parameters include the transformation from the local world frame to ground ENU frame, as well as clock biases and clock drifts. The transformation from ground ENU frame to ECEF frame ($\mathbf{R}_{WG}^{WE}, \mathbf{p}_{WG}^{WE}$) is treated as a known quantity rather than a parameter, because the ground ENU frame can be located anywhere on the ground that is not too far away from the origin of the local world frame for SLAM, for example 1 km. Although GNSS measurements and images are collected at different moments, we do not introduce additional parameters concerning pose and velocity at GNSS measurement moments. In other words, we incorporate intermediate GNSS measurements, which is inspired by [19]. Note that just like in the GNSS-SLAM initialization stage, $\mathbf{R}_{WL}^{WG}$ only has 1 DoF, so it is parameterized only by the rotation angle around the Z axis, which writes as κ_{WL}^{WG}.

The cost function $c(\mathscr{X})$ to be minimized is

$$
\begin{aligned}
c(\mathscr{X}) = & c^{reproj}(\mathscr{X}) + c^{IMU}(\mathscr{X}) + \sum_{e=0}^{E-1} \sum_{i \in \mathscr{S}(e)} \mathbf{e}_{i,e}^{P\,T} \mathbf{W}^P \mathbf{e}_{i,e}^P \\
& + \sum_{e=0}^{E-1} \sum_{i \in \mathscr{S}(e)} \mathbf{e}_{i,e}^{D\,T} \mathbf{W}^D \mathbf{e}_{i,e}^D + \mathbf{e}^{marg\,T} \mathbf{e}^{marg},
\end{aligned}
\tag{5}
$$

where E is the number of GNSS epochs in the sliding window and $\mathscr{S}(e)$ is the set of observable GNSS satellites at epoch e. $\mathbf{e}_{i,e}^P$ and $\mathbf{e}_{i,e}^D$ are the pseudorange residual and the Doppler shift residual for satellite i observed at epoch e respectively. $\mathbf{e}^{marg}$ denotes the marginalization residual. The reprojection factors $c^{reproj}(\mathscr{X})$ and IMU factors $c^{IMU}(\mathscr{X})$ are identical to those in VINS-mono [16], and the readers may refer to (14), (16) and (17) in VINS-mono [16] for details. The IMU pre-integration performed in our proposed approach is also identical to that in VINS-mono [16].

As for the GNSS factors, the pseudorange residual $\mathbf{e}_{i,e}^P$ and the Doppler shift residual $\mathbf{e}_{i,e}^D$ write as

$$
\begin{aligned}
\mathbf{e}_{i,e}^P = & \| \mathbf{p}_{s_i,e}^{WE} - (\mathbf{R}_{WG}^{WE} \mathbf{R}_{WL}^{WG} \mathbf{p}_{g_{m(e)}}^{WL} + \mathbf{R}_{WG}^{WE} \mathbf{p}_{WL}^{WG} + \mathbf{p}_{WG}^{WE}) \| + c\delta t_{j(i),e}^r \\
& - c\delta t_{i,e}^s - \rho_{i,e}, \\
\mathbf{e}_{i,e}^D = & \mathbf{l}_{i,e}^T(\mathbf{v}_{s_i,e}^{WE} - \mathbf{R}_{WG}^{WE} \mathbf{R}_{WL}^{WG} \mathbf{v}_{g_{m(e)}}^{WL}) + c\dot{\delta t}_{j(i),e}^r - c\dot{\delta t}_{i,e}^s - (-\lambda_{i,e} D_{i,e}),
\end{aligned}
\tag{6}
$$

where

$$
\begin{aligned}
\mathbf{l}_{i,e} = & \frac{\mathbf{p}_{s_i,e}^{WE} - (\mathbf{R}_{WG}^{WE} \mathbf{R}_{WL}^{WG} \mathbf{p}_{g_{m(e)}}^{WL} + \mathbf{R}_{WG}^{WE} \mathbf{p}_{WL}^{WG} + \mathbf{p}_{WG}^{WE})}{\| \mathbf{p}_{s_i,e}^{WE} - (\mathbf{R}_{WG}^{WE} \mathbf{R}_{WL}^{WG} \mathbf{p}_{g_{m(e)}}^{WL} + \mathbf{R}_{WG}^{WE} \mathbf{p}_{WL}^{WG} + \mathbf{p}_{WG}^{WE}) \|}, \\
\mathbf{p}_{g_{m(e)}}^{WL} = & \mathbf{p}_{b_k}^{WL} + \mathbf{v}_{b_k}^{WL} \Delta t_{k,m(e)} + \mathbf{R}_{b_k}^{WL}(\boldsymbol{\alpha}_{b_{m(e)}}^{b_k} + \mathbf{R}(\boldsymbol{\gamma}_{b_{m(e)}}^{b_k}) \mathbf{p}_g^b) \\
& - \frac{1}{2} \mathbf{g}^{WL} \Delta t_{k,m(e)}^2, \\
\mathbf{v}_{g_{m(e)}}^{WL} = & \mathbf{v}_{b_k}^{WL} + \mathbf{R}_{b_k}^{WL}(\boldsymbol{\beta}_{b_{m(e)}}^{b_k} - \mathbf{R}(\boldsymbol{\gamma}_{b_{m(e)}}^{b_k})[\mathbf{p}_g^b]_\times(\hat{\boldsymbol{\omega}}_{m(e)} - \mathbf{b}_{\omega_k})) \\
& - \mathbf{g}^{WL} \Delta t_{k,m(e)}.
\end{aligned}
\tag{7}
$$

In (6) and (7), c denotes the speed of light, $\mathbf{p}_{s_i,e}^{WE}$ and $\delta t_{i,e}^s$ are the position and clock bias of satellite i at GNSS

epoch e respectively, and $\mathbf{v}_{s_i,e}^{WE}$ and $\dot{\delta t}_{i,e}^s$ are the velocity and clock drift of satellite i at epoch e respectively. The above four quantities are computed directly from the broadcast emphemeris. The subscript $j(i)$ means the constellation corresponding to satellite i, and the subscript $m(e)$ denotes the measurement moment of GNSS epoch e. The parameters $\delta t_{j(i),e}^r$ and $\dot{\delta t}_{j(i),e}^r$ are the receiver clock bias and clock drift w.r.t GNSS constellation $j(i)$ at epoch e respectively. $\rho_{i,e}$, $\lambda_{i,e}$ and $D_{i,e}$ are the pseudorange measurement, wavelength of the carrier, and Doppler shift measurement of satellite i at epoch e respectively. $\Delta t_{k,m(e)}$ is the time interval from the exposure moment of image k to the moment $m(e)$. $\mathbf{p}_g^b$ is the translational component of IMU-GNSS receiver extrinsic parameters, which is assumed to have been accurately calibrated beforehand in this paper. $\boldsymbol{\alpha}_{b_{m(e)}}^{b_k}$, $\boldsymbol{\beta}_{b_{m(e)}}^{b_k}$ and $\boldsymbol{\gamma}_{b_{m(e)}}^{b_k}$ are the nominal states coming from IMU pre-integration from the exposure moment of image k to the moment $m(e)$, and they correspond to displacement, change in velocity, and relative rotation respectively. $\mathbf{R}(\boldsymbol{\gamma}_{b_{m(e)}}^{b_k})$ is the rotation matrix converted from the rotation quaternion $\boldsymbol{\gamma}_{b_{m(e)}}^{b_k}$. For the details of IMU-preintegration the readers may refer to (3) and (5) in VINS-mono [16]. $\hat{\boldsymbol{\omega}}_{m(e)}$ is the angular velocity measurement from IMU at moment $m(e)$, and $\mathbf{g}^{WL}$ is the gravity vector in frame $\{WL\}$, which writes as $\begin{bmatrix} 0 & 0 & g \end{bmatrix}^T$ with g being the magnitude of gravity. In (7), an approximation is applied that the gyroscope bias at image exposure moment $\mathbf{b}_{w_k}$ substitutes for the gyroscope bias at GNSS measurement moment $m(e)$, because gyroscope bias is a slow time-varying quantity.

In fact, $\mathbf{W}^P$ and $\mathbf{W}^D$ in (5) are both 1×1 matrices which can be regarded as scalars. In our implementation, we adopt $\mathbf{W}^P = 1$ and $\mathbf{W}^D = 4$, assuming the median errors of pseudorange measurements and pseudorange rates converted from Doppler shift measurements to be $1m$ and $0.5m/s$ respectively. The optimization is performed in a sliding window by minimizing the cost function $c(\mathscr{X})$ using Ceres Solver [20].

C. Sliding Window and Marginalization

In the sliding window in our approach illustrated in Fig.2, the parameters related to GNSS, i.e. GNSS receiver clock bias and clock drift, are *attached* to the image frame just before it. There may exist more than one GNSS parameter node attached to a certain image frame. If the second latest frame is a keyframe, we keep it in the sliding window, and marginalize the oldest keyframe and the GNSS parameters and feature points attached to it, as well as the reprojection factors, IMU factors and GNSS factors related to them. If the second latest frame is not a keyframe, we simply remove the frame and all its corresponding visual measurements. However, the GNSS parameters attached to the second latest frame will not be removed, and they will be attached to the third latest frame instead, i.e. removed and added in Fig.2. The GNSS factors that are originally attached to the second latest frame will be changed, for the reason that the IMU pre-integration in the above GNSS factors need to be recomputed because their starting moments will switch from
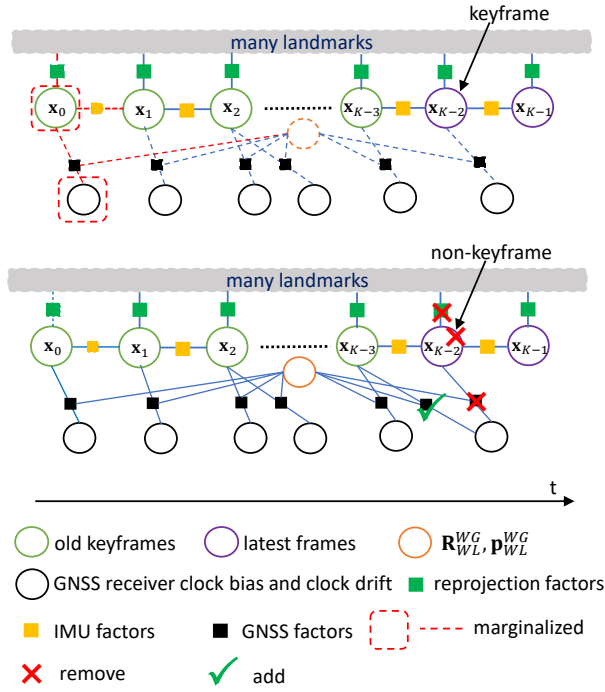
Fig. 2. Marginalization in sliding window. The hollow circles represent the parameters (nodes), and the solid rectangles represent the factors (edges). The camera-IMU extrinsic parameters and the marginalization factor are not shown here for clarity of explanation.

the second latest frame to the third latest frame. The GNSS measurements in the above GNSS factors remain unchanged.

The marginalization in our approach is carried out using Schur-Complement. The readers may refer to [21] and [16] for details of applying marginalization. Note that since marginalization is applied, the GNSS measurements that have slid out of the sliding window still take effect on the state estimation in the current sliding window. Therefore $\mathbf{R}_{WL}^{WG}$ and $\mathbf{p}_{WL}^{WG}$ are still observable even if there are no GNSS measurements in the current sliding window.

D. Removing Noisy GNSS Measurements

In urban canyons, raw GNSS measurement can be pretty noisy. Therefore, before the GNSS measurements enter into the sliding window for optimization, a filtering is carried out by one of the following two possible methods. The first method is to perform single point positioning, as well as velocity determination using raw Doppler shift measurements [17]. For the second method, if all GNSS measurements are excluded during the past 5 seconds, single point positioning and velocity determination using raw Doppler shift measurements are performed. Otherwise a joint optimization is performed inside a sliding window similar to Sect. IV-B, but the optimization here differs with Sect. IV-B in that only the GNSS measurements at the current epoch are involved in the optimization. For both methods, we compute the residuals of pseudoranges and Doppler shift measurements after the optimization. The pseudoranges and Doppler shifts are filtered out if their corresponding residuals exceed the threshold

T^P and T^D respectively. If too few GNSS measurements are left after filtering out noisy measurements at a certain GNSS epoch, we will remove all the GNSS measurements at the epoch altogether. The first method has the advantage in speed and the second one has the advantage in accuracy. The reason we perform a GNSS single point positioning if all GNSS measurements are excluded during the past 5 seconds for the second method is that, in this way we prevent good GNSS measurements from being excluded because of the incorrectly estimated current position. A discussion on the above two methods is presented in Sect. V-B. In our implementation, $T^P = 10m$ and $T^D = 3m/s$. Fig.3 shows the GNSS single point positioning results and the results of our filtering operation by the first method. From Fig.3 we can see that the GNSS epochs when measurements are all filtered out generally experience less accurate single point positioning results than the epochs when measurements are partially filtered out or not filtered out.
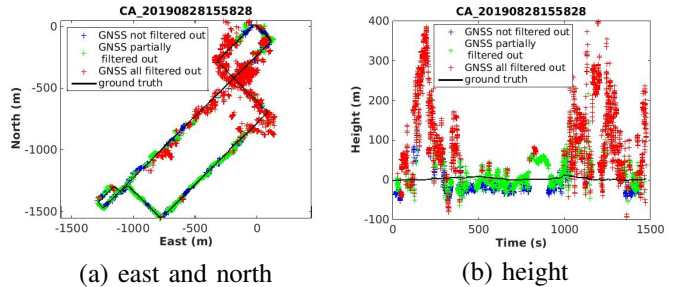

(a) east and north

(b) height

Fig. 3. GNSS single point positioning results w.r.t east and north directions (a), and w.r.t height direction (b). The red, green and blue markers are epochs when GNSS measurements are all filtered out, partially filtered out and not filtered out respectively.

E. Some Implementation Details

- At the initial stage just after GNSS-SLAM initialization, a prior factor constraining the parameters $(\mathbf{R}_{WL}^{WG}, \mathbf{p}_{WL}^{WG})$ is added to the sliding window, because at the initial stage the marginalization factor does not contain enough information from GNSS measurements in the past, and thus the estimation of $(\mathbf{R}_{WL}^{WG}, \mathbf{p}_{WL}^{WG})$ is sensitive to noises in GNSS measurements. After the GNSS factors in more than T^N GNSS epochs have been marginalized, the prior factor is permanently removed from the sliding window. In our implementation, $T^N = 30$.

- According to [18] and [22], VI-SLAM system is not well-constrained before the platform has made a turn for the first time. Therefore, we employ the strategy proposed in [22] but in a much simpler way that we start to adjust the camera-IMU extrinsic parameters $(\mathbf{R}_c^b, \mathbf{p}_c^b)$ in optimization after the platform has made a turn. Here a *turn* means a motion with large rotation, e.g. the vehicle turning left or turning right at a crossroad.

- It is a vulnerable scene for our approach when the car stops at a crossroad. In our implementation when the estimated speed of the second latest frame is under $0.5m/s$, the GNSS factors in the sliding window

TABLE I

COMPARISON OF MEAN ABSOLUTE ERROR (MAE), ROOT MEAN SQUARE ERROR (RMSE) AND TRAJECTORY COMPLETENESS (TC) OF RESULTING TRAJECTORIES FROM DIFFERENT APPROACHES

Data Sequence	Length	Approach	MAE in Translation (m)			MAE in Rotation (deg)			RMSE in Trans. (m)	TC
			X	Y	Z	Z	Y	X		
CA_20190828155828	5.9km	VINS-mono [16]	93.75	77.38	11.16	2.106	1.036	2.501	142.74	99.97%
		GNSS SPP	**6.294**	**5.269**	11.96	*	*	*	28.12	46.85%
		VINS-Fusion (with GNSS) [8]	11.59	9.256	33.94	5.225	10.72	11.82	55.15	98.73%
		Proposed	7.063	7.637	**4.047**	**1.499**	**0.323**	**2.143**	**14.33**	99.97%
CA_20190828173350	3.2km	VINS-mono [16]	49.17	30.79	**2.686**	2.251	**0.644**	2.206	64.84	99.76%
		GNSS SPP	88.00	48.29	181.26	*	*	*	861.72	50.73%
		VINS-Fusion (with GNSS) [8]	15.69	15.80	30.17	21.39	26.59	16.87	51.59	99.77%
		Proposed	**2.897**	**11.48**	3.523	**1.006**	0.668	**2.098**	**14.34**	99.76%
CA_20190828184706	1.8km	VINS-mono [16]	6.669	6.080	1.347	**2.608**	0.455	2.059	11.29	98.39%
		GNSS SPP	2.270	2.190	5.514	*	*	*	8.394	93.03%
		VINS-Fusion (with GNSS) [8]	2.120	2.729	4.519	6.130	15.17	9.207	7.518	100.00%
		Proposed	**2.099**	**1.692**	**0.623**	4.671	**0.367**	**1.863**	**3.590**	98.39%
CA_20190828190411	1.0km	VINS-mono [16]	7.128	7.816	**1.156**	2.191	1.266	2.499	11.88	99.49%
		GNSS SPP	3.100	3.117	11.03	*	*	*	15.72	100.00%
		VINS-Fusion (with GNSS) [8]	6.698	2.833	10.05	10.61	10.08	8.418	18.96	99.37%
		Proposed	**1.604**	**1.476**	2.735	**2.135**	**0.814**	**2.151**	**4.455**	99.56%

Here *GNSS SPP* means GNSS single point positioning. *Proposed* means the proposed approach in this paper. The symbol * means the result is not available. GNSS single point positioning only provides positioning results, and hence its rotational accuracy is not available.

are temporarily inactivated, but they still participate in marginalization as if they were in the sliding window.

V. EXPERIMENTS

The experiments are conducted on the public dataset UrbanLoco [23], which is collected in highly urbanized areas in San Francisco and Hong Kong. The images in Hong Kong data are collected using a fisheye sky camera, and the images in San Francisco data are collected using six 360-degree view cameras. According to [23], the ground truth in UrbanLoco dataset is provided by Novatel SPAN-CPT, which is an RTK-IMU navigation system. According to [23], when the satellite observability is satisfactory, the ground truth error is within 2 cm, and the calibration certificate dictates that the error after 10 seconds of GNSS backout is within 12 cm. However, there does exist something wrong with the ground truth in the sequence CA_20190828190411, and hence we have manually excluded the incorrect part of ground truth in the sequence CA_20190828190411 according to the Google Earth. Our approach is evaluated on 4 sequences collected in San Francisco, because the other 3 sequences in San Francisco contain tunnels, which are very tough scenerios for both camera and GNSS signals, and VINS-mono [16] evaluation results on only the above 4 sequences are reported in the paper [23]. The 4 sequences contain both urban canyons and scenes with low-rise buildings. All the experiments presented are performed on a PC with Intel Core i7 3.6GHz×6 core CPU and 64GB memory.

A. Evaluation on Accuracy of Trajectories

Our proposed approach is compared with state-of-the-art VI-SLAM approach VINS-mono [16], GNSS single point positioning results computed using RTKLIB [24], and the loose coupling approach VINS-Fusion (with GNSS) [8]. The accuracies of the above approaches are compared in

(a) CA_20190828155828

(b) CA_20190828173350

(c) CA_20190828184706
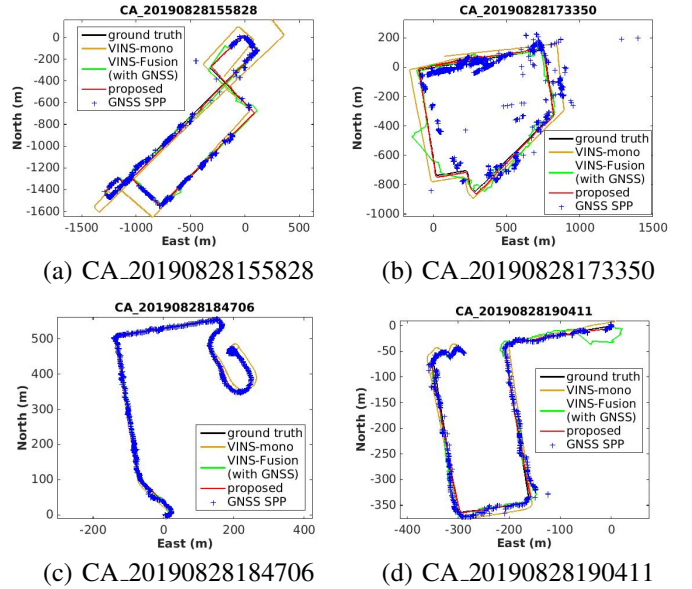

(d) CA_20190828190411

Fig. 4. Comparison of trajectories in east and north directions. In figure (b) a few very noisy GNSS SPP results that lie outside the figure are omitted. All trajectories are aligned with the ground truth.

Table I. In Table I, all the four approaches are evaluated by ourselves. Although the evaluation results of VINS-mono [16] is presented in [23], it is reevaluated by us because the incorrect part of ground truth is excluded in the sequence CA_20190828190411. *GNSS SPP* is the single point positioning result produced using RTKLIB [24]. For VINS-Fusion (with GNSS) [8], the inputted GNSS positions are the single point positioning results from RTKLIB [24], i.e. the results of *GNSS SPP*. In this subsection, for our proposed approach, the second method to remove the noisy GNSS

(a) CA_20190828155828   (b) CA_20190828173350

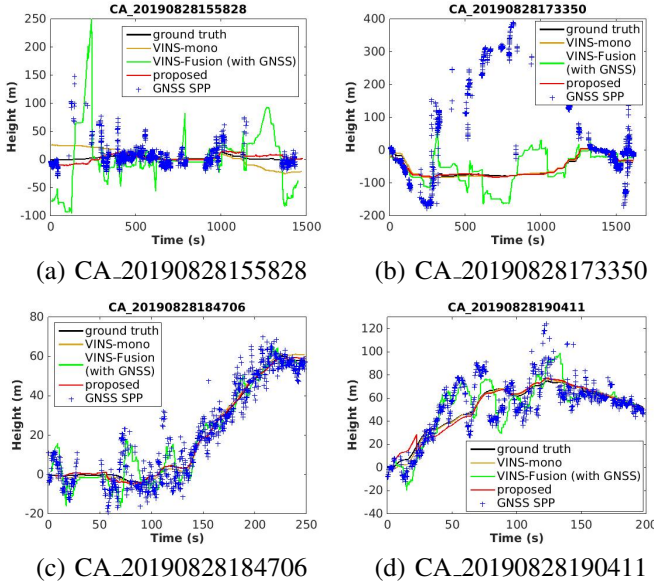(c) CA_20190828184706   (d) CA_20190828190411

Fig. 5. Comparison of trajectories in height direction. In figure (a) and (b) a few very noisy GNSS SPP results that lie outside the figure are omitted. All trajectories are aligned with the ground truth.

measurements described in Sect. IV-D is applied. For fairness of comparison, we employ the frontal camera and the IMU in VINS-mono [16], VINS-Fusion (with GNSS) [8] and our proposed approach. Also for fairness of comparison, because the resulting trajectory of VINS-mono [16] is aligned to the ground truth by a rigid body transformation [25] before evaluation, the other three approaches are also aligned to the ground truth by a rigid body transformation before evaluation. For each approach the loop closure option is turned off. For *GNSS SPP* and our proposed approach, the same raw GNSS measurements are utilized, including the measurements from both GPS and GLONASS. Same as the evaluation of VINS-mono [16] in paper [23], our proposed approach is also evaluated at about half-real time playback rate, and VINS-mono [16] is reevaluated at half-real time playback rate. The comparisons of trajectories among different approaches w.r.t the east and north directions, as well as w.r.t. the height direction are also shown in Fig.4 and Fig.5 respectively.

From Table I, Fig.4 and Fig.5 we can see that our proposed approach outperforms the other three approaches in terms of both translation and rotation accuracy. Table I reports the mean absolute error (MAE) in translation and in rotation, the root mean square error (RMSE) in translation, as well as the trajectory completeness. RMSE in translation is precisely the absolute trajectory error (ATE) presented in [25]. Trajectory completeness is computed as follows: We sample the time interval every 0.1 second from the beginning to the end of each sequence. For every sampled moment, if and only if position is successfully calculated at any moment that is within 3 seconds before or after it, the moment is regarded as successfully positioned. Trajectory completeness is the ratio of the number of successfully positioned sampled moments

to the number of all the sampled moments. From Table I the most evident conclusion that can be drawn is that the translation RMSE of our proposed approach is dramatically smaller than those of other approaches.

Through tightly coupling raw GNSS measurements, our approach achieves global positioning results, whose resulting trajectories on the two sequences are projected onto Google Earth as shown in Fig.6.
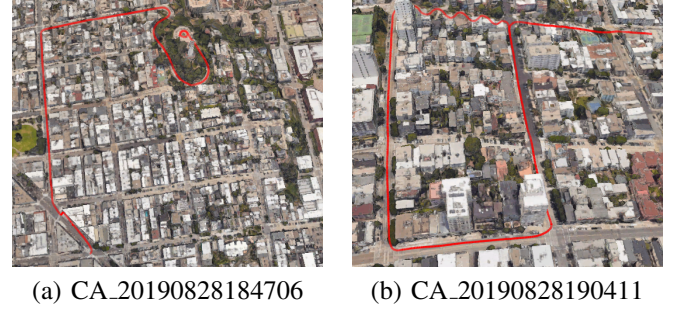

(a) CA_20190828184706   (b) CA_20190828190411

Fig. 6. Our resulting trajectories of (a) CA_20190828184706 and (b) CA_20190828190411 projected onto Google Earth.

B. Comparison on Two Noisy GNSS Measurements Removal Methods

A comparison on the two noisy GNSS measurements removal methods described in Sect. IV-D is supported here in terms of absolute trajectory error (ATE) and time consumption. From Table II we can see that the second one has a better accuracy, at the cost of more time consumption.

TABLE II

COMPARISON OF ATE AND MEAN TIME CONSUMPTION (MTC)

Sequence	Length	Method	ATE (m)	MTC (ms)
CA_20190828-155828	5.9km	GNSS solver	19.081	**0.389**
		Mixed solver	**14.328**	24.663
CA_20190828-173350	3.2km	GNSS solver	20.280	**0.341**
		Mixed solver	**14.337**	26.080
CA_20190828-184706	1.8km	GNSS solver	**3.395**	**0.336**
		Mixed solver	3.590	30.141
CA_20190828-190411	1.0km	GNSS solver	5.941	**0.347**
		Mixed solver	**4.455**	29.473

Here *mean time consumption (MTC)* means the mean time consumption of one execution of excluding the noisy GNSS measurements. *GNSS solver* and *Mixed solver* represent the first method and the second method in Sect. IV-D respectively.

VI. CONCLUSIONS

In this paper, we propose an optimziation-based visual-inertial SLAM tightly coupled with raw GNSS measurements. Feature points, IMU measurements as well as pseudo-ranges and Doppler shift measurements from GNSS, which are captured at different moments, are integrated by means of optimization in a sliding window. Experimental results prove that our proposed approach outperforms state-of-the-art visual-inertial SLAM, GNSS single point positioning, as well as the loose coupling approach [8] on public dataset.

REFERENCES

[1] M. Schreiber, H. Königshof, A. Hellmund, and C. Stiller. Vehicle localization with tightly coupled gnss and visual odometry. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, pages 858–863, 2016.

[2] P. V. Gakne and K. O'Keefe. Tightly-coupled gnss/vision using a sky-pointing camera for vehicle navigation in urban areas. *Sensors*, 18(4):1244, 2018.

[3] Y. Yu, W. Gao, C. Liu, S. Shen, and M. Liu. A gps-aided omnidirectional visual-inertial state estimator in ubiquitous environments. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7750–7755, 2019.

[4] W. Lee, K. Eckenhoff, P. Geneva, and G. Huang. Intermittent gps-aided vio: Online initialization and calibration. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5724–5731, 2020.

[5] J. Rehder, K. Gupta, S. Nuske, and S. Singh. Global pose estimation with limited gps and long range visual odometry. In *2012 IEEE International Conference on Robotics and Automation*, pages 627–633, 2012.

[6] R. Mascaro, L. Teixeira, T. Hinzmann, R. Siegwart, and M. Chli. Gomsf: Graph-optimization based multi-sensor fusion for robust uav pose estimation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1421–1428, 2018.

[7] W. Yan, L. Bastos, J. A. Gonçalves, A. Magalhães, and T. Xu. Image-aided platform orientation determination with a gnss/low-cost imu system using robust-adaptive kalman filter. *GPS Solutions*, 22(1):12, 2018.

[8] T. Qin, S. Cao, J. Pan, and S. Shen. A general optimization-based framework for global pose estimation with multiple sensors, 2019.

[9] W. Wen, X. Bai, Y. C. Kan, and L. Hsu. Tightly coupled gnss/ins integration via factor graph and aided by fish-eye camera. *IEEE Transactions on Vehicular Technology*, 68(11):10651–10662, 2019.

[10] D. H. Won, E. Lee, M. Heo, S. Sung, J. Lee, and Y. J. Lee. Gnss integration with vision-based navigation for low gnss visibility conditions. *GPS solutions*, 18(2):177–187, 2014.

[11] D. H. Won, E. Lee, M. Heo, S. Lee, J. Lee, J. Kim, S. Sung, and Y. J. Lee. Selective integration of gnss, vision sensor, and ins using weighted dop under gnss-challenged environments. *IEEE Transactions on Instrumentation and Measurement*, 63(9):2288–2298, 2014.

[12] T. Li, H. Zhang, Z. Gao, X. Niu, and N. El-Sheimy. Tight fusion of a monocular camera, mems-imu, and single-frequency multi-gnss rtk for precise navigation in gnss-challenged environments. *Remote Sensing*, 11(6):610, 2019.

[13] G. Huang. Visual-inertial navigation: A concise review. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9572–9582, 2019.

[14] W. Wen, T. Pfeifer, X. Bai, and L. Hsu. It is time for factor graph optimization for gnss/ins integration: Comparison between FGO and EKF. *CoRR*, 2020.

[15] Z. Gong, R. Ying, F. Wen, J. Qian, and P. Liu. Tightly coupled integration of gnss and vision slam using 10-dof optimization on manifold. *IEEE Sensors Journal*, 19(24):12105–12117, 2019.

[16] T. Qin, P. Li, and S. Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018.

[17] G. Xie et al. Principles of gps and receiver design. *Publishing House of Electronics Industry, Beijing*, 7:61–63, 2009.

[18] K. J. Wu, C. X. Guo, G. Georgiou, and S. I. Roumeliotis. Vins on wheels. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5155–5162, 2017.

[19] S. Ch'ng, A. Khosravian, A. Doan, and T. Chin. Outlier-robust manifold pre-integration for ins/gps fusion. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7489–7496, 2019.

[20] S. Agarwal, K. Mierle, and Others. Ceres solver. `http://ceres-solver.org`.

[21] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale. Keyframe-based visual–inertial odometry using nonlinear optimization. *The International Journal of Robotics Research*, 34(3):314–334, 2015.

[23] W. Wen, Y. Zhou, G. Zhang, S. Fahandezh-Saadi, X. Bai, W. Zhan, M. Tomizuka, and L. T. Hsu. Urbanloco: A full sensor suite dataset for mapping and localization in urban scenes. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2310–2316, 2020.

[24] T. Takasu. Rtklib: An open source program package for gnss positioning. *Tech. Rep., 2013. Software and documentation*, 2011.

[25] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 573–580, 2012.

[26] G. Cioffi and D. Scaramuzza. Tightly-coupled fusion of global positional measurements in optimization-based visual-inertial odometry. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5089–5095, 2020.

[22] J. Liu, W. Gao, and Z. Hu. Visual-inertial odometry tightly coupled with wheel encoder adopting robust initialization and online extrinsic calibration. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5391–5397, 2019.