GNSS-Stereo-Inertial SLAM for Arable Farming

Javier Cremona[a,*], Javier Civera[b], Ernesto Kofman[a], Taihú Pire[a]

[a] *CIFASIS, French Argentine International Center for Information and Systems Sciences (CONICET-UNR), Rosario, Argentina*
[b] *University of Zaragoza, Zaragoza, Spain*

Abstract

The accelerating pace in the automation of agricultural tasks demands highly accurate and robust localization systems for field robots. Simultaneous Localization and Mapping (SLAM) methods inevitably accumulate drift on exploratory trajectories and primarily rely on place revisiting and loop closing to keep a bounded global localization error. Loop closure techniques are significantly challenging in agricultural fields, as the local visual appearance of different views is very similar and might change easily due to weather effects. A suitable alternative in practice is to employ global sensor positioning systems jointly with the rest of the robot sensors. In this paper we propose and implement the fusion of GNSS, stereo views and inertial measurements for localization purposes. Specifically, we incorporate, in a tightly-coupled manner, GNSS measurements into the stereo-inertial ORB-SLAM3 pipeline. We thoroughly evaluate our implementation in the sequences of the Rosario dataset, recorded by an autonomous robot in soybean fields, and our own in-house data. Our data includes measurements from a conventional GNSS, rarely included in evaluations of state-of-the-art approaches. We characterize the performance of GNSS-Stereo-inertial SLAM in this application case, reporting pose error reductions between 10 % and 30 % compared to visual-inertial and loosely-coupled GNSS-stereo-inertial baselines. In addition to such analysis, we also release the code of our implementation as open source.

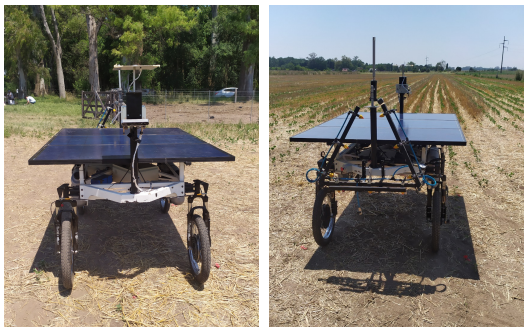
Keywords: GNSS-Stereo-Inertial SLAM, Agricultural Robotics, Precision Agriculture.

1. Introduction

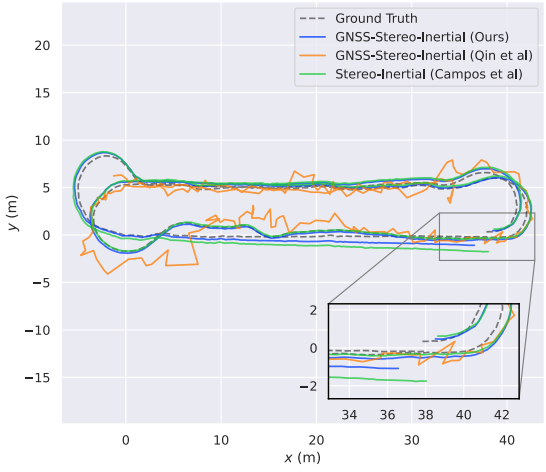
Over the last decades, several agricultural tasks such as sowing, weed detection and removal or harvesting are being progressively automated targeting sustainable and environmentally friendly production. The use of autonomous robots in an agricultural environment has gained relevance, as it enables an efficient use of resources (Auat Cheein & Carelli, 2013; Bac et al., 2014). In general, in order to fully automate these and other agricultural tasks, the robot needs to know its pose relative to the environment in which it is navigating.

A localization system must have a very high degree of robustness and accuracy for a mobile robot to navigate safely without damaging the environment or itself. For most environments and tasks, a single sensor may not offer a sufficiently reliable robot pose estimate. As a few illustrative examples, GNSS sensors in outdoor environments do not accumulate error (drift) but they present considerable variance in their global position readings and may suffer frequent signal loss. State-of-the-art methods based on visual sensors perform badly if images have insufficient or repetitive textures, which is common in agricultural environments. Lighting can also be a problem if it is insufficient or excessive, and abrupt robot motion can cause image blur that degrades the estimation performance. Finally, interoceptive sensors that measure the internal state of the robot, such as the encoders in the wheel motors or inertial measurement units (IMU), are accurate for short-term motion estimation but

*Corresponding author.
Email addresses:
cremona(at)cifasis-conicet(dot)gov(dot)ar (Javier Cremona), jcivera(at)unizar(dot)es (Javier Civera), kofman(at)cifasis-conicet(dot)gov(dot)ar (Ernesto Kofman), pire(at)cifasis-conicet(dot)gov(dot)ar (Taihú Pire)



(a) (b)



(c)

Figure 1: (a) and (b): Frontal and back views of our field robot and the arable field environment in which we navigate. (c): Trajectory estimated by our GNSS-stereo-inertial SLAM framework, along with GNSS-RTK ground truth, visual-inertial ORB-SLAM3 (Campos et al., 2021) and VINS-Fusion (Qin et al., 2019)

drift after a few metres. Summing up, as all sensors have different and complementary advantages and disadvantages, it is essential for field robotics to properly fuse the measurements of multiple sensors to achieve robust and accurate pose estimates. This is particularly relevant to allow the robot to navigate over long periods of time (long-term navigation) and to keep the error bounded locally and globally.

SLAM, standing for Simultaneous Localization and Mapping, stands for the set of methods targeting global localization and mapping from a set of onboard sensors in a mobile agent (Cadena et al., 2016). A large number of visual-inertial SLAM pipelines have been proposed in the last decade (Mur-Artal & Tardós, 2017a; Qin et al., 2018; Campos et al., 2021). Many of them demonstrate high accuracy and robustness in indoor and urban environments. However, when it comes to the agricultural environment, they present problems in correctly estimating the pose of the robot. Among others, agricultural environments are challenging for visual navigation due to insufficient and/or repetitive texture and direct sunlight. Adding inertial measurements provides a slight improvement in the estimation. Nevertheless, as shown in (Cremona et al., 2022), state-of-the-art visual-inertial systems accumulate significant errors after navigating a few minutes on arable lands. Robust SLAM systems such as ORB-SLAM3 (Campos et al., 2021) can eliminate drift when revisiting already mapped places, but the so-called loop closing offers a poor performance on agricultural fields due to insufficiently discriminative visual appearances. A reasonable alternative, that we use in this work, is to employ measurements from global positioning sensors such as GNSS to allow the robot to navigate for long periods without accumulating drift.

This paper presents a GNSS-stereo-inertial SLAM implementation that fuses GNSS, visual and inertial measurements using a tightly-coupled approach. Specifically, we extend the state-of-the-art ORB-SLAM3 (Campos et al., 2021) with GNSS factors. The global positioning measurements are incorporated into the mapping thread, so that it performs periodic corrections in the local map and hence also correct the current camera pose in the tracking thread. In this manner, we can achieve drift-less trajectories without depending on the ability of the system to close loops based on visual appearance. We evaluated our implementation on the agricultural dataset known as Rosario

Dataset (Pire et al., 2019) and an additional in-house dataset, which contains data from a wheeled robot in a soybean field (see Figure 1a-1b for a picture of our robot). In both cases, we show how our implementation is able to effectively fuse GNSS readings outperforming the original stereo-inertial ORB-SLAM3. The contribution of the work can be summarized as follows:

- Implementation of a GNSS-Stereo-Inertial framework.
- Evaluation of our GNSS-Stereo-Inertial framework tightly-coupled fusion in agricultural environments, incorporating real conventional GNSS measurements instead of simulated ones, which are rarely included in evaluations of state-of-the-art approaches.
- Public release of our implementation as open-source[1] , in order to facilitate its usage, extensions and comparisons and evaluations by the robotics community.

The article is organized as follows: Section 2 discusses related work on multi-modal sensor fusion. In Section 3, we describe the proposed GNSS-Stereo-Inertial framework. In Section 4, we present and discuss the experimental results of our GNSS-Stereo-Inertial implementation on real data in an agricultural field. Finally, we present our conclusions in Section 5.

2. Related Work

Sensor fusion methods can be broadly divided into two groups, *loosely-coupled* and *tightly-coupled*. Loosely-coupled methods are those that omit correlations between measurements from different sensors. This simplifies the fusion, as the estimation from each sensor can run separately and the estimates be fused afterwards. Most of these approaches are based on filters, such as the Extended Kalman Filter (EKF), that sequentially updates the system state integrating previous information. This is however suboptimal compared to tightly-coupled methods (Strasdat et al., 2012), which model the correlations between state variables and sensor measurements. In this last case, the measurements from all sensors are jointly integrated in the same optimization problem. As a drawback, tightly-coupled solutions generally have a higher computational cost than loosely-coupled ones. In the rest of the section, we refer the most related works to ours, from the loosely-coupled to the tightly-coupled ones.

Weiss et al. (2012) propose an EKF-based estimation method for Micro Air Vehicles (MAV). Its contribution is a modular loosely-coupled method that is capable of fusing visual, inertial and external positioning sensor (such as GPS or a laser telemetry tracking system) information. The results show that the proposed method allows state predictions to be made up to 1 kHz for MAV control tasks, being robust to low frequency measurements of 1 Hz, delays of up to 500 ms in the measurements and noise with standard deviations up to 20 cm. Shen et al. (2014) present a similar loosely-coupled approach but using an Unscented Kalman Filter (UKF), in order to better address the non-linearities in the sensor models. Wei et al. (2011) use stereo cameras to estimate the motion of a ground robot, considering motions only in the horizontal plane, and using an EKF to fuse global GPS measurements in a loosely-coupled manner, reducing the drift. Won et al. (2014a,b) propose a selective integration method for GNSS, visual and inertial measurements to improve localization accuracy under GNSS-challenged environments. The authors introduced a new performance index to recognize poor environments based on the geometrical distribution of the satellites and the local image features.

Li et al. (2019) present a multi-state constraint Kalman filter (MSCKF) approach to fuse monocular, inertial and raw GNSS-RTK measurements. The MSCKF makes use of a measurement model that does not require to include the feature landmarks in the state vector of the EKF, improving the robustness and computational complexity of the system. Salehi et al. (2017) use a mixture of tightly-coupled and loosely-coupled techniques for the fusion of visual and GPS measurements. An exhaustive optimization restricted to a temporal window of recent visual measurements is used, while measurements outside the window are marginalized by obtaining estimates of relative motion between poses. This allows improving computational times, preventing the computational complexity to scale. Yu et al. (2019) present a GPS-assisted visual-inertial estimation framework for omnidirectional platforms. It extends VINS-MONO (Qin

[1] `https://github.com/CIFASIS/gnss-stereo-inertial-fusion`

et al., 2018) to support multiple cameras, fuses visual and inertial information in a tightly-coupled manner, combined with a loosely-coupled approach to incorporate the measurements provided by GPS. Later, the same authors present GVINS (Cao et al., 2022), a framework based on non-linear optimization. GVINS tightly fuses GNSS raw measurements with visual and inertial information for state estimation. The GNSS pseudorange and Doppler shift measurements are modelled under a probabilistic factor graph framework along with visual and inertial constraints. The same approach is applied in (Liu et al., 2021).

Lynen et al. (2013) present Multi-Sensor Fusion (MSF), a modular sensor fusion system based on an EKF filter where inertial information is used at the prediction step. The information coming from the different sensors is modeled in a general manner as relative and/or absolute pose estimates, thus allowing to fuse measurements coming from a large number of sensors using a loosely-coupled approach. The work places particular emphasis on modelling the temporal arrival of the measurements by applying a technique known as Stochastic Cloning able to address asynchronous sensor fusion. Mascaro et al. (2018) present the Graph-Optimization based Multi-Sensor Fusion (GOMSF) framework which solves the fusion of pose estimates in different coordinate systems. Visual-inertial estimates from the MSF in local coordinates are merged with measurements in global coordinates from a GPS.

Lee et al. (2020) present a GPS-VIO system that fuses visual-inertial data with intermittent GPS measurements. The authors proposed a GPS-IMU online calibration approach for the time offset and extrinsics estimation. In (Boche et al., 2022) a tightly coupled visual-inertial-GPS system is presented. The system is based on OKVIS2 (Leutenegger, 2022). In the work a new global reference frame initialization has been introduced. It incorporates measurement uncertainties to decide whether the extrinsic transformation between the global and visual-inertial reference frame becomes observable. (Han et al., 2022) implement a system that integrates GNSS measurements into ORB-SLAM3 (Campos et al., 2021). In contrast to our research, their approach defines a residual that combines GNSS and IMU pre-integration measurements, along with implementing online calibration for the GNSS-IMU extrinsic. Remarkably, their system was evaluated within indoor and urban environments, where the GNSS signal can be susceptible to disruptions, but without facing the visual challenges typically present in agricultural fields.

In contrast to the previously mentioned works, this paper presents a tightly-coupled GNSS-stereo-inertial SLAM to tackle localization in agricultural environments. The proposed framework extends the Visual-Inertial SLAM system ORB-SLAM3 (Campos et al., 2021) with GNSS measurements. We built on top of ORB-SLAM3 since it has a fair performance in agricultural environments (Cremona et al., 2022). Our implementation is publicly released as open source to facilitate its use, extension and reproduction of the results by the robotics community.

3. Proposed GNSS-Stereo-Inertial Framework

This section presents the technical aspects of our implementation. Firstly, we introduce the notation and conventions adopted that are necessary to fully detail the model of our GNSS factor. Later, we briefly introduce ORB-SLAM3 (Campos et al., 2021), the state-of-the-art framework Visual-Inertial SLAM that we use in our method. We refer the reader to the original ORB-SLAM3 publication for the full details on such framework. Finally, we describe the formulation of our GNSS factor.

3.1. Notation

Figure 2 shows the coordinate frames used in this work. W represents the world frame and B represents the body frame, that we place in the IMU sensor. $\mathbf{a}^S$ represents the coordinates of a geometry entity $\mathbf{a}$ with respect to the reference frame S. $\mathbf{R}_B^W \in SO(3)$ refers to the rotation of B with respect to W, and $\mathbf{t}_B^W \in \mathbb{R}^3$ represents the translation of the reference frame B expressed in the frame W. The rigid transformation formed by the rotation $\mathbf{R}_B^W$ and the translation $\mathbf{t}_B^W$ is denoted as $\mathbf{T}_B^W \in SE(3)$, and transforms points in homogeneous coordinates from the reference frame B to the reference frame W. For global positioning measurements, $\mathbf{t}_A^B \in \mathbb{R}^3$ is the position of the GNSS antenna in the body frame, and is assumed to be known from a calibration stage. All GNSS measurements are transformed to the local Cartesian frame that we denote as A_0. We detail below how we choose such reference frame.

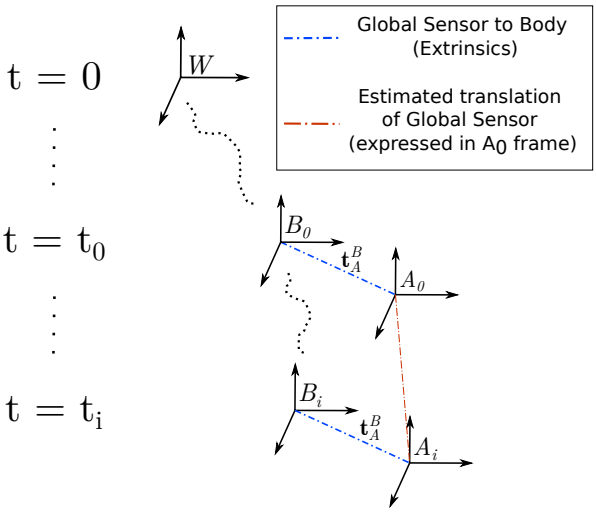


Figure 2: Reference Frames used in this work. W represents the world frame and B_i represents the body frame at time t_i. t_0 corresponds to time of the arrival of the first GNSS measurement. A_0 is an East-North-Up (ENU) local Cartesian frame whose position is given by this first GNSS measurement, i.e. the position of the antenna at time t_0. The position of the GNSS antenna in the body frame is represented with a translation $\mathbf{t}_A^B \in \mathbb{R}^3$ and shown with a blue line, and can be obtained from the calibration of the system. The red line represents the estimated translation of the GNSS antenna from time t_0 to time t_i. This 3D vector is compared with the GNSS measurements in the GNSS error residual $\mathbf{r}_{\mathcal{G}_i}$. Both vectors are expressed in A_0 frame.

3.2. ORB-SLAM3

ORB-SLAM3 is a state-of-the-art visual-inertial SLAM framework evolved from ORB-SLAM2 (Mur-Artal & Tardós, 2017b) and ORB-SLAM-VI (Mur-Artal & Tardós, 2017a). With respect to ORB-SLAM-VI, ORB-SLAM3 proposes a substantially more robust inertial initialization based on maximum-a-posteriori estimates. As it is common in current SLAM systems, the processing is split into multiple threads to exploit multi-core architectures. Specifically, ORB-SLAM3 implements a tracking thread, a local mapping thread and a loop closure and map merging thread. The tracking thread estimates the pose of the current frame by minimizing the reprojection error and incorporating IMU constraints into the optimization by preintegration (Forster et al., 2017). It also contains the heuristics for deciding whether a frame becomes a keyframe. The mapping thread main task is a visual-inertial bundle adjustment on a sliding window of keyframes, although it also performs auxiliary map management tasks such as point and keyframe culling. Finally, the loop closure and map merging thread ensures the global consistency of large maps by recognizing revisited places and correcting the drift, and joining separate maps if a common overlap is detected.

From the results in (Cremona et al., 2022), ORB-SLAM3 presents an acceptable accuracy in arable lands for short camera trajectories, but long-term navigation is still challenging. The authors propose a novel loop closure algorithm to correct the drift. However, even with such improvement, loop closure keeps being challenging due to the similarity in appearance of the local visual features. As a result, visual SLAM systems may accumulate drift when loop closures are not detected or the estimation may be corrupted by false loop detections.

3.3. GNSS-Stereo-Inertial Fusion

In this work, we formulate a tightly-coupled approach for fusing visual, inertial and GNSS data. Firstly, GNSS measurements are associated to the timestamp of a keyframe according to their temporal proximity. If there is a keyframe with a temporal difference under a specific threshold, the GNSS constraint is set to this keyframe. GNSS readings that are not close in time to any keyframe are discarded (see an illustration of this approach in Figure 3). While this is an approximation, we found that, given the high variance of conventional GNSS, a sufficiently small threshold and appropriate keyframe management policy makes its effect negligible.

The first GNSS reading that is associated with a keyframe determines the position of A_0, the Cartesian frame for our global position measurements (see Figure 2). We choose A_0 as a East-North-Up (ENU) local Cartesian frame. The subsequent GNSS measurements are transformed to be expressed in A_0, and we refer to them as $\hat{\mathbf{z}}_i$, where t_i is the timestamp of the corresponding keyframe. This is done once the IMU is initialized. If the map is reset, the process of selecting A_0 is repeated.

Our GNSS-Stereo-Inertial fusion is done in the local bundle adjustment of a sliding window of keyframes and 3D points observed from them. Figure 4 shows the factor graph corresponding to such optimization. The state variables to optimize are $\mathcal{X} = \{\mathcal{X}_B, \mathcal{L}\}$, where $\mathcal{X}_B = [\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N]$ is the set of sensor states for a window covering the last N keyframes and $\mathcal{L} = [\mathbf{y}_1, \ldots, \mathbf{y}_j, \ldots, \mathbf{y}_M]$ is the set of landmarks states that were measured during those last N keyframes. The sensor state $\mathbf{x}_i$ at

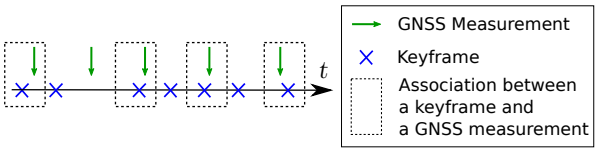


Figure 3: Representation of the temporal association between keyframes and GNSS measurements. Keyframes are depicted with blue crosses on the temporal line and GNSS measurements are depicted with green arrows. The dotted box represents the association between a keyframe and a measurement. Note that there are keyframes without a corresponding GNSS measurement and that GNSS measurements can be discarded if they are further than a specific temporal threshold from any keyframe.

the time instant i is

$$\mathbf{x}_i = [\mathbf{T}_{B_i}^W, \mathbf{v}_i^\top, \mathbf{b}_{a_i}^\top, \mathbf{b}_{g_i}^\top], \tag{1}$$

which contains the sensor rigid transformation with respect to the world frame $\mathbf{T}_{B_i}^W \in SO(3)$, its local velocity $\mathbf{v}_i \in \mathbb{R}^3$ and the accelerometer and gyroscope bias $\mathbf{b}_{a_i} \in \mathbb{R}^3$ and $\mathbf{b}_{g_i} \in \mathbb{R}^3$. Landmarks are represented by their Euclidean coordinates in the world frame, i.e., $\mathbf{y}_j = [X^W, Y^W, Z^W]^\top \in \mathbb{R}^3$

In comparison to ORB-SLAM3, a GNSS error term is added to the cost function. Note that, as shown in Figure 3, some keyframes may not have an associated GNSS measurement. Then, our GNSS-Stereo-Inertial mapping optimization can be stated as follows

$$\begin{aligned}\hat{\mathcal{X}} = \arg\min_{\mathcal{X}} \Bigg(& \sum_{i=1}^{N} \|\mathbf{r}_{\mathcal{I}_{i-1,i}}\|^2_{\Sigma^{-1}_{\mathcal{I}_{i-1,i}}} + \\ & + \sum_{j=1}^{M} \sum_{i \in \mathcal{K}_j} \rho\left(\|\mathbf{r}_{\mathcal{V}_{ij}}\|_{\Sigma^{-1}_{\mathcal{V}_{ij}}} + \right) \\ & + \sum_{i \in \mathcal{N}^*} \rho\left(\|\mathbf{r}_{\mathcal{G}_i}\|_{\Sigma^{-1}_{\mathcal{G}_i}} + \right) \Bigg),\end{aligned} \tag{2}$$

where $\mathcal{N}^*$ is the set of keyframes that have an associated GNSS measurement. The three addends correspond, respectively, to the inertial, visual and GNSS constraints. For the sake of completeness, we will detail the three of them, although the first two are used exactly as proposed in ORB-SLAM3 and the third one is our novel contribution.

The inertial residual is defined as follows

$$\mathbf{r}_{\mathcal{I}_{i-1,i}} = [\mathbf{r}^\top_{\Delta\mathbf{R}_{i-1,i}}, \mathbf{r}^\top_{\Delta\mathbf{v}_{i-1,i}}, \mathbf{r}^\top_{\Delta\mathbf{p}_{i-1,i}}]^\top, \tag{3}$$

where $\mathbf{r}_{\Delta\mathbf{R}_{i-1,i}}$, $\mathbf{r}_{\Delta\mathbf{v}_{i-1,i}}$ and $\mathbf{r}_{\Delta\mathbf{p}_{i-1,i}}$ correspond to orientation, velocity and position residuals that have the following form

$$\begin{aligned}\mathbf{r}_{\Delta\mathbf{R}_{i-1,i}} &= \log\left(\Delta\mathbf{R}^\top_{i-1,i}\mathbf{R}^\top_{i-1}\mathbf{R}_i\right) \\ \mathbf{r}_{\Delta\mathbf{v}_{i-1,i}} &= \mathbf{R}_i^\top\left(\mathbf{v}_i - \mathbf{v}_{i-1} - \mathbf{g}\Delta t_{i-1,i}\right) - \Delta\mathbf{v}_{i-1,i} \\ \mathbf{r}_{\Delta\mathbf{p}_{i-1,i}} &= \mathbf{R}_i^\top\left(\mathbf{p}_i - \mathbf{p}_{i-1} - \mathbf{v}_i\Delta t_{i-1,i} - \frac{1}{2}\mathbf{g}\Delta t^2_{i-1,i}\right) - \\ &\quad - \Delta\mathbf{p}_{i-1,i}.\end{aligned} \tag{4}$$

The terms denoted as $\Delta\mathbf{R}_{i-1,i}$, $\Delta\mathbf{v}_{i-1,i}$ and $\Delta\mathbf{p}_{i-1,i}$ come from the preintegration of the IMU readings between the time instants $i-1$ and i, and are computed together with their on-manifold covariance $\Sigma_{\mathcal{I}_{i-1,i}}$ according to (Forster et al., 2017). $\mathbf{g}$ stands for the gravity direction, which is set at the system bootstrapping.

The visual residual $\mathbf{r}_{\Delta\mathbf{v}_{i-1,i}}$ is

$$\mathbf{r}_{\mathcal{V}_{ij}} = \mathbf{u}_{ij} - \pi\left(\mathbf{T}_B^C \mathbf{T}_B^{W\,-1} \tilde{\mathbf{y}}_j\right), \tag{5}$$

where $\tilde{\mathbf{y}}_j$ stands for the homogeneous representation of the j^{th} landmark, $\pi(\cdot)$ for the pinhole projection model of a 3D point in homogeneous coordinates in a stereo image, and $\mathbf{u}_{ij}$ the measured image coordinates of the j^{th} landmark in the i^{th} stereo keyframe. The visual covariance of image landmarks $\Sigma_{\mathcal{V}_{ij}}$ is set to the standard 1-pixel standard deviation isotropic Gaussian.

Finally, the GNSS error residual is

$$\mathbf{r}_{\mathcal{G}_i} = \hat{\mathbf{z}}_i - \mathbf{R}_W^{A_0}\left(\mathbf{R}_{B_i}^W \mathbf{t}_A^B + \mathbf{t}_{B_i}^W - \left(\mathbf{R}_{B_0}^W \mathbf{t}_A^B + \mathbf{t}_{B_0}^W\right)\right). \tag{6}$$

The second term represents the translation vector of the global sensor (in this case, the GNSS antenna) at time instant i in the reference frame A_0, as can be seen in Figure 2. $\mathbf{R}_{B_0}^W$ and $\mathbf{t}_{B_0}^W$, which are the relative rotation and translation between the body and the world frame at time t_0, are kept constant during the optimization. $\mathbf{R}_W^{A_0}$ is computed by aligning the first 20 GNSS measurements with the poses estimated by ORB-SLAM3 in the same time period using Umeyama's method (Umeyama, 1991). After estimating this rotation, it is kept fixed during the whole optimization process. The covariance matrix $\Sigma_{\mathcal{G}_i}$ is set from the specifications sheet of our GNSS device in each Cartesian axis

$$\Sigma_{\mathcal{G}_i} = \begin{bmatrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_y^2 & 0 \\ 0 & 0 & \sigma_z^2 \end{bmatrix}. \tag{7}$$

This covariance matrix is defined relative to a tangential plane through the GNSS reported position. The values are expressed in ENU frame. Finally,

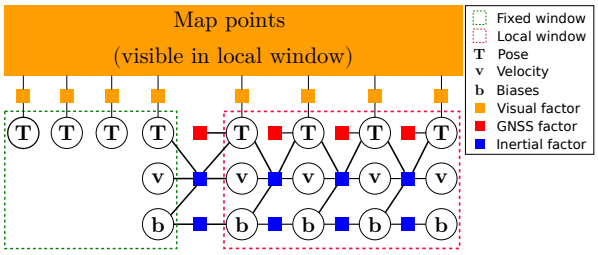


Figure 4: Factor Graph corresponding to the Local Bundle Adjustment of our GNSS-Stereo-Inertial SLAM. In comparison to ORB-SLAM3, a GNSS factor (in red) is added to the cost function. The Local Window is composed by the N last keyframes. The fixed window contains keyframes outside the local window that are connected in the covisibility graph to any local keyframe. These keyframes remain fixed during optimization. Additionally, the keyframe N+1 is included in the fixed window as it constrains the IMU states.

the Jacobian with respect to the pose error state is defined as

$$\frac{\partial \mathbf{r}_{\mathcal{G}_i}}{\partial \delta \mathbf{T}_{B_i}^{W}} = \left[\mathbf{R}_{W}^{A_0} \mathbf{R}_{B_i}^{W} \left[\mathbf{t}_{A}^{B} \right]^{\times} \quad -\mathbf{R}_{W}^{A_0} \right], \qquad (8)$$

where δ indicates that the derivative is computed with respect to a right perturbation in the pose.

4. Experimental Evaluation

This section shows the experimental results of the implementation proposed in Section 3. The framework is evaluated on the Rosario Dataset (Pire et al., 2019), a set of agricultural data captured by a weed removal robot. Later, an evaluation of the system in a soybean field is presented, using the same weed removal robot. The difference between the latter test and the evaluation on the Rosario Dataset is that new sensors are available, including measurements from a conventional GNSS. For the temporal association between keyframes and GNSS measurements explained in Section 3.3, a threshold of 0.035 seconds is chosen in all experiments.

4.1. Rosario Dataset

The Rosario Dataset (Pire et al., 2019) is a set of data captured by the sensors of a weed removal robot developed by the CIFASIS institute (CONICET-UNR) in Rosario, Argentina. It is composed of six sequences captured in a soybean field. The sequences contain stereo images of 672×376 px captured at 15 Hz, measurements from an IMU with a frequency of 142 Hz including gyroscope and accelerometer, wheel odometry obtained at 10 Hz and GNSS-RTK measurements at 5 Hz. The GNSS-RTK data are used as positional ground-truth.

Table 1: Mean (standard deviation) of the Absolute Trajectory Error (ATE) [m] for stereo-inertial ORB-SLAM3 (Campos et al., 2021), a loosely-coupled GNSS-stereo-inertial system (Qin et al., 2019) and our tightly-coupled GNSS-stereo-inertial framework in the six sequences of the Rosario Dataset. Best results are in **bold.**

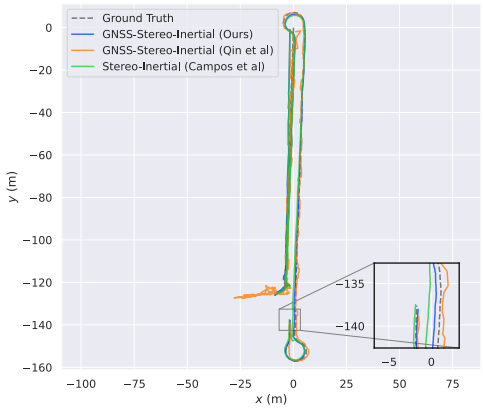
Sequence	Stereo-Inertial (Campos et al., 2021)	GNSS-Stereo-Inertial (Qin et al., 2019)	GNSS-Stereo-Inertial (Ours)
01	0.90 (0.34)	1.44 (2.06)	**0.86 (0.26)**
02	1.33 (0.75)	**0.90 (0.40)**	0.94 (0.56)
03	1.12 (0.65)	1.34 (1.91)	**0.99 (0.56)**
04	1.09 (0.65)	1.42 (1.20)	**1.04 (0.60)**
05	0.89 (0.55)	1.43 (1.91)	**0.76 (0.38)**
06	2.48 (1.40)	1.81 (0.87)	**1.23 (0.70)**

Since the Rosario Dataset does not have conventional GNSS measurements, we simulate noisy GNSS measurements by corrupting the ground-truth with zero-mean Gaussian noise, as in (Cioffi & Scaramuzza, 2020). We use isotropic Gaussian noise $\mathbf{n}_p \sim \mathcal{N}\left(\mathbf{0}, \sigma_p^2 \cdot \mathbf{I}\right)$, with a standard deviation $\sigma_p = 0.5\,\text{m}$. We selected this value from observing the covariance of the conventional GNSS used in the experiments in Section 4.2.

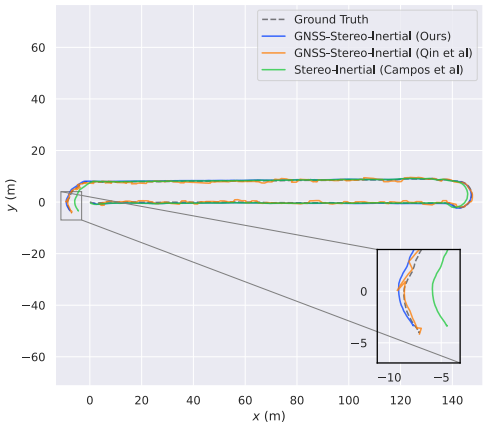
We compared our GNSS-Stereo-Inertial implementation against Stereo-Inertial ORB-SLAM3 and a loosely-coupled GNSS-Stereo-Inertial system known as VINS-Fusion (Qin et al., 2019). VINS-Fusion was chosen because it is a state-of-the-art system that takes as input the same GNSS measurements as our system, i.e. latitude, longitude and altitude. Each system was run five times in each of the Rosario sequences, and Table 1 presents the lowest ATE error of the five executions for each framework. ATE has been computed after the estimated trajectories were aligned with the ground-truth GNSS readings using Umeyama's method (Umeyama, 1991). The corresponding trajectories are presented in Figure 5.

4.2. Data with conventional GNSS in soybean fields

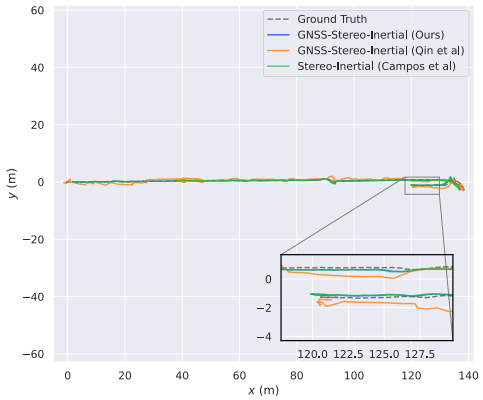
In the experiments from the previous section, noisy GNSS measurements had to be simulated from GNSS-RTK ones, as the dataset does not contain conventional GNSS measurements. In this section we present an evaluation with conventional GNSS measurements. For this, we equipped our weed removal robot with such sensor and deployed it again in a soybean field. On board the robot there is a ZED stereo camera which captures images 1280×720 px at 15 Hz, an Emlid Reach GNSS operating at a frequency of 5 Hz, and an InvenSense MPU-9250 IMU set at 200 Hz. The covariance of



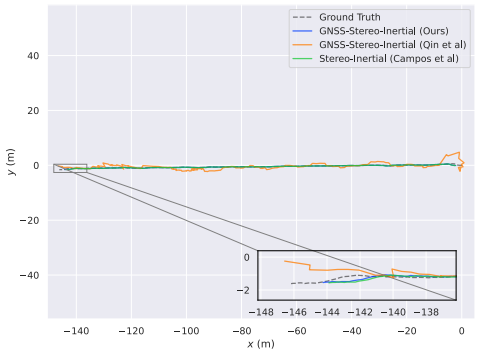
(a) Sequence 01



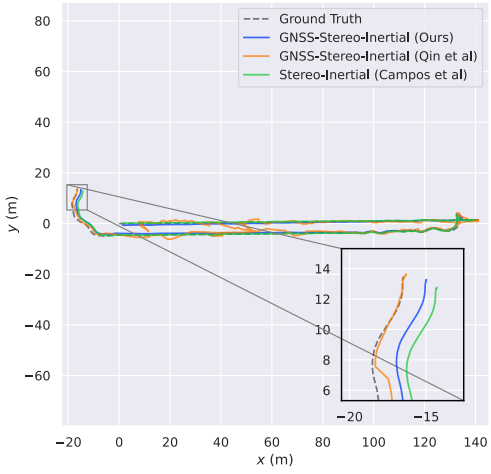
(b) Sequence 02



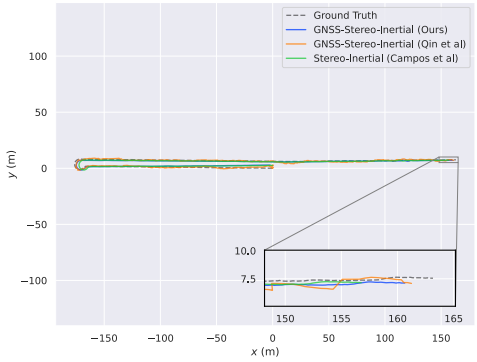
(c) Sequence 03



(d) Sequence 04



(e) Sequence 05



(f) Sequence 06

Figure 5: Results from Stereo-Inertial ORB-SLAM3 (Campos et al., 2021), a loosely-coupled GNSS-Stereo-Inertial system (Qin et al., 2019) and our tightly-coupled GNSS-Stereo-Inertial system on the Rosario Dataset. The estimated trajectories are aligned with the ground-truth using Umeyama's method (Umeyama, 1991).

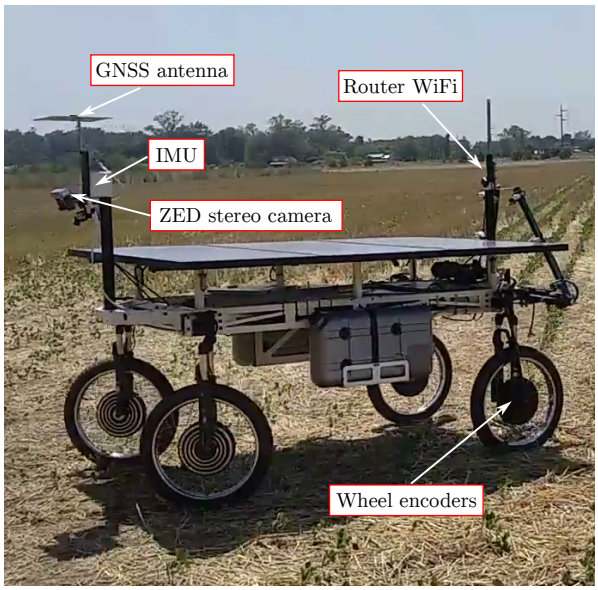


Figure 6: Weed removal robot used in our in-house dataset in a soybean field. We equipped the robot with a ZED stereo camera, an Emlid Reach GNSS receiver, and an InvenSense MPU-9250 IMU. Furthermore, wheel odometry can be obtained from the wheel encoders.

the conventional GNSS measurements is offered by the driver of the GNSS receiver. In addition, GNSS-RTK provides positional ground-truth. Figure 6 shows the robot configuration in the soybean field. We commanded the robot to record three data sequences. The corresponding GNSS-RTK trajectories are shown in Figure 7 and images samples captured by the ZED camera can be seen in Figure 8.

On this data we ran the three frameworks mentioned in the previous experiment. The results of this experiment are shown in Table 2, while the trajectories can be seen in the Figure 9. Estimated trajectories were aligned again with the ground-truth using Umeyama's method. Finally, the reconstructed map and the trajectory estimated by our our tightly-coupled GNSS-Stereo-Inertial SLAM for sequence B is shown in Figure 10, as a qualitative illustration of the mapping capability of our framework.

4.3. Discussion

As can be seen in the results, our implementation clearly outperforms the stereo-inertial configuration of ORB-SLAM3 and the loosely-coupled approach in (Qin et al., 2019). As a very relevant note, we ran the full stereo-inertial ORB-SLAM3

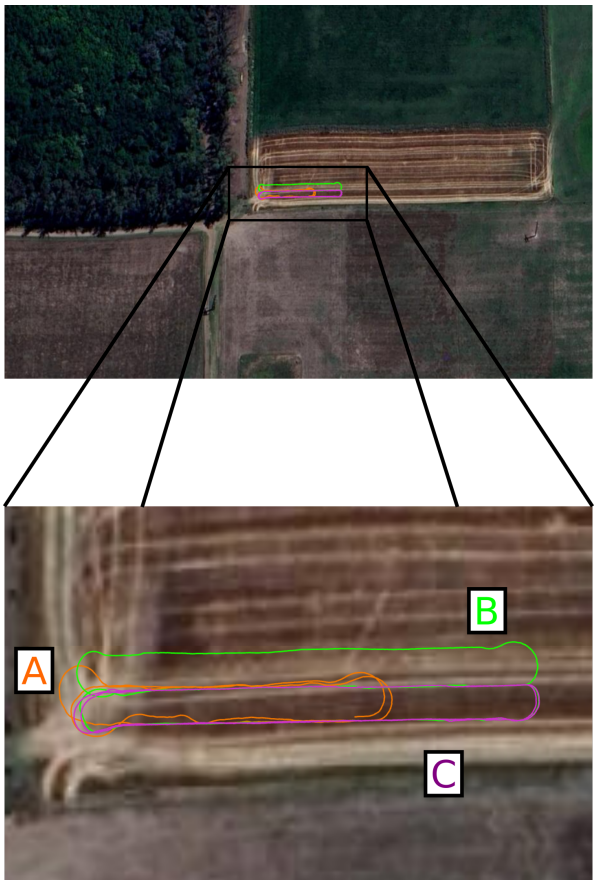


Figure 7: GNSS-RTK trajectories for the sequences A (orange), B (green) and C (purple) of our in-house recordings in the soybean field.

Table 2: Mean (standard deviation) of the Absolute Trajectory Error (ATE) [m] for stereo-inertial ORB-SLAM3 (Campos et al., 2021), a loosely-coupled GNSS-stereo-inertial system (Qin et al., 2019) and our tightly-coupled GNSS-stereo-inertial framework in the in-house recordings in soybean fields. Best results are in **bold.**

Sequence	Stereo-Inertial (Campos et al., 2021)	GNSS-Stereo-Inertial (Qin et al., 2019)	GNSS-Stereo-Inertial (Ours)
A	0.64 (0.33)	1.08 (0.78)	**0.44 (0.16)**
B	0.43 (0.18)	5.58 (3.57)	**0.36 (0.13)**
C	0.46 (0.12)	16.90 (7.67)	**0.39 (0.12)**

in our configuration sequences with loop closure capabilities and, with its configuration by default, it was unable to detect previously visited locations and hence close loops due to insufficiently discriminative visual appearances of the agricultural environment (*perceptual aliasing*). Although the default configuration for the loop closure parameters might be loosened to detect a higher number of loop closures, that would also produce a higher number of false positives (due again to percep-

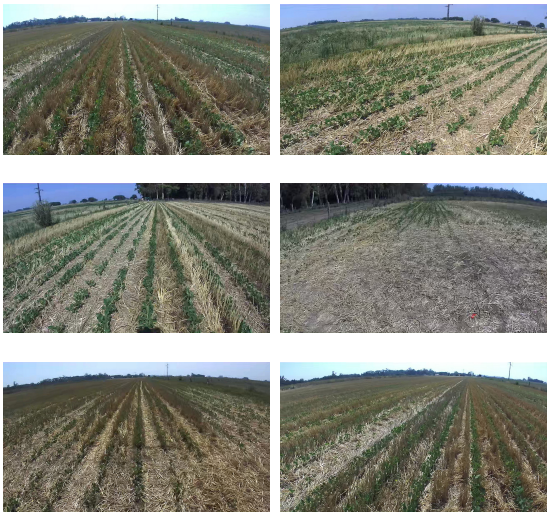


Figure 8: Sample images from our in-house dataset. Note the repetitive textures, a challenge for visual SLAM.

tual aliasing) that would corrupt the estimation. These challenges are the main motivation for incorporating global positioning sensors in agricultural environments, allowing to reduce the drift without depending on visual features. Very interestingly, we found in our experiments that only one third of the optimized keyframes had associated GNSS measurements. This may indicate that high-frequency GNSS measurements are not necessary to improve the estimation of visual-inertial SLAM, and a sparse subset of them might suffice to offer a reasonable performance.

Unlike the loosely-coupled system, our implementation returns smoother trajectories. Moreover, since the fusion is loosely-coupled, the global position measurements correct the estimate without considering the continuous motion of the robot and act as an interpolation between the underlying visual-inertial system and the GNSS measurement. Even though in sequence 02 of the Rosario Dataset, the loosely-coupled fusion system obtains a lower error, in the trajectory of the Figure 5 it can be observed that the estimation looks bumpy. Smooth pose estimation, like the one offered by our tightly-coupled approach, is more suitable for use in a navigation control algorithm.

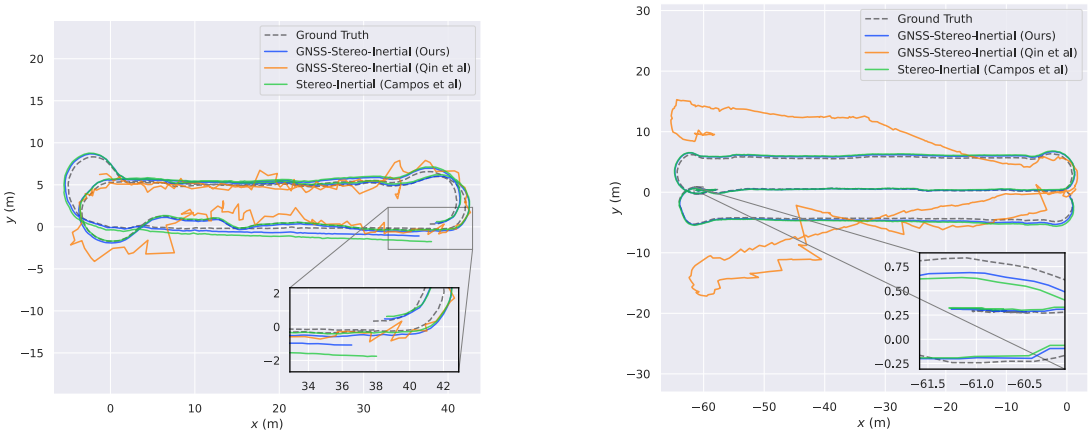
Regarding the experiment with conventional GNSS measurements, it should be pointed out that the loosely-coupled system lost the visual-inertial tracking in the three sequences. This indicates that it is important not only to focus on global measurements, but also to have a robust visual-inertial fusion. In our case, we use ORB-SLAM3 as the underlying system, as a result of having analyzed the performance of different visual-inertial systems in previous research (Cremona et al., 2022). As a conclusion, in addition to a tight coupling of the sensor data, the robustness of the visual-inertial estimates are also relevant for practical implementations in agricultural applications.

An important consideration is the modelling of the noise of GNSS measurements. Based on previous works (Cioffi & Scaramuzza, 2020; Boche et al., 2022), the uncertainty was modelled as additive isotropic Gaussian noise. This is a simple model that arises naturally from the GNSS device data, as the device drivers generally provide a covariance of the position. Other ways of modelling the noise of GNSS measurements in the context of pose estimation are worth studying, as when comparing the simulated signal in the section 4.1 experiment with the conventional GNSS signal used in the field experiments, differences in their behaviour were observed. When the conventional GNSS signal was inspected in detail, a bias was found, mainly at altitude, which could be verified by the GNSS-RTK. Therefore, this topic should be addressed in future work.

5. Conclusions

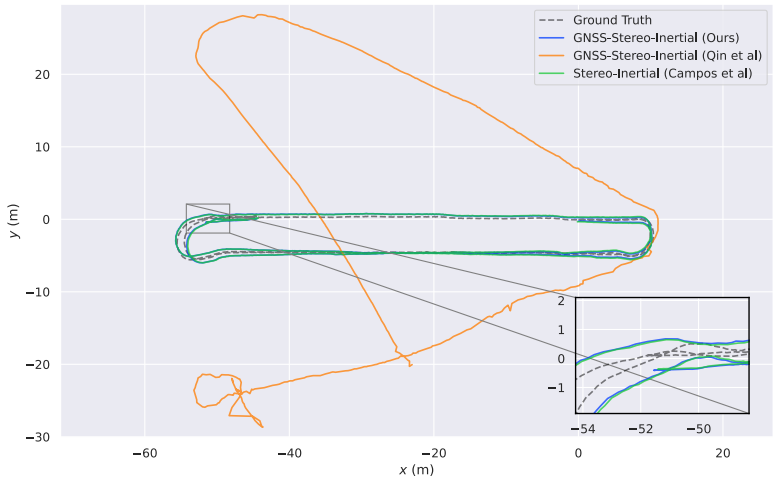
This work presents a GNSS-stereo-inertial SLAM framework that fuses in a tightly-coupled manner the information from a stereo camera, an IMU and a conventional GNSS sensor. In order to report the most competitive results, we implement our GNSS factor on top of the ORB-SLAM3 framework, the top performer in the evaluation of (Cremona et al., 2022). As we are motivated by long-term autonomous navigation in arable farms, we present results in the Rosario Dataset and in-house sequences from an agricultural robot. Very importantly, several works in the literature evaluate GNSS-stereo-inertial SLAM methods by emulating conventional GNSS measurements while we use a real sensor, so we are the first ones in reporting results in realistic conditions in agricultural scenes.

Our results show that there is a consistent gain in accuracy if GNSS measurements are tightly fused with visual and inertial ones in the local mapping optimization of a SLAM system. Very importantly, not only the localization errors are reduced but also



(a) Sequence A

(b) Sequence B



(c) Sequence C

Figure 9: Results from Stereo-Inertial ORB-SLAM3 (Campos et al., 2021), the loosely-coupled GNSS-Stereo-Inertial system of (Qin et al., 2019) and our tightly-coupled GNSS-Stereo-Inertial implementation on our in-house recordings in soybean fields, using conventional GNSS. Estimated trajectories are aligned with the ground-truth using Umeyama's method. Note the smaller errors of tightly-coupled approaches, and how our GNSS fusion improves over the stereo-inertial baseline.

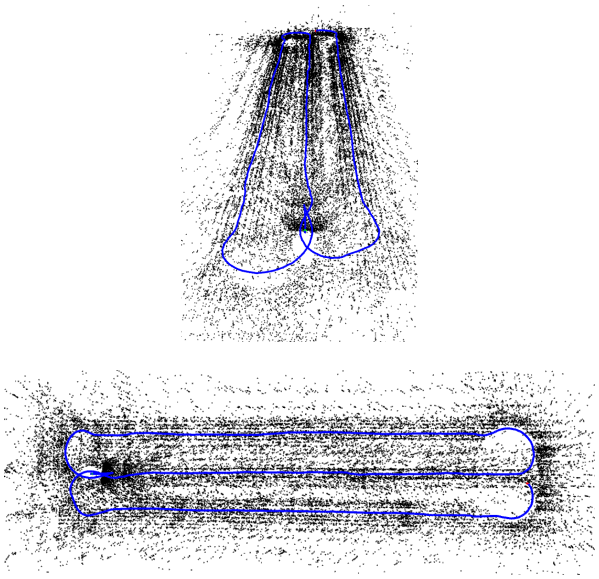


Figure 10: Map estimated by our our tightly-coupled GNSS-Stereo-Inertial SLAM for sequence B of our in-house dataset, seen from tilted and top views. The black points correspond to the tracked visual features, and the blue line to the estimated trajectory.

their variance between runs, indicating a looser dependence from the visual features used.

As an additional contribution of this work, we release our implementation for the benefit of the agricultural robotics community.

Acknowledgments

This work was partially supported by CONICET (Argentina) (PUE 0015-2016), by the Santa Fe province (Argentina) Government under Grant PEICID-2021-170, by the Spanish Government under Grants PGC2018-096367-B-I00 and PID2021-127685NB-I00 and by the Aragon Government under Grant DGA T45 17R/FSE.

References

Auat Cheein, F. A., & Carelli, R. (2013). Agricultural Robotics: Unmanned Robotic Service Units in Agricultural Tasks. *IEEE Industrial Electronics Magazine*, *7*, 48–58. doi:10.1109/MIE.2013.2252957.

Bac, C. W., van Henten, E. J., Hemming, J., & Edan, Y. (2014). Harvesting Robots for High-value Crops: State-of-the-art Review and Challenges Ahead. *Journal of Field Robotics*, *31*, 888–911. doi:10.1002/rob.21525.

Boche, S., Zuo, X., Schaefer, S., & Leutenegger, S. (2022). Visual-Inertial SLAM with Tightly-Coupled Dropout-Tolerant GPS Fusion. URL: https://arxiv.org/abs/2208.00709. doi:10.48550/ARXIV.2208.00709.

Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., & Leonard, J. J. (2016). Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age. *IEEE Trans. Robotics*, *32*, 1309–1332. doi:10.1109/TRO.2016.2624754.

Campos, C., Elvira, R., Rodríguez, J. J. G., M. Montiel, J. M., & D. Tardós, J. (2021). ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM. *IEEE Trans. Robotics*, *37*, 1874–1890. doi:10.1109/TRO.2021.3075644.

Cao, S., Lu, X., & Shen, S. (2022). GVINS: Tightly Coupled GNSS-Visual-Inertial Fusion for Smooth and Consistent State Estimation. *IEEE Transactions on Robotics*, *38*, 2004–2021. doi:10.1109/TRO.2021.3133730.

Cioffi, G., & Scaramuzza, D. (2020). Tightly-coupled Fusion of Global Positional Measurements in Optimization-based Visual-Inertial Odometry. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)* (pp. 5089–5095). doi:10.1109/IROS45743.2020.9341697.

Cremona, J., Comelli, R., & Pire, T. (2022). Experimental evaluation of Visual-Inertial Odometry systems for arable farming. *Journal of Field Robotics*, *39*, 1123–1137. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.22099. doi:10.1002/rob.22099.

Forster, C., Carlone, L., Dellaert, F., & Scaramuzza, D. (2017). On-Manifold Preintegration for Real-Time Visual–Inertial Odometry. *IEEE Trans. Robotics*, *33*, 1–21. doi:10.1109/TRO.2016.2597321.

Han, S., Deng, F., Li, T., & Pei, H. (2022). Tightly Coupled Optimization-based GPS-Visual-Inertial Odometry with Online Calibration and Initialization. arXiv:2203.02677.

Lee, W., Eckenhoff, K., Geneva, P., & Huang, G. (2020). Intermittent GPS-aided VIO: Online Initialization and Calibration. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)* (pp. 5724–5731). doi:10.1109/ICRA40945.2020.9197029.

Leutenegger, S. (2022). OKVIS2: Realtime Scalable Visual-Inertial SLAM with Loop Closure. URL: https://arxiv.org/abs/2202.09199. doi:10.48550/ARXIV.2202.09199.

Li, T., Zhang, H., Gao, Z., Niu, X., & El-sheimy, N. (2019). Tight Fusion of a Monocular Camera, MEMS-IMU, and Single-Frequency Multi-GNSS RTK for Precise Navigation in GNSS-Challenged Environments. *Remote Sensing*, *11*. doi:10.3390/rs11060610.

Liu, J., Gao, W., & Hu, Z. (2021). Optimization-Based Visual-Inertial SLAM Tightly Coupled with Raw GNSS Measurements. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)* (pp. 11612–11618). doi:10.1109/ICRA48506.2021.9562013.

Lynen, S., Achtelik, M. W., Weiss, S., Chli, M., & Siegwart, R. (2013). A robust and modular multi-sensor fusion approach applied to MAV navigation. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)* (pp. 3923–3929). doi:10.1109/IROS.2013.6696917.

Mascaro, R., Teixeira, L., Hinzmann, T., Siegwart, R., & Chli, M. (2018). GOMSF: Graph-Optimization Based Multi-Sensor Fusion for robust UAV Pose estimation. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)* (pp. 1421–1428). doi:10.1109/ICRA.2018.8460193.

Mur-Artal, R., & Tardós, J. D. (2017a). Visual-Inertial Monocular SLAM With Map Reuse. *(IEEE) Robotics and Automation Letters*, *2*, 796–803. doi:10.1109/LRA.2017.2653359.

Mur-Artal, R., & Tardós, J. D. (2017b). ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Trans. Robotics*, *33*, 1255–1262. doi:10.1109/TRO.2017.2705103.

Pire, T., Mujica, M., Civera, J., & Kofman, E. (2019). The Rosario Dataset: Multisensor Data for Localization and Mapping in Agricultural Environments. *Intl. J. of Robotics Research*, *38*, 633–641. doi:10.1177/0278364919841437.

Qin, T., Cao, S., Pan, J., & Shen, S. (2019). A General Optimization-based Framework for Global Pose Estimation with Multiple Sensors. arXiv:1901.03642.

Qin, T., Li, P., & Shen, S. (2018). VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator. *IEEE Trans. Robotics*, *34*, 1004–1020. doi:10.1109/TRO.2018.2853729.

Salehi, A., Gay-bellile, V., Bourgeois, S., & Chausse, F. (2017). A hybrid bundle adjustment/pose-graph approach to VSLAM/GPS fusion for low-capacity platforms. In *IEEE Intelligent Vehicles Symposium (IV)* (pp. 1728–1735). IEEE. doi:10.1109/IVS.2017.7995957.

Shen, S., Mulgaonkar, Y., Michael, N., & Kumar, V. (2014). Multi-sensor fusion for robust autonomous flight in indoor and outdoor environments with a rotorcraft MAV. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)* (pp. 4974–4981). doi:10.1109/ICRA.2014.6907588.

Strasdat, H., Montiel, J., & Davison, A. J. (2012). Visual SLAM: Why filter? *Image and Vision Computing*, *30*, 65–77. doi:10.1016/j.imavis.2012.02.009.

Umeyama, S. (1991). Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans. Pattern Anal. Machine Intell.*, *13*, 376–380. doi:10.1109/34.88573.

Wei, L., Cappelle, C., Ruichek, Y., & Zann, F. (2011). Intelligent Vehicle Localization in Urban Environments Using EKF-based Visual Odometry and GPS Fusion. *IFAC Proceedings Volumes*, *44*, 13776–13781. doi:10.3182/20110828-6-IT-1002.01965. 18th IFAC World Congress.

Weiss, S., Achtelik, M. W., Chli, M., & Siegwart, R. (2012). Versatile distributed pose estimation and sensor self-calibration for an autonomous MAV. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)* (pp. 31–38). doi:10.1109/ICRA.2012.6225002.

Won, D. H., Lee, E., Heo, M., Lee, S.-W., Lee, J., Kim, J., Sung, S., & Lee, Y. J. (2014a). Selective Integration of GNSS, Vision Sensor, and INS Using Weighted DOP Under GNSS-Challenged Environments. *IEEE Transactions on Instrumentation and Measurement*, *63*, 2288–2298. doi:10.1109/TIM.2014.2304365.

Won, D. H., Lee, E., Heo, M., Sung, S., Lee, J., & Lee, Y. J. (2014b). GNSS integration with vision-based navigation for low GNSS visibility conditions. *GPS Solutions*, *18*, 177–187. URL: https://doi.org/10.1007/s10291-013-0318-8. doi:10.1007/s10291-013-0318-8.

Yu, Y., Gao, W., Liu, C., Shen, S., & Liu, M. (2019). A GPS-aided Omnidirectional Visual-Inertial State Estimator in Ubiquitous Environments. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)* (pp. 7750–7755). doi:10.1109/IROS40897.2019.8968519.