A GPS-aided Omnidirectional Visual-Inertial State Estimator in Ubiquitous Environments

Yang Yu[1], Wenliang Gao[1], Chengju Liu[2], Shaojie Shen[1], and Ming Liu[1]

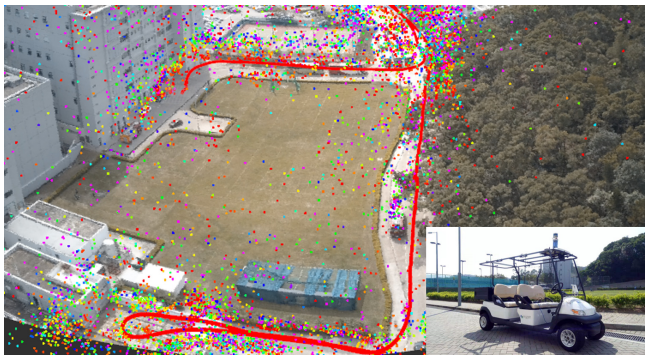
***Abstract*— The visual-inertial navigation system (VINS) has been a practical approach for state estimation in recent years. In this paper, we propose a general GPS-aided omnidirectional visual-inertial state estimator capable of operating in ubiquitous environments and platforms. Our system consists of two parts: 1) the pre-processing of omnidirectional cameras, IMU, and GPS measurements, and 2) the sliding window based nonlinear optimization for accurate state estimation. We test our system in different conditions including an indoor office, campus roads, and challenging open water surface. Experiment results demonstrate the high accuracy of our approach than state-of-the-art VINSs in all scenarios. The proposed odometry achieves drift ratio less than 0.5% in 1200 m length outdoors campus road in overexposure conditions and 0.65% in open water surface, without a loop closure, compared with a centimeter accuracy GPS reference.**

I. INTRODUCTION

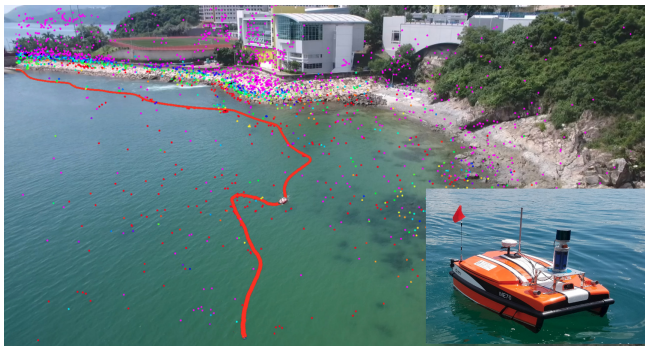
A. Motivation

Accurate state estimation is a prerequisite in many robotic applications such as unmanned aerial vehicles (UAVs), unmanned ground vehicles (UGVs), and unmanned surface vessels (USVs). Visual-inertial navigation system (VINS) estimators have led the trend in state estimation in the past decade with impressive progress by the community [1]–[3]. However, existing VINS estimators suffer from problems caused by operating environments and sensor configuration, limiting their usage in real-world robotic applications. For stabilizing perceptions, current VINS estimators are tested in restricted environments and with specific camera mechanical configuration [4]. In outdoor experiments, VINS estimators are facing challenges such as overexposure, featureless frames, and tiny pixel parallax for faraway features, resulting in the loss of stable features tracking. The configuration of cameras on UGVs or USVs, facing the front, strengthen these negative impacts. In open water environments near the coast for USVs, reliable visual measurements are only gained from nearby static objects from the shore. Current VINS approaches drift easily when the USV is making turns or when cameras are facing the sea surface.

It is reasonable to deduce that visual-inertial odometry (VIO) can achieve an outstanding performance outdoors as long as enough stable features can be observed continuously. Omnidirectional perception is then necessary to solve outdoor VINS estimation problems. Due to the inevitable drifts of large-scale outdoor trajectory estimation, precise global measurements, such as from a GPS, could help improve the accuracy of the system. Therefore, a GPS-aided omnidirectional VINS is a potential approach for state estimation tasks in challenging outdoor environments.



(a) Hybrid view of the ground experiment by our UGV platform



(b) Hybrid view of the open water experiment by our USV platform

Fig. 1: The hybrid views of outdoor experiments. Red curves are estimated trajectories, and colored spots indicate 3D features by height.

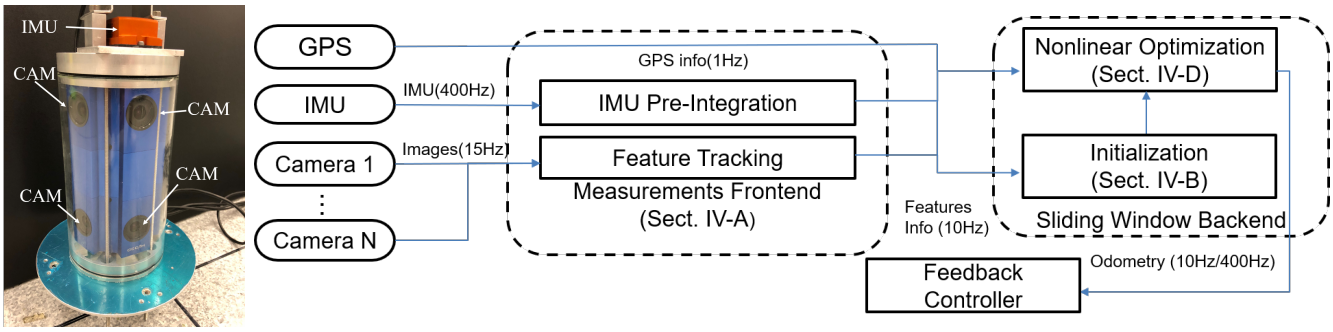
B. Contributions

In this work, we propose a novel GPS-aided omnidirectional visual-inertial state estimator, which can perform accurate motion estimation in indoors and challenging outdoor environments. We extend VINS-MONO [2] to support flexible numbers of pinhole cameras, reconstruct the feature extraction and measurement model for omnidirectional visual perception based on a unit sphere. The tightly-coupled nonlinear optimization module fuses the visual and inertial measurements, with loosely-coupled GPS refinement to decrease

This work was supported by the National Natural Science Foundation of China (Grant No. U1713211, 61573260), and the Basic Research Project of Shanghai Science and Technology Commission (Grant No. 16JC1401200, 18DZ1200804).

[1]Yang YU, Wenliang Gao, Shaojie Shen, Ming Liu are with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong, China. (email:{yang.yu,wenliang.gao}@connect.ust.hk,{eeshaojie, eelium}@ust.hk)

[2]Chengju Liu is with College of Electrical and Information Engineering, Tongji University, China. (e-mail: liuchengju@tongji.edu.cn)



(a) Sensors Structure

(b) System Pipeline

Fig. 2: (a) Hardware structure of the omnidirectional cameras with an IMU. 10 cameras are distributed as a cylinder wrapped with a plastic shell. (b) The pipeline of the proposed GPS-aided omnidirectional visual-inertial state estimator.

the drifts caused by the accumulated error in large-scale outdoors. Note that our proposed estimator also supports the GPS-blocked version which achieves a novel performance than state-of-the-art VINS algorithms.

We demonstrate the performance of our integrated system in a variety of indoor and outdoor conditions. The trajectory can be estimated close to the centimeter accuracy GPS reference with a drift ratio less than 0.7% in challenging large-scale outdoor experiments. Accordingly, we identify our contributions as follows:

- A precise and robust omnidirectional visual inertial system with a flexible number of pinhole camera configuration, online camera-IMU calibration, and fast initialization from non-stationary states.
- A tightly-coupled nonlinear optimization for visual and inertial measurements with loosely-coupled GPS refinement for accurate odometry generation.
- A variety of cross-platform experiments in the indoor, challenging outdoor ground and open water environments with novel performance.

C. Outline

In Sect IV-A, the processing of sensor measurement is presented. System initialization, in Sect IV-B, generates the roughly-estimated IMU bias, local pose state, velocity, and gravity. Then these parameters will be optimized with the feature inverse depth, calibration parameters, and frame rotation by the nonlinear optimization (Sect IV-D). The residual functions for all measures are also defined in this section. The outlier rejection is introduced in Sect IV-C. We demonstrate experiments results of indoor and outdoor environments in Sect V. The conclusions are discussed at the end of the paper.

II. RELATED WORK

Research on vision-based state estimation is extensive. We focus on applications of omnidirectional cameras and related research about visual-inertial state estimation.

There are several kinds of omnidirectional cameras. A dioptric camera uses a shaped lens, like a fish-eye lens [5], reaching a field of view (FOV) larger than 180 degrees. A catadioptric camera combines a standard camera with a shaped mirror, such as a parabolic, hyperbolic, or elliptical mirror, and provides a 360 degrees horizontal FOV and a more than 100 degrees vertical FOV [6]. Polydioptric cameras use multiple cameras, usually pinhole cameras, with overlapping FOVs [7]. Steffen et al. [8] also used polydioptric model with fish-eye cameras for omnidirectional system. However, they did not integrate the IMU measurements. In this paper, we use a polydioptric omnidirectional model with pinhole cameras because of the simple calibration and the flexibility of configuration. In addition, the overexposure and tracking loss of one camera will not affect the performance of other cameras.

A straightforward method to fuse images and IMU data is to treat the IMU as an independent module to refine the vision-only estimated odometry, which is known as loosely-coupled sensor fusion. The tightly-coupled vision-inertial method, which unites the IMU and visual measurements to be optimized jointly, is believed to have a better performance than the former method due to less information loss [9] [10]. There are two kinds of tightly-coupled state estimation: extended Kalman filter (EKF) based [3] [11], and graph-optimization-based [1] [2]. The EFK-based approach has lower computational cost but suffers from estimation inconsistency caused by EKF linearization [12]. Bundle adjustment is utilized for graph-optimization-based estimation to maintain a bounded-size sliding window of recent states for solving a nonlinear least-squares problem. Although its higher computational resource cost for iterations compared with the EKF-based approach, the optimization-based method achieves superior results [13] and real-time performance due to the improvement in hardware computational power. In this paper, therefore, we adopt the optimization-based state estimation approach.

III. OVERVIEW

Our proposed GPS-aided omnidirectional visual-inertial odometry system is shown in Fig. 2(b). Five pairs of stereo cameras, which are vertically distributed as a cylinder with little overlap, track features in all directions. Each camera works separately so the system can generate odometry as long as there is at least one camera working properly. We implement the omnidirectional VINS in monocular version, as Omni-Mono with the top 5 circular cameras, and stereo version, as Omni-Stereo with all 10 cameras (Fig 2(a)).

Stereo cameras are employed since the rigid baseline between two cameras provides scale information and more stable feature matching, which improves system performance in challenging outdoors. Furthermore, the system initialization process can be reduced to quite a short time by the rigid stereo constraints, especially on the seawater surface.

We assume that the intrinsic parameters of the cameras are known. Initial calibration parameters between cameras and the IMU are calculated by Kalibr toolbox [14] and our optimization module supports the online self-calibration for revision. We also assume the GPS signal is available at the beginning of the initialization for the GPS-aided versions.

TABLE I: Nomenclature.

Notation	Explanation
$\mathbf{F}$	Frame, where $\mathbf{F}_B$, $\mathbf{F}_W$, and $\mathbf{F}_G$ represent the IMU body frame, the world frame and the global East-North-Up (ENU) frame, respectively;
$\mathbf{R}$	Rotation matrix in $SO(3)$, where $\mathbf{R}_b^a$ represents the rotation from $\mathbf{F}_b$ to $\mathbf{F}_a$;
$\mathbf{q}$	Quaternion under Hamilton notation, with $\mathbf{q}_b^a$ corresponding to $\mathbf{R}_b^a$;
$\mathbf{p}$, $\mathbf{v}$	Translation and velocity vector in $\mathbb{R}^3$, where $\mathbf{p}_b^a$ and $\mathbf{v}_b^a$ represents the translation and velocity, respectively, from $\mathbf{F}_b$ to $\mathbf{F}_a$ defined in $\mathbf{F}_a$;
$\mathbf{m}$	Translation vector in $\mathbb{R}^3$ for GPS measurement, where $\mathbf{m}_s^A$ represents the translation in $\mathbf{F}_A$ with state s;
$\mathbf{b}$	IMU bias in $\mathbb{R}^3$, where $\mathbf{b}_{a_t}$, $\mathbf{b}_{g_t}$ represents the IMU acceleration bias and gyroscope bias at time t;
$\tilde{\mathbf{a}}, \tilde{\boldsymbol{\omega}}$	Raw measurements of acceleration and angular velocity, respectively, in $\mathbb{R}^3$ from the IMU;
$\mathbf{g}^W$	Gravity vector in the world frame;
$\mathcal{X}$	States to be estimated, including IMU states $\mathbf{x}_k$, calibration parameters $\mathbf{t}_{c^i}^B$ between the ith camera and the IMU, the inverse depth of features, and the rotation from $\mathbf{F}_W$ to $\mathbf{F}_G$;
$\mathbf{t}_{c^i}^B$	Calibration parameters between the i^{th} camera and IMU;
$\mathbf{x}_k$	State vector of IMU when the k^{th} image is captured;
$\mathbf{r}$	Residual, where $\mathbf{r}_\mathcal{B}$, $\mathbf{r}_\mathcal{C}$, and $\mathbf{r}_\mathcal{G}$ represent the IMU residual, the camera residual and the GPS residual.
$\tilde{\mathbf{z}}$	The raw sensor measurements;
P_l	The l^{th} feature observed from the camera;
c_k^i	Camera i, where k indicates the k^{th} frame

The nomenclature is defined in Table I. We indicate $\tilde{(\cdot)}$ as the noise measurement or estimate of the accurate quantity, and $\otimes$ as multiplication between two quaternions.

IV. Methodology

A. Measurements Preprocessing

There are three kinds of measurements in our system: the sparse features from images, the IMU and the GPS measurements. The images and IMU measurements are preprocessed then incorporated into the estimation. The sliding window runs on the image frequency, and the low-frequency GPS measurements are bonded with the same time stamp image.

1) Visual Feature Extraction For Multi-cameras: We evenly divide each new incoming image into 3×3 subsections and equally extract features from each part. New corner features are detected by Harris Corner Detect [15] to maintain a maximum number of 80 features in each camera for real-time performance, and the existing features are tracked by the KLT sparse optical flow algorithm [16]. Features

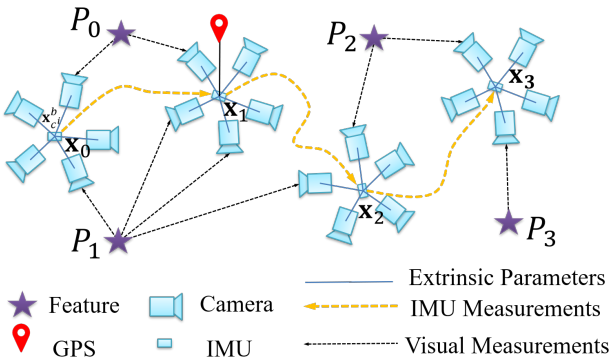


Fig. 3: Illustration of the sliding window for GPS-aided omnidirectional VIO. The GPS signal is available for at most one frame in the window because of the low frequency. For Omni-Stereo, the camera indicates the stereo cameras pair.

are maintained with their specific indexes and are sent to the optimization module as an omnidirectional point cloud. Note that we only implement the feature matching between the vertical stereo cameras pair, also by KLT algorithm, due to the few overlapping between the horizontal cameras. Features observed from a vertical stereo pair are independent with the features from the other stereo pairs.

2) IMU Pre-Integration: Multiple IMU measurements will be pre-integrated during $[k, k+1]$ to avoid re-propagation. We utilize the IMU pre-integration model presented in [17], which involves the IMU bias correction. The measurements of acceleration and angular velocity from the IMU, $\tilde{\mathbf{a}}$ and $\tilde{\boldsymbol{\omega}}$, in the body frame are defined as

$$\begin{aligned} \tilde{\mathbf{a}}_t &= \mathbf{a}_t + \mathbf{b}_{a_t} + \mathbf{R}_W^t \mathbf{g}^W + \mathbf{n}_a \\ \tilde{\boldsymbol{\omega}}_t &= \boldsymbol{\omega}_t + \mathbf{b}_{g_t} + \mathbf{n}_g, \end{aligned} \tag{1}$$

combined with Gaussian acceleration noise $\mathbf{n}_a$ and gyroscope noise $\mathbf{n}_g$. To avoid the large computation of re-propagation, we translate the world frame to the body frame.

B. System Initialization

1) Omnidirectional Initialization: The accuracy of the initialization process dramatically affects the results of the VINS estimator, especially in real-world applications . The performance of monocular VIO initialization is limited by scale recovery. Sufficient motion and rotation are required to accumulate enough keyframes to refine the state parameters, which takes several seconds for initialization [2]. Regarding the stereo system, with the known calibration parameters between two cameras, the initialization needs only solve the IMU bias, velocity, and gravity.

We first introduce the initialization for GPS-blocked versions of proposed approaches. For the Omni-Mono, we follow the same initialization process as [2]. For the Omni-Stereo, we construct the stereo visual initialization of the sliding window, including pose and feature position, and then align it with the IMU integration results. Each frame taken from the stereo cameras can be treated as a keyframe because of the known calibration parameters. Through loosely-coupled alignment between the stereo visual odometry and

the IMU metric pre-integration information, the gravity, velocity, and gyroscope bias can be roughly recovered. According to our test, the initialization of the stereo system is less than 1 second without sufficient movements required. After initialization, the Z-axis of the world frame precisely aligns with the gravity vector.

2) Rotation From World frame to Global frame: The initial point is defined as the GPS measurement at the origin of the world frame. The rotation $\mathbf{R}_W^G$ is initialized by the initial point $\mathbf{m}_{init}^G$ and the next GPS point $\mathbf{m}_{next}^G$ with the assumption that there is only yaw rotation from the global frame to the world frame, and the distance between the two GPS points meets the minimum threshold. In this way, we can obtain the initial estimate of rotation by

$$\theta = \arccos \frac{(\mathbf{m}_{next}^G - \mathbf{m}_{init}^G) \cdot \mathbf{p}_B^W}{\|(\mathbf{m}_{next}^G - \mathbf{m}_{init}^G)\| \|\mathbf{p}_B^W\|}, \tag{2}$$

where θ is the yaw angle between the world frame and the global frame and $\|\cdot\|$ is the l^2 norm.

C. Outlier Rejection

Reliable visual measurements are only from nearby static objects. They can help the calibration parameters converge quickly and provide strong constraints to the state parameters. Matched features from the sea surface, the sky, and faraway islands will influence the estimation negatively and are treated as unreliable features. We define inconsistent matched and unreliable features as outliers.

Outlier rejection is firstly performed using 2D-2D RANSAC with the fundamental matrix model [18] in the measurement preprocessing stage introduced in Sect. IV-A. Then we adopt 3D-2D PnP RANSAC [19] outlier rejection based on the 3D feature position, which is estimated in the local sliding window, and on the 2D coordinates in the camera plane. We also utilize the Huber loss in the optimization module to reduce adverse effects by outliers with large residuals. Finally, we remove the features with unreasonable large depth values after each optimization iteration.

D. Nonlinear Optimization

We propose a sliding-window-based nonlinear optimization for real-time state estimation, as shown in Fig. 3. We utilize the IMU measurement model in [2] for IMU residual.

1) Formulation: The state vector is defined as

$$\begin{aligned}
&\mathcal{X} = [\mathbf{x}_n, \cdots, \mathbf{x}_{n+N}, \mathbf{t}_{c^0}^B, \mathbf{t}_{c^1}^B, \cdots, \mathbf{t}_{c^I}^B, d_m, \cdots, d_{m+M}, \mathbf{q}_W^G], \\
&\mathbf{x}_k = [\mathbf{p}_{B_k}^W, \mathbf{v}_{B_k}^W, \mathbf{q}_{B_k}^W, \mathbf{b}_a, \mathbf{b}_g], k \in [n, n+N], \\
&\mathbf{t}_{c^i}^B = [\mathbf{p}_{c^i}^B, \mathbf{q}_{c^i}^B], i \in [0, I],
\end{aligned} \tag{3}$$

where d_m is the inverse depth of the m^{th} feature from its first observation. I, N, and M is the total number of cameras, IMU states in the sliding window and features, [20]. The calibration parameters $\mathbf{t}_c^B$ will be updated in optimization when they converge to a reasonable value.

We extend the measurement residuals to handle the multiple cameras, IMU and GPS for the maximum posterior (MAP) estimation:

$$\begin{aligned}
\mathcal{X}^* = \arg\min_{\mathcal{X}} \Bigg\{ &\|\mathbf{r}_p - \mathbf{H}_p \mathcal{X}\|^2 + \sum_{k \in \mathcal{B}} \|\mathbf{r}_{\mathcal{B}}(\tilde{\mathbf{z}}_{B_{k+1}}^{B_k}, \mathcal{X})\|^2_{\Theta_{B_{k+1}}^{B_k}} \\
&+ \sum_{i \in I} \sum_{(l,j) \in \mathcal{C}^i} \|\mathbf{r}_{\mathcal{C}^i}(\tilde{\mathbf{z}}_l^{c_j^i}, \mathcal{X})\|^2_{\Theta_l^{c_j^i}} \\
&+ \|\mathbf{r}_{\mathcal{G}}(\tilde{\mathbf{z}}_{G_u}^{G_0}, \mathcal{X})\|^2_{\Theta_{G_u}^{G_0}} \Bigg\},
\end{aligned} \tag{4}$$

where $[\mathbf{r}_p, \mathbf{H}_p]$ is the prior information [2]. $\mathcal{B}, \mathcal{C}$, and $\mathcal{G}$ are the set of IMU, camera, and GPS measurements, respectively. $\mathcal{C}^i$ is the set of all observations by camera i. And j, l are image frame indexes and feature indexes, respectively. G_0, G_u are the initial and current GPS measurements. $\Theta_{B_{k+1}}^{B_k}$ and $\Theta_l^{c_j^i}$ are the covariance matrix of the IMU and visual measurements, respectively. The covariance of the GPS measurements $\Theta_{G_u}^{G_0}$ is measured from the GPS directly.

To solve this nonlinear MAP problem, Ceres Solver [21] is used. We adopt the Huber loss as the loss function to penalize outliers with large residual for better system robustness. The inverse matrices of Θ for different sensor measurements are employed as the regularization weights to balance the loss. Because of the low frequency of the GPS sensor, the GPS residual is only added to the optimization module when the GPS data is received with acceptable covariance in the current frame. Otherwise, the nonlinear optimization only considers the marginalization, IMU and visual residuals.

2) Omnidirectional Multi-Camera Measurement Model: To utilize the benefits of the omnidirectional camera, the camera measurement residual is defined on a unit sphere, as proposed in [17]. The camera residual is defined as:

$$\mathbf{r}_{\mathcal{C}^i}(\tilde{\mathbf{z}}_l^{c_j^i}, \mathcal{X}) = \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 \end{bmatrix}^T \cdot (\tilde{P}_l^{c_j^i} - \frac{P_l^{c_j^i}}{\|P_l^{c_j^i}\|}), \tag{5}$$

$$\tilde{P}_l^{c_j^i} = \begin{bmatrix} \tilde{x}_l^{c_j^i} & \tilde{y}_l^{c_j^i} & \tilde{z}_l^{c_j^i} \end{bmatrix}^T \tag{6}$$

is the observation of the l^{th} feature in the j^{th} frame from the i^{th} camera. $\mathbf{h}_1$ and $\mathbf{h}_2$ are two arbitrarily selected orthogonal bases which span the tangent plane of $\tilde{P}_l^{c_j^i}$.

Note that we also store the first captured camera index, m^{th}, and frame indexes, v^{th}, of the l^{th} feature. Thus, the formulation of camera measurements of the l^{th} feature is

$$\begin{aligned}
P_l^{c_j^i} &= -\mathbf{R}_{c^i}^{B^{-1}} \mathbf{p}_{c^i}^B \\
&\quad + \mathbf{R}_{c^i}^{B^{-1}} \left(-\mathbf{R}_{B_j}^{W^{-1}} \mathbf{p}_{B_j}^W + \mathbf{R}_{B_j}^{W^{-1}} P_l^W \right), \\
P_l^W &= \mathbf{p}_{B_v}^W + \mathbf{R}_{B_v}^W (\mathbf{p}_{c^m}^B + \mathbf{R}_{c^m}^B \frac{1}{\lambda_l} \cdot P_l^{c_v^m}),
\end{aligned} \tag{7}$$

where $P_l^{c_v^m}$ is the noiseless first observation of the l^{th} feature that happens in the v^{th} frame, from the m^{th} camera. Especially, as for the feature of the monocular camera, the first captured camera is the current captured camera, $m = n$.

3) GPS Measurement Model: Different from the IMU residual, the GPS residual is defined based on the initial and current GPS measurements, not on the two continuous sliding windows, because of the low GPS frequency compared with the sliding window. For any current GPS point, we have

$$\mathbf{m}^G_{cur} - \mathbf{m}^G_{init} = \mathbf{R}^G_B \mathbf{m}^B_{cur} - \mathbf{R}^G_W \mathbf{m}^W_{init} + \mathbf{R}^G_W \mathbf{p}^W_B, \quad (8)$$

where $\mathbf{m}^B_{cur}$ and $\mathbf{m}^W_{init}$ are the same because of the fixed translation from the GPS to the IMU regardless of the coordinates. $\mathbf{m}^G_{cur}$ and $\mathbf{m}^G_{init}$ are GPS positions in the global frame. In this way, residual can be defined as

$$\begin{aligned}\mathbf{r}_{\mathcal{G}}(\tilde{\mathbf{z}}^{G_0}_{G_u}, \mathcal{X}) = \mathbf{m}^G_{cur} - \mathbf{m}^G_{init} \\ - \mathbf{R}^G_W(\mathbf{R}^W_B \mathbf{m}^B_{cur} - \mathbf{m}^W_{init} + \mathbf{p}^W_B).\end{aligned} \quad (9)$$

V. EXPERIMENTS

A. Implementation Details

An OCCAM Omni-Stereo camera, an Xsens MTi-10 IMU and a low-cost G-STAR IV GPS comprise our system (Fig 2(a)). A real-time kinematic (RTK) GPS, COMNAV T300 GNSS, with centimeter accuracy is adopted as the outdoor evaluation reference. The multi-camera system has 10 cameras, and each camera captures monochrome 752 $\times$ 480 images at 15 Hz. Cameras are hardware synchronized by the manufacturer. The IMU provides the acceleration and angular velocity at 400 Hz and the GPS signal frequency is at 1 Hz. Our UGV platform is a golf car and USV platform is the OceanAlpha as shown in Fig. 1.

TABLE II: Indoor Experiment Results

Algorithm	Drift(m)	Drift-Ratio(%)
MSCKF-VIO Stereo	0.426	2.43
OKVIS Stereo	0.581	3.31
VINS-FUSION Stereo	0.302	1.72
Our Stereo	0.15	0.85
Our Omni-Mono	0.06	0.33
Our Omni-Stereo	**0.05**	**0.30**

TABLE III: Ground Experiment Results

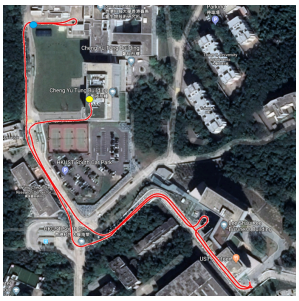
Trajectory	Total Length (m)	GPS Aided	Omni-Mono		Omni-Stereo	
			Drift (m)	Ratio (%)	Drift (m)	Ratio (%)
Campus	1207.9	NO	31.32	2.59	11.64	0.96
		YES	17.96	1.48	**5.50**	**0.46**

B. Indoor Experiments

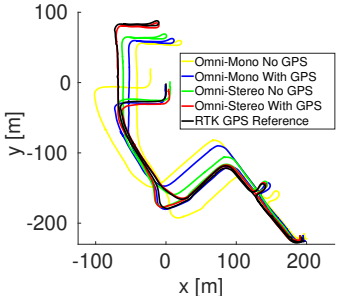
With a motion tracking system as the ground truth, the indoor experiments are executed in our lab office. We intend to prove the accuracy of our omnidirectional VIO system in GPS-blocked conditions by moving the sensors with hands. We compare our proposed Omni-Mono and Omni-Stereo with MSCKF-VIO [22], OKVIS [9] and VINS-FUSION [23]. We also run our Omni-Stereo with only one stereo pair as a reference. All algorithms are tested without loop closure. We execute the algorithms 10 times with the same data, and the drift results are shown in Table II with total length of 17.52 m. The drift is calculated based on the absolute trajectory error (ATE). Our one pair stereo VIO has a lower drift ratio than other stereo frameworks. The Omni-Mono and Omni-Stereo both greatly improve accuracy.

TABLE IV: Open Water Experiment Results

Trajectory	Total Length (m)	GPS Aided	Omni-Mono		Omni-Stereo	
			Drift (m)	Ratio (%)	Drift (m)	Ratio (%)
U-Turn	284.30	NO	4.55	1.60	2.41	0.84
		YES	3.25	1.14	**1.85**	**0.65**



(a) Estimated Trajectory aligned with satellite



(b) Trajectory comparison.

Fig. 4: (a) The estimated trajectory aligned in the satellite map. The yellow and blue spots are the start and end points. (b) Comparison of the estimated trajectories.

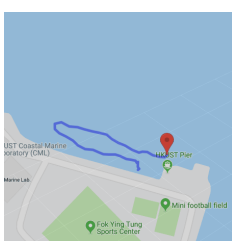
C. Outdoor Ground Experiments

The outdoor experiments are first executed on the HKUST campus road. All the stereo frameworks mentioned in indoor experiments drifted hugely for this task because of the overexposure, long distance motion, and feature detection failure. We compare the performance of the proposed Omni-Mono and Omni-Stereo with GPS enabled and disabled.

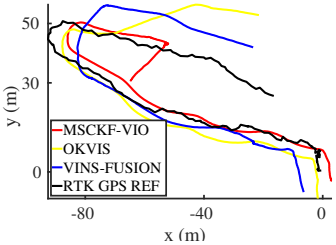
The results are shown in Fig. 4. We align the estimated trajectory [24] in the satellite map in Fig. 4(a) . Although the drifting rates are higher than those in the indoor experiments, the estimated odometry accurately aligns with the RTK-GPS reference in large-scale experiments by extracting enough features in all directions without a loop closure module. The drift results are also calculated based on the ATE and are shown in Table III. Omni-Stereo has better performance in such large-scale path estimation than Omni-Mono. The GPS-aided approach helps refine the drifts for both approaches.

D. Outdoor Open Water Experiments

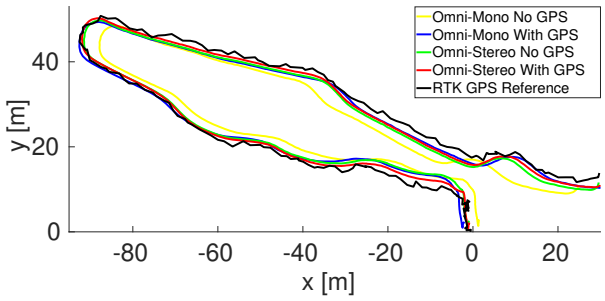
We test our algorithms on a USV platform in an open water area. We operate the USV following the coast with a U-turn which is shown by the GPS in Google Maps in Fig. 5(a). The velocity of the USV is around 1.5 meters per second. Note that the USV works on the sea surface without a stationary initial state. All stereo frameworks mentioned in Table II drifted a lot when the USV facing to the sea surface showing in Fig. 5(b). From the results in Table IV, our proposed omnidirectional VIO could recover the trajectories precisely for the complex U-turn path. Omni-Stereo has a better performance than Omni-Mono because of the additional cameras and stereo constraints.



(a) Map Reference



(b) Stereo Frameworks Drifted Results



(c) Trajectory Comparision

Fig. 5: (a) GPS trajectory shown in Google Maps. The red pin is the end point. (b) Stereo frameworks drift with the U-Turn path. (c) Comparison of our estimated trajectories.

E. Discussion

For the indoor experiments with blocked GPS measurements, our proposed Omni-Mono and Omni-Stereo improve the accuracy of state estimation with the drift ratio around 0.3% without loop closure, which outperform state-of-the-art. For the outdoor experiments without tremendous stationary features and stable lighting conditions, our proposed system still achieves remarkable results while other stereo systems fail. The Omni-Stereo achieves a drift ratio less than 1% without GPS and 0.5% with the benefit of GPS in the large-scale overexposure outdoor environment (Table III). In the open water experiments, limited reliable visual features are only from the coast, so the number of reliable features is much lower compared with the indoor and outdoor ground conditions, leading higher drift ration results around 0.8% shown in Table IV. The long-distance features, strong exposure under the sun and the low camera resolution may also lead to these results. The performance has the potential to be better with high-resolution cameras.

VI. CONCLUSIONS AND FUTURE WORKS

In this paper, a novel GPS-aided omnidirectional VIO state estimator, suitable for a variety of severe operating environments, has been presented. We extended state-of-the-art monocular VIO to adopt a flexible number of cameras to compose omnidirectional monocular and omnidirectional stereo systems. We also added the GPS factor into the nonlinear-optimization-based, multi-sensor fusion to reduce the drift in outdoor large-scale trajectories. The accuracy and robustness of the proposed system have been demonstrated with indoor and outdoor experiments. We open source our implementation[1]. Further research is still necessary for better system performance, and we are interested in loop closure and mapping integration to our next proposed system.

REFERENCES

[1] S. Shen *et al.*, "Tightly-coupled monocular visual-inertial fusion for autonomous flight of rotorcraft mavs," in *2015 IEEE International Conference on Robotics and Automation*, 2015, pp. 5303–5310.
[2] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.
[3] M. Li and A. I. Mourikis, "High-precision, consistent ekf-based visual-inertial odometry," *The International Journal of Robotics Research*, vol. 32, no. 6, pp. 690–711, 2013.
[4] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2014.
[5] M. Ramezani *et al.*, "Pose estimation by omnidirectional visual-inertial odometry," *Robotics and Autonomous Systems*, pp. 26–37, 2018.
[6] D. Scaramuzza and R. Siegwart, "Appearance-guided monocular omnidirectional visual odometry for outdoor ground vehicles," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1015–1026, Oct 2008.
[7] J. Schneider *et al.*, "Fast and effective online pose estimation and mapping for uavs," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 4784–4791.
[8] S. Urban and S. Hinz, "Multicol-slam-a modular real-time multi-camera slam system," *arXiv preprint arXiv:1610.07336*, 2016.
[9] S. Leutenegger *et al.*, "Keyframe-based visualinertial odometry using nonlinear optimization," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, 2015.
[10] G. Huang *et al.*, "Optimal-state-constraint ekf for visual-inertial navigation," in *Robotics Research*. Springer, 2018, pp. 125–139.
[11] M. Bloesch *et al.*, "Robust visual inertial odometry using a direct ekf-based approach," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2015, pp. 298–304.
[12] J. A. Hesch *et al.*, "Consistency analysis and improvement of vision-aided inertial navigation," *IEEE Transactions on Robotics*, 2014.
[13] K. 1Eckenhoff *et al.*, "High-accuracy preintegration for visual-inertial navigation," in *Algorithmic Foundations of Robotics*, 2016.
[14] P. Furgale and R. Siegwart, "Unified temporal and spatial calibration for multi-sensor systems," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov 2013, pp. 1280–1286.
[15] C. G. Harris *et al.*, "A combined corner and edge detector." in *Alvey vision conference*, vol. 15, no. 50. Citeseer, 1988, pp. 10–5244.
[16] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. of the Intl. Joint Conf. on Artificial Intelligence*, Vancouver, Canada, Aug. 1981, pp. 24–28.
[17] Y. Lin *et al.*, "Autonomous aerial navigation using monocular visual-inertial fusion," *Journal of Field Robotics*, vol. 35, pp. 23–51, 2018.
[18] Hartley *et al.*, *Multiple View Geometry in Computer Vision*, 2nd ed. New York, NY, USA: Cambridge University Press, 2003.
[19] V. Lepetit *et al.*, "Epnp: An accurate o(n) solution to the pnp problem," *International Journal of Computer Vision*, vol. 81, 02 2009.
[20] Z. Yang *et al.*, "Self-calibrating multi-camera visual-inertial fusion for autonomous mavs," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2016, pp. 4984–4991.
[21] S. Agarwal and K. Mierle, "Ceres solver," http://ceres-solver.org.
[22] K. Sun *et al.*, "Robust stereo visual inertial odometry for fast autonomous flight," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 965–972, 2018.
[23] T. Qin, S. Cao, J. Pan, and S. Shen, "A General Optimization-based Framework for Global Pose Estimation with Multiple Sensors," *arXiv e-prints*, p. arXiv:1901.03642, Jan. 2019.
[24] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 376–380, April 1991.

[1]https://github.com/gaowenliang/vins_so