2024 International Conference on Unmanned Aircraft Systems (ICUAS)
June 4-7, 2024. Chania, Crete, Greece

LiDAR Stereo Visual Inertial Pose Estimation Based on Feedforward and Feedbacks

Wenyu Yang[1], Haochen Hu[1], Kwai-wa Tse[1], Shengyang Chen[2], Weisong Wen[1], and Chih-yung Wen[1]

***Abstract*— In this paper, we present a LiDAR Visual Inertial Odometry (LVIO) based on feedforward and feedbacks. Compared to traditional Kalman filter-based methods or optimization-based methods for sensor fusion, the proposed system achieves sensor fusion through feedforward and feedbacks. This system, named Feedforward-feedback LiDAR Visual Inertial System (FLiVIS) consists of a Visual Inertial Odometry (VIO) subsystem and a LiDAR Inertial Odometry (LIO) subsystem, these two subsystems are coupled through complementary filters. Instead of directly integrating gyroscope data and accelerometer data, our framework leverages the complementary nature of gyroscope and accelerometer measurements. FLiVIS is evaluated on public datasets, it achieves a relative translation error of 0.68% on the KITTI dataset and 0.138 m absolute translation error on the NTU-Viral dataset, respectively. The experiment results demonstrate the accuracy and robustness of FLiVIS with respect to other state-of-the-art counterparts. FLiVIS is capable of accommodating both multi-line spinning LiDARs and emerging solid-state LiDARs, which employ distinct scanning patterns. Additionally, it can perform real-time operations on a range of platforms, from laptops to upboard processors.**

I. Introduction

Simultaneous Localization and Mapping (SLAM) [1] is crucial for robot operation in unknown environments, such as Unmanned Aerial Vehicle (UAV)s search and rescue, and building inspection tasks. Common SLAM systems utilize Light Detection and Ranging (LiDAR), and visual sensors for exteroceptive perception and inertial sensors for interoceptive state estimation. Previous research frameworks primarily relied on a single external perception sensor (like a camera or LiDAR) combined with an Inertial Measurement Unit (IMU). For UAVs, cost-effective configurations often include VIO, Stereo Visual Inertial Odometry (SVIO) , and LIO.

As the demand for operating intelligent robots in the real world increases, these robots often need to work in challenging environments. Above mentioned systems that rely on a single exteroceptive sensor often cannot provide the accuracy and robustness needed for pose estimation in these challenging conditions. For example, VIO is prone to failure in outdoor environments where the lighting conditions may change rapidly, and the performance of LIO is affected by the environment geometry. To tackle this, several recent works proposed LVIO that fuses the camera and LiDAR measurement. Existing systems usually focus on the fusion camera measurements and LiDAR measurements. Similar to the fusion of a single exteroceptive sensor with an IMU, Kalman filter or factor graph optimization approaches are usually applied to fuse camera and LiDAR measurement.

Contemporary filter-based VIO or LIO systems predominantly employ either a filter-based or an optimization-based framework for the fusion of different data. Both frameworks hinge on the identification of the most suitable prediction that fulfills each individual measurement, such as those derived from visual cameras or an Inertial Measurement Unit. As the input data escalates, there is a corresponding increase in the demand for computational power. This poses a significant challenge for robotic systems, especially those with constraints on size, weight, and power, such as Unmanned Aerial Vehicles. High-performance pose estimators, in these instances, can consume an excessive amount of computational resources.

Building on our previous work [2], we propose a feedforward-feedback based framework. Unlike other frameworks where gyroscope data is used to predict rotation and accelerometer measurements are used to calculate pose through integration, our framework leverages the complementary nature of gyroscope and accelerometer measurements. Gyroscope data is accurate in the short term but drifts over time (low-frequency noise), while accelerometer data is noisy in the short term but accurate over time (high-frequency noise). We use measurements from both the gyroscope and accelerometer to jointly predict the robot's rotation. We introduce a feedforward-feedback-based complementary filter framework, integrating IMU data and visual LiDAR into the pose estimation pipeline. This structure combines the advantages of the Kalman filter method and includes several decoupled models to partition the process of LiDAR stereo vision inertial fusion. The complementary filter is used to process IMU data, and the feedforward and feedbacks are used to combine measurements from different sensors. With a better initial value, the update step of sensor fusion can converge more quickly and achieve more accurate pose estimation. This approach enhances the precision and robustness of pose estimation, which is crucial for the operation of robots in complex environments.

In summary, the aim of this work is to tackle the issue of fast and robust LiDAR-visual-inertial pose estimation, especially for systems that are restricted by resource limitations.

*This work was supported by the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University

[1]W. Yang, H. Hu, K-W. Tse, W. Wen, and C-Y. Wen are with the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China. (E-mail: allen.yang@connect.polyu.edu.hk, haru-haochen.hu@connect.polyu.hk, kwai-wa.tse@connect.polyu.hk, welson.wen@polyu.edu.hk, cywen@polyu.edu.hk)

[2]S. Chen is with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China. (E-mail: 17903070r@connect.polyu.hk)

2024 International Conference on Unmanned Aircraft Systems (ICUAS) | 979-8-3503-5788-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICUAS60882.2024.10556962

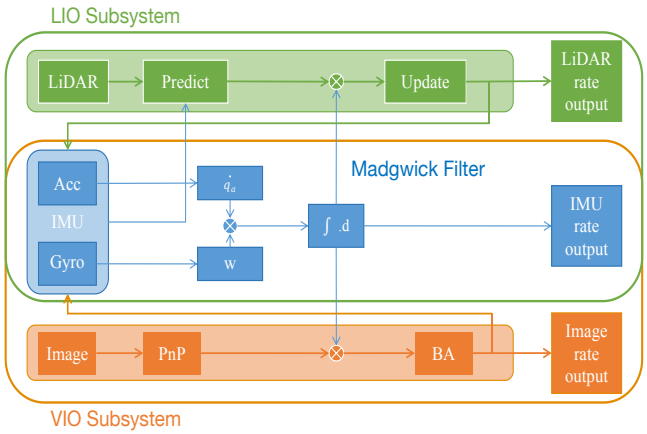


Fig. 1. System architecture of the proposed pipeline. The inputs are the measurements from the LiDAR, IMU, and camera. The IMU data is processed by the Madgwick filter, and the fusion of IMU estimation and the other two sensors is achieved through feedforward and feedbacks

Fig. 1 shows the system architecture of FLiVIS, the whole system consists of a VIO subsystem and a LIO subsystem. The VIO subsystem is based on FLVIS [2] and the LIO subsystem is based on Fast-lio2 [3]. The proposed system is evaluated on the KITTI dataset [4] and NTU VIRAL dataset [5] to test its applicability in autonomous driving and UAV inspection tasks. The results demonstrate that FLiVIS is comparable to state-of-the-art systems, and it can achieve real-time processing on various platforms, from laptops to upboard processors.

II. Related Work

A. Visual(-Inertial) Odometry

V-SLAM has been widely applied in robotics and AV/VR applications. With regard to different sensor properties, it has evolved significantly with variations including Mono(monocular), stereo, and RGBD V-SLAM. Notable systems such as PTAM [6], and ORB-SLAM [7] have been developed in the past decade. Based on the extraction of feature points, V-SLAM methodologies can be categorized into two distinct categories: feature-based and direct methods. The feature-based approach [8] extracts salient feature points, a process that can be computationally intensive and time-consuming. Conversely, the direct method [9] circumvents the feature extraction step, focusing solely on the computation of pixel gradients, thereby offering a more expedient solution. However, this efficiency comes at the cost of increased susceptibility to failure.

A common enhancement across these systems is the fusion with IMU [10], which has been observed to substantially improve the robustness of the SLAM system. This improvement has been particularly beneficial for UAV navigation due to its lightweight and robust nature, thereby broadening the scope of its applications. Significant VIO systems are the Vins series [11], Openvins [12], MSCKF [13], and FLVIS [2].

B. LiDAR(-Inertial) Odometry

Previously, due to factors such as large size and high cost, LiDAR was primarily used in vehicle autonomous driving, with less application in UAVs. However, in recent years, with the advancement of LiDAR hardware technology , its size has been continuously reduced and the price has also decreased. Consequently, there are now examples of LiDAR being used in UAVs [14]. This demonstrates the expanding applicability of LiDAR technology in the field of UAVs.

The majority of existing LiDAR odometry methodologies encompass two primary components: scan registration and mapping [15]. Scan registration is facilitated by the Iterative Closest Point (ICP) or Normal Distribution Transformation (NDT) [16] process. The ICP process iteratively searches the K Nearest Neighbors (KNN) [17] to accomplish the registration of point cloud scans between consecutive frames, while the NDT method directly matches the probability density function of the current scan to the last scan or the map. Different from the scan registration step, mapping operates at a lower frequency [18], registering newly added point clouds into the local map. This two-step structure forms the backbone of contemporary LiDAR odometry systems, enabling efficient spatial recognition and navigation.

The integration of IMU can significantly enhance the accuracy and robustness of LiDAR odometry [3]. This is achieved by compensating for motion distortion in a LiDAR scan and providing an optimal initial pose required for the ICP process. Recent advancements in LiDAR-inertial fusion have led to the development of more tightly coupled systems that perform odometry within a compact local map,composed of a fixed number of recent LiDAR scans or keyframes. This scan-to-local-map registration typically yields higher accuracy than scan-to-scan registration by utilizing more recent information. For instance, LIO-SAM [19] requires a nine-axis IMU to generate attitude measurements as a prior for scan registration within a small local map. Similarly, LINS [20] incorporates a tightly coupled iterated Kalman filter and a robocentric formula into the LiDAR pose optimization process. However, due to the small size of the local map used to achieve real-time performance, the odometry tends to drift rapidly. This necessitates a low-rate mapping process, such as map refining (as in LINS), sliding window joint optimization (as in [21]), and factor graph smoothing (as in LIO-SAM [19]).

C. LiDAR-Visual-Inertial Odometry

The utilization of multiple sensors in LiDAR-Visual-Inertial SLAM systems enhances their capability to navigate various challenging environments, even in instances where one sensor fails or is partially degraded [22]. This advantage has motivated the development of several LiDAR-Visual-Inertial SLAM systems within the research community, demonstrating the adaptability and resilience of these systems in diverse environmental conditions. For instance, R2LIVE [22] employs a different approach, extracting features from both LiDAR and images and then conducting re-projection error within the framework of an Iterated Error-

State Kalman Filter. Notably, R2LIVE provides a tightly-coupled system for fusing LiDAR-Visual-Inertial sensors. In LVI-SAM [23], the accuracy of the VIO is enhanced by extracting depth information of visual features using LiDAR measurement data. Correspondingly, the LIO utilizes the estimates from VIS as initial values to aid in scan matching. Loop closure is initially identified by VIO and further improved by LIO. These diverse systems underscore the dynamic evolution and application of LiDAR-Visual Inertial SLAM methodologies in the field.

Similar to VIO and LIO, most LVIO fuses data from different sensors by Kalman filter-based [13] approach or optimization-based [24] approach. The Markov assumption in the Kalman filter-based method makes it prone to failure in noisy data, while the optimization-based methods are usually computationally costly [25].

In this work, we exploit the complementarity of IMU measurement [26] and fuse the information from different sensors by feedforward and feedbacks. A better initial value is fed into visual or LiDAR pose estimation in the pipeline, thus the system is lightweight and robust. The main contributions of this work are listed as follows:

- A novel pose estimation framework that fuses the LiDAR visual and inertial measurement through designed feedforward and feedbacks.
- Performance evaluation on both the KITTI autonomous driving dataset and NTU-Viral UAV dataset. The benchmark evaluation results are comparable to the top-ranked existing algorithms.

III. System Overview and Notation

The proposed LiDAR stereo visual-inertial system, as shown in Fig. 1, consists of an LIO subsystem and a VIO subsystem. The coordinate system on the upper right part is attached to the IMU center and designated as the IMU frame (i). The LiDAR center is denoted as ($LiDAR$), and the corresponding optical center coordinates of the two cameras are camera0 frame (c_0) and camera1 frame (c_1), respectively. The world frame is defined as a local East, North, and Up (ENU) Cartesian coordinate system, which means that the direction of gravity is opposite the z-axis. A transformation matrix $\boldsymbol{T} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0} & 1 \end{bmatrix} \in \mathbb{R}^{4\times4}$ consist of a rotation matrix $\boldsymbol{R} \in \mathbb{R}^{3\times3}$ and a translation vector $\boldsymbol{t} \in \mathbb{R}^3$. The transformation from the IMU frame to the world frames is represented by the manifold on the special Euclidean group $SE(3)$. For example, T_{wi} refers to the transformation from the IMU frame to the world frame, and R_wi refers to the rotation matrix from the IMU frame to the world frame. It can also be parameterized using the quaternion $q_{wi} = (q_w, q_x, q_y, q_z)^\top$. The $p_{wi} = (p_x, p_y, p_z)^\top = t_{wi}$ refers to the translation from the IMU frame to the world frame.

$$T_{wi} = \begin{pmatrix} R_{wi} & t_{wi} \\ 0_{1\times3} & 1 \end{pmatrix} = (q_{wi}^{\oplus^\top} q_{wi}^{+}, p_{wi}). \tag{1}$$

We refer readers to [27] for the conversion between the rotation matrix and the quaternion. The egomotion can be described using: rotation q_{wi}, position p_{wi} and velocity v_{wi}. Notably, these states are defined in the world frame. For simplicity, subscripts and superscripts are neglected, and q, p, and v are used hereinafter. The state of the system also contains the IMU-related states in the IMU frame: the accelerometer measurement a and the gyroscope measurement $w = (0, w_x, w_y, w_z)^\top$, the accelerometer bias b_a, and the gyroscope bias b_g. the quaternion derivative describing the rate of change of the earth frame relative to the sensor frame $\dot{q}$ can be calculated [28] as (2):

$$\dot{q} = \frac{1}{2} q \otimes w. \tag{2}$$

In summary, the full state of the system can be represented by $x = (q, p, v, a, w, b_a, b_g)^\top$. The state will be updated when IMU, visual, or LiDAR measurement is available. We refer readers to [2] for the IMU propagation model.

IV. Methodology

In this section, we introduce the sensor fusion method used in this paper. In Section IV-A, we first compare the similarities and differences between the complementary filter, Kalman filter, and optimization-based methods. We believe the main difference lies in the perspective from which sensor fusion is viewed: the complementary filter approaches it from the frequency domain and signal perspective, while the Kalman filter and optimization-based methods model the sensor fusion problem from the perspective of maximum a posterior probability. In Section IV-B, we use the one-dimensional complementary filter as an example to derive the different forms of signal fusion from the frequency domain to the time domain. In Section IV-C, we demonstrate the Madgwick filter in three-dimensional space based on IMU data and clarify its role in FLiVIS. Finally, we explain our visual and LiDAR measurement model in section IV-D.

A. Complementary Filter

In conventionally Kalman filter-based or optimization-based methods, only gyroscope measurement is used for orientation estimation. The accelerometer data is ignored in the orientation update step. Note that, in the presence of gravity, the accelerometer measurement can be used for orientation estimation under the following two assumptions:

1) During the moving process of the unmanned vehicle, the forces are essentially balanced, i.e., the vehicle is moving at uniform speed or hovering at a fixed point. That is, the accelerometer measurement in the global coordinate is $(0, 0, g)^\top$.
2) The sensor measurement data only involves high-frequency or low-frequency noise, that is, the sensor measurement equation is as follows:

$$\begin{aligned} \hat{a}(t) &= a(t) + n_H(t) \\ \hat{w}(t) &= w(t) + n_L(t), \end{aligned} \tag{3}$$

here, the variable with a caret ($\hat{}$) represents the measured value, while the variable without a caret represents the ground truth value. Note that the two assumptions hold for most autonomous vehicle operation periods, we fuse $\hat{a}$ into the orientation estimation pipeline.

Although subject to noise, orientation estimates from accelerometer measurements exhibit sustained accuracy over long periods. Conversely, orientation estimates derived from gyroscope measurements offer precision over short periods but are susceptible to drift over long durations. These distinct characteristics can be exploited and harnessed within the frequency domain [29]. Specifically, orientation estimates predicated on gyroscope measurements demonstrate advantageous properties at high frequencies, thus the gyroscope measurement is filtered by a High Pass Filter (HPF). In contrast, orientation estimates obtained from accelerometer measurements exhibit favorable attributes at low frequencies, thereby the accelerometer measurement is filtered by a Low Pass Filter (LPF). This approach facilitates the effective utilization of the unique properties inherent to each sensor's measurements provided by the IMU, thereby optimizing system performance. The diagram of a complementary filter on IMU measurement is shown in Fig. 2. The mathematical derivation of the one-dimensional complementary filter and Madgwick filter are detailed in IV-B and IV-C respectively.

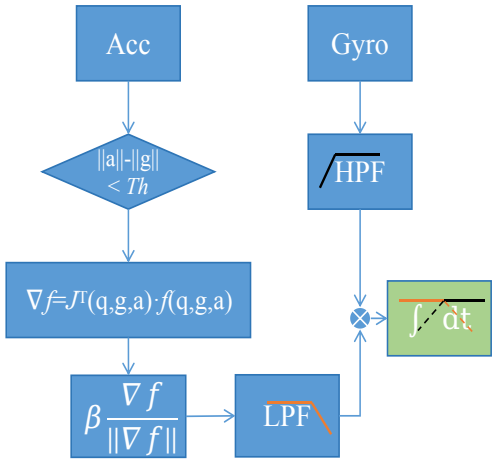


Fig. 2. Madgwick filter forms the perspective of the complementary filter.

B. One Dimensional Complementary Filter from Frequency Domain to Time Domain

Consider the one-dimensional case, assuming the IMU rotates along one axis (taking the X-axis as an example here). We denote the estimation of an angle θ from gyroscope measurement by $\hat{\theta}_w$ and the estimation from accelerometer measurement by $\hat{\theta}_a$. $\hat{\theta}_w$ is obtained by integrating gyroscope measurement w_{gyro} and $\hat{\theta}_a$ is obtained by calculating the angle between gravity and the acceleration direction. The Laplace transforms of θ, $\hat{\theta}_a$, $\hat{\theta}_w$, and $\hat{w}$ are denoted by $\Theta(s)$, $\hat{\Theta}_a(s)$, $\hat{\Theta}_w(s)$, and $\hat{W}(s)$, respectively. The complementary filter computes $\Theta(s)$ as:

$$\begin{aligned}\Theta(s) &= LPF(s)\hat{\Theta}_a(s) + HPF(s)\hat{\Theta}_w(s) \\ &= LPF(s)\hat{\Theta}_a(s) + (1 - LPF(s))\frac{1}{s}\hat{W}(s). \end{aligned} \tag{4}$$

Where $LPF(s)$ is a low-pass filter and $HPF(s) = 1 - LPF(s)$ is hence a high-pass filter. The insight behind the name complementary filter of (4) is that it leverages the complementarity nature of the measurement provided by the gyroscope and the accelerometer.

Taking $LPF(s) = \frac{C(s)}{s+C(s)}$ as a first-order low-pass filter, the output of the complementary filter can be written as:

$$\Theta(s) = \frac{C(s)}{s + C(s)}\hat{\Theta}_a(s) + \frac{s}{s + C(s)}\frac{1}{s}\hat{W}(s). \tag{5}$$

Set $C(s)$ in the low-pass filter as $\frac{1}{a}$, (5) becomes:

$$\Theta(s) = \frac{1}{as+1}\hat{\Theta}_a(s) + \frac{1}{s}\frac{as}{as+1}\hat{W}(s). \tag{6}$$

Substituting $s = \frac{1-z^{-1}}{T}$ into (6), we have the Z-transform of the complementary filter:

$$\Theta(z) = \frac{1}{a\frac{1-z^{-1}}{T} + 1}\hat{\theta}_a(z) + \frac{a}{a\frac{1-z^{-1}}{T} + 1}\hat{W}(z). \tag{7}$$

Convert the Z-transform (7) to the discrete-time difference form, we have:

$$\begin{aligned}\theta(t) &= \frac{T}{a+T}\hat{\theta}_a(t) + \frac{a}{a+T}(\theta(t-1) + Tw(t)) \\ &= (1-\gamma)\hat{\theta}_a(t) + \gamma(\theta(t-1) + Tw(t)), \end{aligned} \tag{8}$$

where γ is $\frac{a}{a+T}$. (4) and (8) are equivalent, except that they represent signal fusion from the perspectives of the frequency domain and time domain, respectively. In the frequency domain, the signal is filtered using a high-pass or low-pass filter to remove noise. In the time domain, it is represented as a weighted complementary filter. In (8), the parameter $\gamma = \frac{a}{a+T}$, represents the weight assigned to the accelerometer, which corresponds to the low-frequency signal. The choice of this weight is related to the cutoff frequency, denoted as a of the low-pass filter. This allows us to derive the complementary filter from the frequency domain to the time domain for one dimension. Having completed the derivation of the complementary filter for one dimension, we now move on to deriving the 3D complementary filter using IMU data as outlined in Section IV-C.

C. Complementary Filter on Quaternions

Having the one-dimensional complementary filter, we can extend it to three-dimensional rotation. In order to achieve a high-precision three-dimensional attitude estimation, we introduce the Madgwick filter here. Madgwick filter is primarily used for attitude estimation, utilizes data from the accelerometer, and applies it to the orientation propagation. It becomes particularly effective when the visual-inertial sensor is nearing a uniform state and demonstrating steady motion as the assumption in IV-A. In scenarios where there is an

absence of external acceleration (i.e., the sensor is either in a state of uniform motion or remains steady), the gravitational field is expected to align with the acceleration measurement field. This alignment is a critical aspect of maintaining the accuracy of the attitude estimation, and can be described by:

$$\begin{aligned} a &= q_{iw}^{\oplus\top} q_{iw}^{+} g \\ &= R_{iw}(q) g. \end{aligned} \tag{9}$$

The R_{iw} can be explicitly rewrite as:

$$\begin{bmatrix} 1-2q_y^2-2q_z^2 & 2q_xq_y-2q_zq_w & 2q_xq_z-2q_yq_w \\ 2q_xq_y+2q_zq_w & 1-2q_x^2-2q_z^2 & 2q_yq_z+2q_xq_w \\ 2q_xq_z-2q_yq_w & 2q_yq_z+2q_xq_w & 1-2q_x^2-2q_y^2. \end{bmatrix} \tag{10}$$

By the alignment of the orientations of the gravitational field and the acceleration measurement field, it becomes possible to feed back the acceleration to the orientation estimation. Specifically, when the norm of the incoming acceleration approximates the magnitude of gravity$\|a\| - \|g\| < \pi_{sh}$, the orientation estimation derived from the accelerometer can be calculated by minimizing the error term by solving (11) and (12).

$$\arg\min_{q_a} f(q_a, g, a). \tag{11}$$

$$f(q_a, g, a) = R_{iw}(q_a) g - a. \tag{12}$$

Where q_a denotes the orientation solved by accelerometer measurement. Assuming that the direction of gravity is aligned with the vertical direction, or z-axis during the UAV flight task. By substituting $g = (0, 0, 1)$ and the normalized accelerometer measurement $\hat{a}$, we can obtain a simplified version of the objective function and the Jacobian, as defined by (13) and IV-C.

$$f(q_a, g, a) = \begin{bmatrix} 2q_xq_z - 2q_yq_w - a_x \\ 2q_yq_z + 2q_xq_w - a_y \\ 1 - 2q_x^2 - 2q_y^2 - a_z \end{bmatrix}, \tag{13}$$

and the Jacobian of $f(q_a, g, a)$ is defined by:

$$J_{q_a}(q_a, g, a) = \begin{bmatrix} -2qy & 2q_z & -2q_w & 2q_x \\ 2q_x & 2q_w & 2q_z & 2q_y \\ 0 & -4q_x & -4q_y & 0. \end{bmatrix} \tag{14}$$

This optimization problem can be solved by the gradient descent algorithm:

$$q_a(n+1) = q_a(n) - \mu \frac{\nabla f(q_a(n), g, a)}{\|\nabla f(q_a(n), g, a)\|}, n = 1, 2, ..., n, \tag{15}$$

where $\nabla f(q_a(n), g, a) = J_q^\top(q_a(n), g, a) f(q_a(n), g, a)$. An appropriate value of μ will ensure the convergence rate of q is limited to the physical orientation rate as this avoids overshooting due to an unnecessarily large step size. The range of a reasonable μ is given by [28] as in (16).

$$\mu = \alpha \|\dot{q}_a\| \Delta t, \quad \alpha > 1 \tag{16}$$

Thus the orientation by the accelerometer is solved and can be further fused with the orientation estimate by the gyroscope measurement. Initially, the fuse orientation is given by (17) as in (4) and (8) :

$$q(t) = \gamma q_a + (1-\gamma) q_w, \quad 0 \le \gamma \le 1. \tag{17}$$

Following [28], (17) is converted to (18)

$$q(n) = q(n-1) + \dot{q} \Delta t, \tag{18}$$

where $\dot{q}$ is given by (19):

$$\dot{q} = \frac{1}{2} q_w \otimes w - \beta \frac{\nabla f(q_a(n), g, a)}{\|\nabla f(q_a(n), g, a)\|}. \tag{19}$$

The orientation given by (18) is fed into the pipeline as shown in Fig. 1 to offer a better orientation initialization for the following steps.

D. Visual and LiDAR measurement model

FLiVIS is composed of VIO and LIO subsystems. The VIO subsystem builds upon our prior work, FLVIS [2], while the LiDAR measurement model in LIO can be substituted with any LiDAR odometry. In this study, we employ measurement models based on Fast-Lio2 [3]. This modular approach allows our LVIO to adapt to a wide range of environments. It's worth noting that the same LiDAR point registration method may yield significantly different results in varying environments. This study emphasizes the use of a complementary filter for IMU data processing and a framework based on feedforward and feedbacks for sensor fusion. As a result, we decouple the visual and LiDAR measurement models from the feedforward-feedback based LVIO system. The following sections will provide a detailed introduction to both the visual and LiDAR measurement models.

1) VIO Subsystem: The visual measurement model, shown in Fig. 3, is based on the extraction of visual feature points and the minimization of re-projection error. The 3D points are obtained through triangulation. The 3D points are then reprojected onto the image plane to compute the re-projection error with their 2D matches. Denote a 3D point $P_i = [X_i, Y_i, Z_i]\top$ and its corresponding 2D point $p_i = [u_i, v_i]\top$. Take the camera intrinsic matrix K_c, we can solve the camera pose by minimizing the re-projection error as follows:

$$T_c = \arg\min_{T_c} \frac{1}{2} \sum_{i=1}^{N} \|p_i - KTP_i\|^2. \tag{20}$$

2) LIO Subsystem: In the LIO subsystem of FLiVIS, the LiDAR measurement model is based on the point-to-plane measurement model of the ikd-tree [3]. The points collected by the LiDAR are first back-propagated through the IMU data and timestamps to obtain undistorted point clouds. As shown in Fig. 4, the nearest point is obtained through the K-nearest-neighbor (KNN) search, Q_j is a point in the K nearest neighbor of the point P_i, and the norm n_i is calculated. Finally, the pose of the LiDAR is obtained by

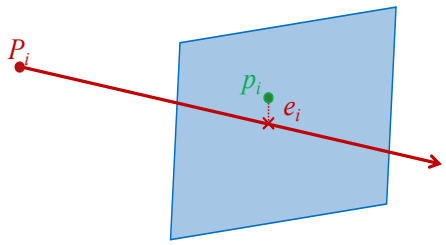


Fig. 3. Visual measurement model, the error is the 2D distance between the 3D waypoint P_i and the 2D image point p_i on the image plane.

minimizing the following cost function:

$$T_{LiDAR} = \arg\min_{T_{LiDAR}} \frac{1}{2} \sum_{i=1}^{N} \|n_i (T_{LiDAR} P_i - Q_j\|^2). \tag{21}$$

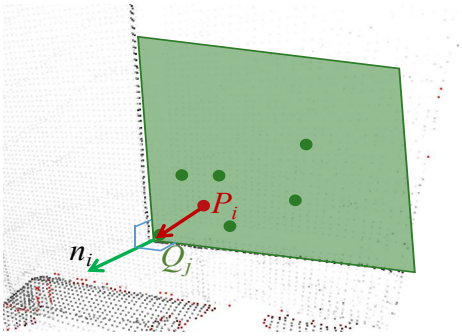


Fig. 4. LiDAR measurement model, the vector between point P_i(in red) in the current scan and its neighbor point Q_j(in green) is orthogonal to the normal vector of the k-nearest neighbors of P_i.

V. Experiments

In this section, we evaluate the proposed system using the public dataset KITTI [4] and NTU-Viral [5] to simulate the running environments of both ground and aerial vehicles. The VIO, LIO subsystems, and the LVIO system are tested. An ablation study is also conducted to demonstrate the effects of the Madgwick filter introduced in IV-C.

A. KITTI Dataset

In this section, we compare the performance of FLiVIS with three state-of-the-art algorithms, namely FLVIS [2], LOAM [15], and line-based LVO [30], utilizing the KITTI dataset [4]. Given that the frequency of the LiDAR, visual, and IMU data is 10HZ, The LIO and VIO subsystems update at the same frequency. Our primary data sources are LiDAR data and stereo images. For FLVIS, we retained the same parameters as our previous setting in our previous work. Fast-Lio, serving as the LO subsystem, forms FLiVIS in conjunction with FLVIS. For LOAM [15], scan-to-scan LiDAR odometry and scan-to-map optimization are employed. The results of the LVO [30] are consistent with those reported in the original paper, as no open-source code is available. It is worth noting that the original paper did not include information regarding rotation error for the line-based LVO method. Thus, a comparison of rotation error between the studies cannot be made.

The relative translation and rotation error is shown in Table I, the LVO-based methods are more accurate than the VO-based and LO-based methods for all sequences. While the line-based LVO [30] achieves the highest accuracy in sequences 04, 06, and 10, the average accuracy of FLiVIS exceeds the line-based LVO [30] by a large margin. It should be noted that the accuracy of LOAM, as presented in Table I, is not as accurate as the results reported on the KITTI odometry leaderboard. This discrepancy can be attributed to the focus on real-time performance. Therefore, the canonical LOAM [15] was chosen as our baseline. The LOAM results showed on the KITTI odometry leaderboard do not operate in real-time. Instead, they were processed at a rate that is 10% of real-time speed, taking approximately one second to process a single scan. For more details, please refer to Section 7.4 of [31].

TABLE I
Quantitative Results on KITTI Dataset

	VO	LO	LVO	
Sequence	FLVIS	LOAM	LVO	FLiVIS
02	2.30/5.7	1.33/15.7	1.38	**0.58/1.5**
04	2.77/5.6	1.39/4.7	**0.42**	0.54/**0.5**
06	2.14/2.4	1.29/5.6	**0.61**	1.1/**1.8**
08	1.22/5.2	0.64/5.2	1.27	**0.32/1.3**
10	1.18/2.5	1.97/6.5	**0.83**	0.84/**0.8**
average	1.92/4.28	1.32/7.54	0.90	**0.68/1.02**

Note that the performance is represented in 'A/B' format, where 'A' indicates the relative translation error in %, and 'B' indicates the relative rotational error in 10^{-3} degrees per meter.

Outdoor scenes pose significant challenges for visual odometry due to factors such as changes in lighting and a lack of feature points. When there are too few feature points within the camera's field of view to perform feature tracking and matching, FLVIS may fail or be subject to drift, Fig. 5 shows a challenging scene for VIO where more than half of the image is occupied with the sky. Visual feature points are concentrated in the middle and lower parts of the image, with a very focused distribution. This can easily lead to divergence during the optimization process, causing drift in visual odometry, and even failure.

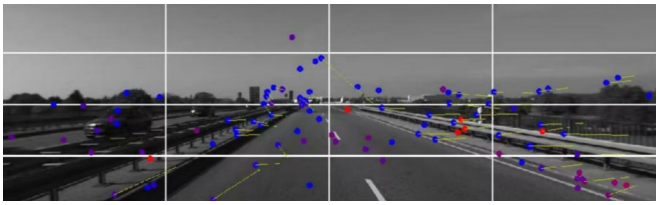


Fig. 5. Challenging scene for feature-based VIO, where the majority of an image is occupied by the sky, with feature points concentrated in the middle and lower parts of the image. This situation can easily lead to the failure of feature tracking algorithms, and the pose estimation algorithm may become unstable, or even diverge.

The system becomes more robust after incorporating LiDAR odometry. When visual odometry encounters a failure, the system re-initializes visual odometry based on the output of LiDAR odometry. A qualitative comparison between the SVIO FLVIS and proposed LVIO FLiVIS can be seen in Fig. 6, it shows the result of the two algorithms on sequence 12 in the KITTI dataset. The error of FLiVIS is smaller than that of FLVIS, and along the whole trajectory, it is aligned with the ground truth.

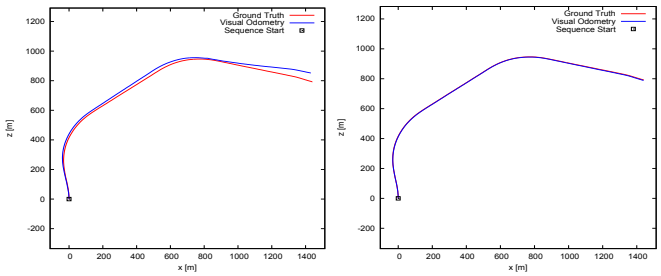


Fig. 6. Qualitative comparison on KITTI dataset. (a) FLVIS result, and (b) FLiVIS result.

B. NTU-Viral Dataset

As stated in [5] Section 1.1, there exists a distinct dichotomy between aerial and terrestrial autonomous vehicles, characterized by the differences in their onboard sensors selection and movement dynamics. In the past, 3D LiDAR was often disregarded for aerial datasets due to payload constraints. The emphasis is typically on high frame-rate camera systems, which provide self-localization capabilities during high-speed and aggressive maneuvers. VIO and V-SLAM systems have been extensively studied to handle increasingly extreme conditions. With the development of LiDAR technology and the challenges of increasingly complex environments, there is now a trend towards equipping UAVs with LiDAR. The NTU-Viral dataset collected data across various outdoor scenarios with a UAV equipped with LiDAR-visual-inertial and UWB sensors. It provides ground truth data for evaluating the accuracy of SLAM algorithms.

The sensor selection for the NTU-Viral dataset is similar to other autonomous driving datasets, but the installation method differs due to the structure of the drone. On the drone, one LiDAR points downwards and another is horizontal. The small baseline does not affect the performance of SVO in indoor environments, but in outdoor environments with greater depth, it poses additional challenges for stereo triangulation.

From the perspective of data collection, there are significant differences between aerial and ground datasets. Ground vehicles can essentially be considered as moving on a two-dimensional plane, while drones experience large changes in roll and pitch. Drones often perform actions of acceleration and sudden stops. The aerial dataset is smaller than the ground vehicle dataset because the endurance time of ground vehicles far exceeds that of small drones. Shorter paths and more intense motion make drone positioning more challenging than ground vehicle positioning.

TABLE II
ABSOLUTE TRAJECTORY ERROR IN (m) ON NTU-VIRAL DATASET

Seq. num	Vins-fusion	M-LOAM	Fast-LVIO	FLiVIS
eee_01	0.608	0.249	0.280	**0.130**
eee_02	0.506	0.166	0.170	**0.125**
eee_03	0.494	0.232	0.230	**0.162**
sbs_01	0.508	0.173	0.290	**0.142**
sbs_02	0.564	0.147	0.220	**0.140**
sbs_03	0.878	0.153	0.220	**0.132**
average	0.593	0.187	0.235	**0.138**

The performance of FLiVIS with SOTA SLAM algorithms using different sensor combinations, i.e. Vins-fusion [24](SVIO), M-LOAM [32] (LO), and Fast-LVIO [25] (LVIO) are evaluated and compare in this section. As can be seen from Table II, FLiVIS achieved the highest accuracy in all sequences, and it performed better than Fast-LiVO, which is also an LVIO. It can be observed that the localization accuracy progressively increases from VIO to LO and then to LVIO. This is because LO results are more reliable under varying lighting conditions, and during intense movements, IMU data provides additional robustness.

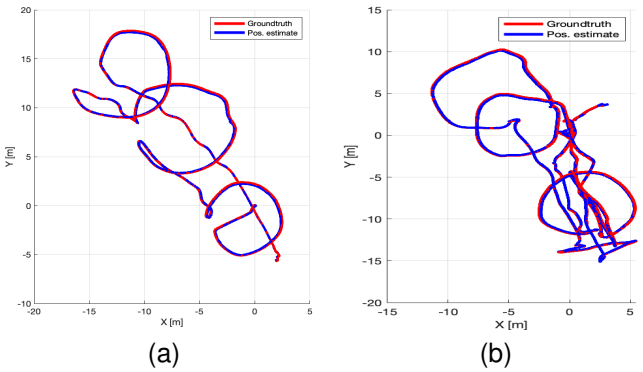
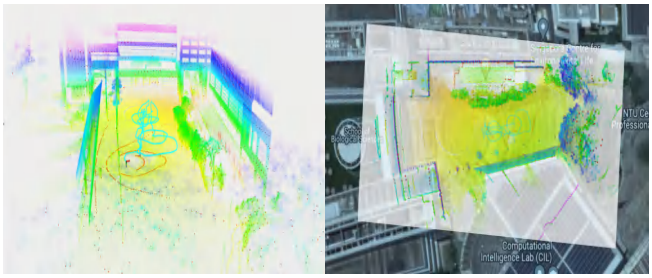


Fig. 7. Trajectory of FLiVIS on the NTU-Viral dataset. (a) sequence eee_03, and (b) sequence sbs_03.

Fig. 7 shows the trajectories of FLiVIS and ground truth in the eee_03 sequence and sbs_03 sequence. Compared with Fig. 6, it can be seen that the drone's motion trajectory is more curved and irregular, while the vehicle's motion trajectory is almost a smooth curve. Moreover, the drone's motion range is smaller, with the overall range within 100 meters, while in Fig. 6, the vehicle's motion range in the x and y directions is over 1000 meters. Despite this, the trajectory derived from FLiVIS is almost identical to the ground truth, demonstrating the accuracy and robustness of the method we proposed.

Fig. 8 shows the LiDAR mapping result and the alignment of LiDAR mapping with Google Maps. It can be seen that the LiDAR map aligns well with Google Maps.



(a) (b)

Fig. 8. Mapping result of School of Biological Sciences, (a) LiDAR mapping result, and (b) top-down view of LiDAR mapping aligns with Google map.

VI. CONCLUSIONS

In this paper, we introduce a LiDAR-stereo-visual-inertial pose estimation method, which comprises an LIO subsystem and a VIO subsystem. The issue of sensor fusion is approached from a frequency domain perspective, and data from different sensors are integrated using a complementary filter. Our proposed system has been tested on the KITI and NTU-Viral datasets, achieving a relative translation error of 0.68 % and an absolute translation error of 0.138 m respectively. These results are competitive with those of other state-of-the-art algorithms. As part of our future work, we plan to directly update visual and LiDAR observations in a joint manner.

REFERENCES

[1] K. Ebadi, L. Bernreiter, H. Biggie, G. Catt, Y. Chang, A. Chatterjee, C. E. Denniston, S.-P. Deschênes, K. Harlow, S. Khattak *et al.*, "Present and future of slam in extreme environments: The darpa subt challenge," *IEEE Transactions on Robotics*, 2023.

[2] S. Chen, Y. Feng, C.-Y. Wen, Y. Zou, and W. Chen, "Stereo visual inertial pose estimation based on feedforward and feedbacks," *IEEE/ASME Transactions on Mechatronics*, 2023.

[3] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang, "Fast-lio2: Fast direct lidar-inertial odometry," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2053–2073, 2022.

[4] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[5] T.-M. Nguyen, S. Yuan, M. Cao, Y. Lyu, T. H. Nguyen, and L. Xie, "Ntu viral: A visual-inertial-ranging-lidar dataset, from an aerial vehicle viewpoint," *The International Journal of Robotics Research*, vol. 41, no. 3, pp. 270–280, 2022.

[6] T. Pire, T. Fischer, G. Castro, P. De Cristóforis, J. Civera, and J. J. Berlles, "S-ptam: Stereo parallel tracking and mapping," *Robotics and Autonomous Systems*, vol. 93, pp. 27–42, 2017.

[7] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[8] R. Mur-Artal and J. D. Tardós, "Probabilistic semi-dense mapping from highly accurate feature-based monocular slam." in *Robotics: Science and Systems*, vol. 2015. Rome, 2015.

[9] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[10] C. Fu, A. Carrio, and P. Campoy, "Efficient visual odometry and mapping for unmanned aerial vehicle using arm-based stereo vision pre-processing system," in *2015 international conference on unmanned aircraft systems (ICUAS)*. IEEE, 2015, pp. 957–962.

[11] 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.

[12] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, "Openvins: A research platform for visual-inertial estimation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 4666–4672.

[13] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint kalman filter for vision-aided inertial navigation," in *Proceedings 2007 IEEE international conference on robotics and automation*. IEEE, 2007, pp. 3565–3572.

[14] R. Milijas, L. Markovic, A. Ivanovic, F. Petric, and S. Bogdan, "A comparison of lidar-based slam systems for control of unmanned aerial vehicles," in *2021 International Conference on Unmanned Aircraft Systems (ICUAS)*. IEEE, 2021, pp. 1148–1154.

[15] J. Zhang and S. Singh, "Loam: Lidar odometry and mapping in real-time." in *Robotics: Science and systems*, vol. 2, no. 9. Berkeley, CA, 2014, pp. 1–9.

[16] J. W. Kim and B. H. Lee, "Robust and fast 3-d scan registration using normal distributions transform with supervoxel segmentation," *Robotica*, vol. 34, no. 7, pp. 1630–1658, 2016.

[17] C. Bai, T. Xiao, Y. Chen, H. Wang, F. Zhang, and X. Gao, "Faster-lio: Lightweight tightly coupled lidar-inertial odometry using parallel sparse incremental voxels," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4861–4868, 2022.

[18] H. Guo, J. Zhu, and Y. Chen, "E-loam: Lidar odometry and mapping with expanded local structural information," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1911–1921, 2022.

[19] T. Shan, B. Englot, D. Meyers, W. Wang, C. Ratti, and D. Rus, "Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping," in *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2020, pp. 5135–5142.

[20] C. Qin, H. Ye, C. E. Pranata, J. Han, S. Zhang, and M. Liu, "Lins: A lidar-inertial state estimator for robust and efficient navigation," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 8899–8906.

[21] H. Ye, Y. Chen, and M. Liu, "Tightly coupled 3d lidar inertial odometry and mapping," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3144–3150.

[22] J. Lin, C. Zheng, W. Xu, and F. Zhang, "R2 live: A robust, real-time, lidar-inertial-visual tightly-coupled state estimator and mapping," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7469–7476, 2021.

[23] T. Shan, B. Englot, C. Ratti, and D. Rus, "Lvi-sam: Tightly-coupled lidar-visual-inertial odometry via smoothing and mapping," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 5692–5698.

[24] T. Qin, J. Pan, S. Cao, and S. Shen, "A general optimization-based framework for local odometry estimation with multiple sensors," *arXiv preprint arXiv:1901.03638*, 2019.

[25] C. Zheng, Q. Zhu, W. Xu, X. Liu, Q. Guo, and F. Zhang, "Fast-livo: Fast and tightly-coupled sparse-direct lidar-inertial-visual odometry," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 4003–4009.

[26] M. Kok, J. D. Hol, and T. B. Schön, "Using inertial sensors for position and orientation estimation," *arXiv preprint arXiv:1704.06053*, 2017.

[27] T. D. Barfoot, *State estimation for robotics*. Cambridge University Press, 2024.

[28] S. O. Madgwick, A. J. Harrison, and R. Vaidyanathan, "Estimation of imu and marg orientation using a gradient descent algorithm," in *2011 IEEE international conference on rehabilitation robotics*. IEEE, 2011, pp. 1–7.

[29] W. T. Higgins, "A comparison of complementary and kalman filtering," *IEEE Transactions on Aerospace and Electronic Systems*, no. 3, pp. 321–325, 1975.

[30] S.-S. Huang, Z.-Y. Ma, T.-J. Mu, H. Fu, and S.-M. Hu, "Lidar-monocular visual odometry using point and line features," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 1091–1097.

[31] J. Zhang and S. Singh, "Low-drift and real-time lidar odometry and mapping," *Autonomous Robots*, vol. 41, pp. 401–416, 2017.

[32] J. Jiao, H. Ye, Y. Zhu, and M. Liu, "Robust odometry and mapping for multi-lidar systems with online extrinsic calibration," *IEEE Transactions on Robotics*, vol. 38, no. 1, pp. 351–371, 2021.