A One-Step Visual-Inertial Ego-Motion Estimation using Photometric Feedback

Shixin Tan, Shangkun Zhong, and Pakpong Chirarattananon, *Member, IEEE/ASME*

***Abstract*—This article presents a robust brightness gradient-based estimation strategy for small aerial robots. The proposed nonlinear observer is capable of estimating the flight altitude and ego-motion by the fusion of monocular vision and inertial measurement unit feedback. The novelty primarily lies in the implementation of the gradient-based featureless approach and the direct use of photometric feedback in the state and output vectors. Under the single-plane assumption, the proposed framework permits the entire estimation process to be accomplished efficiently in a single iterative step without the need for feature detection and tracking or pre-computation of optic flow as commonly seen in conventional methods. The nonlinear and featureless implementation reduces the computational demand, enlarges the region of attraction, and markedly improves the robustness of the ego-motion estimation against scenes with scarce features when compared to Kalman-based estimators and feature-based methods. We conducted extensive flight experiments with different flying patterns and textures to evaluate the performance of the proposed observer. The results reveal that the root-mean-square error in the altitude estimates is approximately 10% for flights at $\approx 40-130$ cm above the ground, comparable to feature-based estimators. Nevertheless, the devised observer does better than feature-based methods when deployed on low-textured scenes or with low-resolution blurry images.**

***Index Terms*—optic-flow, direct gradient-based methods, inertial measurement unit, nonlinear observer**

I. Introduction

MICRO Aerial Vehicles (MAVs) have gained tremendous popularity in recent decades. The advancements are driven by numerous civilian and defense applications, including espionage, transportation, inspection, and agriculture. With the progress in miniaturization and increasing integration of robots in everyday life, MAVs are anticipated to operate in urban and confined environments [1]. To this end, technical challenges related to perception and navigation must be addressed to allow these small flying machines to safely negotiate unstructured and GPS-denied environments.

Unlike terrestrial platforms that prevalently rely on multiple sensors such as LiDAR and cameras to compute RGB-depth images [2], small aerial robots are severely limited in payload and power budget. Among MAVs, an integration of an onboard camera and inertial sensor or Visual-Inertial Systems (VINS) have emerged as a common framework for providing state estimation and localization [3]–[7]. The fusion of readings from an inertial measurement unit (IMU) with visual feedback allows retrieval of the scale factor, leading to accurate pose estimates through iterative refinements. Widely used strategies involve feature detection, tracking, map reconstruction.

Despite the capability of existing map-based VINS to provide rich and accurate information, there exist several shortcomings. The robustness of feature-based methods critically depends on the ability to track features, landmarks, and retain key frames over a prolonged period. In addition to the computational resource required [8], the feature detection and tracking process could be adversely affected by poor image quality, motion blur, and low-textured scenes [9]. Besides, the refinement over a large number of poses and landmarks is demanding. This renders the solution unsuitable for smaller MAVs with severely restricted payload and power [10]–[12].

An alternative lightweight vision-based strategy suitable for ego-motion estimation for small drones is reactive navigation [13]. These algorithms directly leverage apparent motion or optic flow from most recent images to infer ego-motion when other devices are absent. When incorporated with another sensor modality, such as a time-of-flight camera [14], ultrasonic sensor, [15] or IMU [16]–[18], true flight velocity and distance can be estimated.

A. Related works and contributions

This paper addresses the optic flow-based ego-motion estimation problem for small flying robots. The omission of mapping of reactive approaches not only reduces the computational complexity, but also improves the robustness since extended tracking of landmarks is no longer required. Most existing reactive ego-motion estimation algorithms operate in multiple stages. The first step is concerned with the feature identification [19], [20], followed by the computation of optic flow via feature tracking with the Lucas-Kanade algorithm [21]. The optic flow is then fused with IMU measurements through either an optimization-based [17], [22] or Extended Kalman Filter-based [16] algorithm to produce ego-motion estimates. Among these, the findings from [23], [24] suggest that the feature detection is the most costly step, even when the reconstruction of map points are taken into consideration.

To improve the robustness and efficiency of the ego-motion estimation, an inspiration can be drawn from the direct or featureless methods [25], [26] used by map-based VINS. In such cases, the motion is inferred from the dense photometric feedback, bypassing the feature identification and tracking. The image intensity, even from areas where gradients are

The authors are with the Department of Biomedical Engineering, City University of Hong Kong, Hong Kong SAR, China (e-mail: sxtan2-c@my.cityu.edu.hk; shanzhong4-c@my.cityu.edu.hk; pakpong.c@cityu.edu.hk).

small, is directly used in the optimization or filtering step [27]. Since the entire image is exploited, it has been shown to outperform feature-based methods in terms of robustness to motion blur or low-textured images when implemented with map-based localization [28]. Nevertheless, the direct methods suffer from the elevated computational cost from the generation of dense map when applied in the context of localization. This is alleviated by uses of patches of photometric feedback. The patch-based or semi-direct approaches offer a balance between robustness and complexity by making use of small image patches around identified features [4].

To date, uses of the direct approach in lightweight estimators for MAVs are still limited. In [29], the authors devised a featureless method for controlling mobile robots based on time-to-contact. For aerial vehicles, the featureless approach was proposed to replace the optic flow measurements from the feature tracking [30]. By taking image gradients, motion information is extracted. When combined with readings from an IMU through an EKF, estimates of flight altitude were obtained. In these examples, the altitude is recovered in multiple steps as the flow divergence (ratio of velocities in x-y-z directions to the distance) are computed first, followed by the altitude estimation.

This work takes a consolidated approach to tackle the ego-motion estimation based on the featureless method. Unlike previous direct estimation methods, where filters or observers require the knowledge of pre-computed optic flow [17], [22] or flow divergence (ratio of flight velocity to the distance) [30] for each iteration, the entire ego-motion estimation for each iteration herein is accomplished in a single step. To achieve this, photometric feedback from an entire image is part of the state vector and directly used as measurements of the proposed estimator. The evolution of the state, output, and their predictions are tightly coupled with the ego-motion and corresponding image gradients. With IMU readings, the scheme efficiently estimates the inverse altitude, flow divergence, and the plane's normal in a single step assuming there exists only a single plane under the camera's view. The single plane assumption is prevalent among optic flow-based ego-motion estimation algorithms due to the exclusion of mapping [17], [18], [22], [29], [30]. Nevertheless, the proposed framework does not impose a restriction on the camera's motion or the plane's inclination as found in [17], [18], [30].

In the proposed one-step featureless approach, photometric feedback is modeled and included in the state and output. This results in a large output vector whose length is equal to the number of pixels observed. The size of the output renders an EKF-based estimator impractical as the computation of the Kalman gain involves an inverse operation on a matrix of which the dimension is dictated by the length of the output vector. This necessitates the development of a nonlinear observer (NLO) that produces the estimates of flight altitude, flow divergence, and plane's normal. With a convergence proof, the NLO has certain advantages over EKF-based algorithms. It provides a quantifiable convergence rate (but lacking the estimates of the covariance) [22]. Without linearization, a nonlinear approach possesses a large region of attraction and therefore is less sensitive to disturbances or poor initial conditions [17], [22]. Overall, the use of dense photometric feedback provides improved robustness over feature-based methods and the one-step implementation, which cannot be achieved with indirect methods, radically streamlines the estimation. It is conceivable that the efficiency and robustness of the proposed ego-motion estimator potentially renders it a suitable estimator for a lightweight reactive navigation strategy to be employed by MAVs with acute computational constraints such as small or insect-sized flying robots [10]–[12].

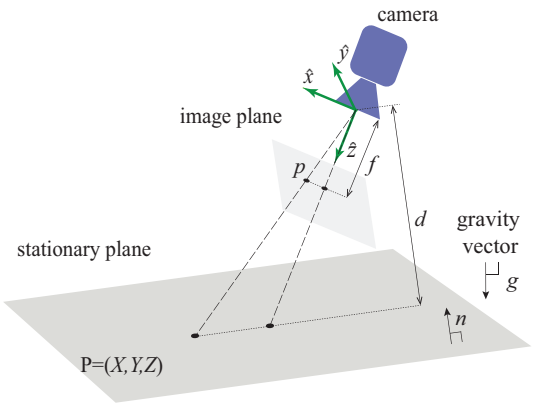


Fig. 1. A diagram of a moving camera, its associated image plane, and the ground. The IMU's frame is assumed coincident with the camera's frame.

This paper is structured as follows. Section II provides descriptions of the camera-centric odometry, optic flow, and formulation of the equations of motion. Section III presents the NLO and its convergence proof based on the state dynamics from Section II in the discrete-time domain. The proposed NLO is validated through a series of flight experiments in section IV, accompanied by the analysis of the results and the comparison to an EKF-based estimator, a feature-based approach, and a map-based state-of-the-art method. Lastly, in section V, a conclusion and future directions are discussed.

II. Ego-motion, Optic Flow, and Dynamics

A. Problem formulation

Consider a scenario where a moving camera observes points on a plane as depicted in Fig. 1. The ego-motion problem considered here begins with the formulation that relates the distance between the camera and the plane, image intensity values, and optic flow. In this work, the distance, relative orientation between the camera and the plane, and flow divergence [24] are regarded as unknowns to be estimated. Photometric feedback or pixel irradiance from entire images is directly treated as measurements. In addition, the proposed method requires the knowledge of the camera-centric angular velocity and linear acceleration which can be provided by an IMU (through an accelerometer and a gyroscope). Without loss of generality, the camera's frame and IMU's frame are assumed to coincide. The framework can be applied in the context of reactive navigation of mobile or aerial robots.

B. Continuous homography constraint

Fig. 1 shows a camera moving with respect to a stationary plane with the camera frame and inertia frame. The camera

frame is defined such that the Z axis aligns with the optical axis. The camera-centric linear and angular velocities with respect to the inertia frame are $\boldsymbol{v} = [v_x, v_y, v_z]^T$and $\boldsymbol{\omega} = [\omega_x, \omega_y, \omega_z]^T$. The apparent velocity of a point $\boldsymbol{P}$=$[X, Y, Z]^T$ on the plane due to the camera motion is $\dot{\boldsymbol{P}} = [\boldsymbol{\omega}]_\times \boldsymbol{P} + \boldsymbol{v}$, where $[\boldsymbol{\omega}]_\times$ is the skew-symmetric matrix associated with $\boldsymbol{\omega}$ [31]. We define $\boldsymbol{n} = [n_x\ n_y\ n_z]^T$ as a camera-centric unit vector normal to the plane. The orthogonal distance d between the camera and the plane satisfies $\boldsymbol{n}^T\boldsymbol{P}/d = 1$. Therefore,

$$\dot{\boldsymbol{P}} = [\boldsymbol{\omega}]_\times \boldsymbol{P} + \frac{\boldsymbol{v}}{d}\boldsymbol{n}^T\boldsymbol{P}. \tag{1}$$

Camera images provide the projection of $\boldsymbol{P}$, or $\boldsymbol{p} = [x, y, f]^T$, where f is the focal length with an unknown scaling parameter λ [31], such that $\lambda\boldsymbol{p} = \boldsymbol{P}$. The velocity of $\boldsymbol{p}$ on the image plane satisfies $\dot{\boldsymbol{P}} = \dot{\lambda}\boldsymbol{p} + \lambda\dot{\boldsymbol{p}}$. With equation (1), this yields

$$\dot{\boldsymbol{p}} = (\dot{\boldsymbol{P}} - \dot{\lambda}\boldsymbol{p})/\lambda = \left([\boldsymbol{\omega}]_\times + \frac{\boldsymbol{v}}{d}\boldsymbol{n}^T\right)\boldsymbol{p} - \dot{\lambda}\boldsymbol{p}/\lambda. \tag{2}$$

The quantity $\dot{\boldsymbol{p}} = [\dot{p}_x, \dot{p}_y, 0]^T$, where $\dot{p}_x$ and $\dot{p}_y$ are the *optic flow* of the point $\boldsymbol{p}$ [31], represents the apparent velocity on the image plane. To eliminate the unknown inverse depth factor λ, the equation above is pre-multiplied with the skew-symmetric matrix $[\boldsymbol{p}]_\times$ and re-writen as:

$$[\boldsymbol{p}]_\times (\dot{\boldsymbol{p}} - [\boldsymbol{\omega}]_\times \boldsymbol{p}) = [\boldsymbol{p}]_\times \frac{\boldsymbol{v}}{d}\boldsymbol{n}^T\boldsymbol{p}. \tag{3}$$

The resultant equation, also known as the continuous homography constraint [22], [31], is suitable for the estimation of d, $\boldsymbol{v}/d$, and $\boldsymbol{n}$ via the one-step direct method below.

C. Optic flow and direct method

Suppose a camera provides images with N pixels in the form of pixel irradiance, such that I_{xy} represents irradiance of the pixel at location (x, y) at time t. The brightness constancy constraint assumes I_{xy} does not change significantly during a short period of time [25] or:

$$\dot{p}_x \frac{\partial I_{xy}}{\partial x} + \dot{p}_y \frac{\partial I_{xy}}{\partial y} + \frac{\partial I_{xy}}{\partial t} = 0, \tag{4}$$

in which $\partial I_{xy}/\partial x$ and $\partial I_{xy}/\partial y$ are the spatial irradiance gradients, and $\partial I_{xy}/\partial t$ is the rate of the pixel irradiance. Here, $\dot{p}_x$ and $\dot{p}_y$ are optic flow as present by $\dot{\boldsymbol{p}}$ in equation (2). We define a 3×1 image gradient vector ∇I= $\begin{bmatrix} \partial I_{xy}/\partial x & \partial I_{xy}/\partial y & 0 \end{bmatrix}^T$ and a basis vector $\boldsymbol{e}_3 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$, equation (4) can be re-written as

$$\frac{\partial I_{xy}}{\partial t} f = -\dot{p}_x \frac{\partial I_{xy}}{\partial x} f - \dot{p}_y \frac{\partial I_{xy}}{\partial y} f = -\boldsymbol{e}_3^T [\nabla I]_\times [\boldsymbol{p}]_\times \dot{\boldsymbol{p}}. \tag{5}$$

Premultiplying equation (3) by $\boldsymbol{e}_3^T[\nabla I]_\times$ and combining it with equation (5) to eliminate of the optic flow terms yields

$$\frac{\partial I_{xy}}{\partial t} f = -\boldsymbol{e}_3^T [\nabla I]_\times [\boldsymbol{p}]_\times \frac{\boldsymbol{v}}{d}\left(\boldsymbol{n}^T\boldsymbol{p} + [\boldsymbol{\omega}]_\times \boldsymbol{p}\right). \tag{6}$$

This provides a direct relation between the camera motion ($\frac{\boldsymbol{v}}{d}\boldsymbol{n}^T$, $\boldsymbol{\omega}$) and the image gradients ($\partial I_{xy}/\partial t$, ∇I). Consequently, no feature detection nor tracking of feature points is required for distance estimation. This markedly reduces the computational demand and enhances the robustness to low-textured scenes and blurred images.

D. Equations of motion

To iteratively estimate d and other quantities of interest from equation (6), we define the flow divergence vector $\boldsymbol{\vartheta}$=$\boldsymbol{v}/d$ and the inverse distance $\alpha = d^{-1}$. Instead of directly dealing with d, the inverse distance parameterization is similar to the treatment in [4], [22] and known to produce better results. The camera provides measurements of the pixel irradiance as an $N \times 1$ vector obtained by stacking I_{xy}'s together: $\boldsymbol{I} = \{I_{xy}\} \in \mathbb{R}^N$. The state and output vectors are defined as

$$\boldsymbol{X} = \begin{bmatrix} \alpha & \boldsymbol{\vartheta}^T & n^T & \boldsymbol{I}^T \end{bmatrix}^T \tag{7}$$

$$\boldsymbol{Y} = C\boldsymbol{X} = \boldsymbol{I}, \tag{8}$$

where $C = [\boldsymbol{0}_{N\times 7}\ \boldsymbol{1}_{N\times N}]$ ($\boldsymbol{1}$ denotes an identity matrix). The time evolution of $\boldsymbol{X}$ is described by the followings:

$$\dot{\alpha} = \left(\boldsymbol{\vartheta}^T\boldsymbol{n}\right)\alpha, \tag{9}$$

$$\dot{\boldsymbol{\vartheta}} = \mathbf{a}\alpha + \left(\boldsymbol{\vartheta}^T\boldsymbol{n}\right)\boldsymbol{\vartheta} + [\boldsymbol{\vartheta}]_\times \boldsymbol{\omega}, \tag{10}$$

$$\dot{\boldsymbol{n}} = [\boldsymbol{n}]_\times \boldsymbol{\omega}, \tag{11}$$

where we have used the fact that $\dot{\boldsymbol{v}} = \mathbf{a} + [\boldsymbol{\omega}]_\times \boldsymbol{v}$ when $\mathbf{a}$ is a camera-centric linear acceleration. Evaluation of $\mathbf{a}$ requires the specific acceleration from the IMU and the knowledge of the gravity vector in the camera frame (but not magnetometer readings for the camera-centric approach). This is achieved with the sensor fusion present in standard IMUs or flight controllers (in which IMU biases are also taken care of). The use of a unit vector $\boldsymbol{n}$ to represent the orientation of the plane avoids issues related to the parameterization of a rotation. Lastly, the dynamics of $\boldsymbol{I}$ is computed from equation (6) as

$$\frac{\partial I_{xy}}{\partial t} = -\left(1/f\right)\boldsymbol{e}_3^T [\nabla I]_\times [\boldsymbol{p}]_\times \left(\boldsymbol{\vartheta}\boldsymbol{n}^T\boldsymbol{p} + [\boldsymbol{\omega}]_\times \boldsymbol{p}\right) \tag{12}$$

Together, equations (9)-(12) capture the dynamics of $\boldsymbol{X}$ as

$$\dot{\boldsymbol{X}} = \mathcal{F}(\boldsymbol{X}, \boldsymbol{\omega}, \mathbf{a}). \tag{13}$$

In this work, both $\boldsymbol{\omega}$ and $\mathbf{a}$ are provided by an IMU and presumed known.

III. Nonlinear Observer

A. Formulation of the NLO

In this section, we establish a discrete-time nonlinear observer for estimating the state vector ($\boldsymbol{X}$) from IMU measurements and image luminosity ($\boldsymbol{Y}$). To begin, let T denote a sample time. Following notations are employed:

- A subscript $(\cdot)_k$ indicates a quantity at the k^{th} time step.
- A variable $\hat{z}_{k|j}$, for example, describes the estimate of z at the k^{th} time step given observations up to and including the j^{th} time step. $\hat{z}_k$ is a shorthand form of $\hat{z}_{k|k}$.
- Given an unknown quantity z_k and its estimate $\hat{z}_k$, the estimation error $\bar{z}_k$ is defined as $\bar{z}_k = z_k - \hat{z}_k$.
- For a vector $\mathbf{z}_i \in \mathbb{R}^{3\times 1}$ for $i = 1, 2, \ldots, N$, an operator $\{\mathbf{z}_i\}_N$ represents a horizontal stacking operation of $\mathbf{z}$ such that $\{\mathbf{z}\}_N = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N] \in \mathbb{R}^{3\times N}$.

In the discrete-time domain, the dynamics described by equation (13) is approximated using the the forward finite difference as the transition model

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k + \mathcal{F}(\boldsymbol{X}_k, \boldsymbol{\omega}_k, \mathbf{a}_k)T. \tag{14}$$

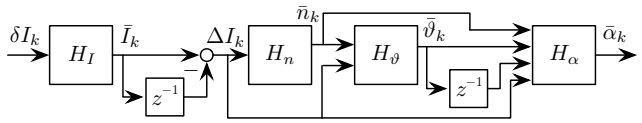


Fig. 2. A block diagram describing the input-output stability of the NLO with z^{-1} denoting the unit delay. Each equivalent subsystem $H_i(\cdot)$ is $\mathcal{L}$ stable, implying the convergence of the estimation errors.

The nonlinear observer serves as the observation model to iteratively predicts and updates the $\hat{\boldsymbol{X}}_k$ according to

$$\begin{aligned}\hat{\boldsymbol{X}}_{k+1|k} &= \hat{\boldsymbol{X}}_k + \mathcal{F}(\hat{\boldsymbol{X}}_k, \omega_k, \mathbf{a}_k)T \\ \hat{\boldsymbol{X}}_{k+1} &= \hat{\boldsymbol{X}}_{k+1|k} + \boldsymbol{G}_k(\boldsymbol{Y}_{k+1} - C\hat{\boldsymbol{X}}_{k+1|k}),\end{aligned} \tag{15}$$

where

$$\boldsymbol{G}_k = \begin{bmatrix} -2\Lambda_\alpha\Gamma_\alpha f \mathbf{a}_{k-1}^T \left\{\boldsymbol{p}^T\hat{\boldsymbol{n}}_k[\boldsymbol{p}]_\times[\nabla I_k]_\times\boldsymbol{e}_3\right\}_N \\ -2\Lambda_\vartheta\Gamma_\vartheta f \left\{\boldsymbol{p}^T\hat{\boldsymbol{n}}_k[\boldsymbol{p}]_\times[\nabla I_k]_\times\boldsymbol{e}_3\right\}_N \\ -2\Lambda_n\Gamma_n f \left\{\boldsymbol{p}\hat{\boldsymbol{\vartheta}}_k^T[\boldsymbol{p}]_\times[\nabla I_k]_\times\boldsymbol{e}_3\right\}_N \\ \mathbf{1}_{N\times N}\end{bmatrix}, \tag{16}$$

when Λ_i's and Γ_i's are symmetric positive definite matrices that satisfy the conditions: (i) $\Lambda_i \to \infty$; and (ii) $\Lambda_i\Gamma_i T \ll 1$. Readers are referred to the Supplemental Materials for the origin of equation (16).

B. Convergence of the NLO

Stability of the NLO is shown in multiple steps. The proof shows that the estimate of each element in $\boldsymbol{X}_k$ sequentially asymtotically converges in the absence of process and measurement noises. With disturbances, the estimation errors remain bounded. For simplicity, process noise and the error induced by the measurement noise in the term ∇I_k are neglected. Otherwise, they can be treated in a similar manner to the measurement noise. With these considerations, the convergences of $\bar{I}_k$, $\bar{\boldsymbol{n}}_k$, $\bar{\boldsymbol{\vartheta}}_k$, and $\bar{\alpha}_k$ are shown in the respective order as schematically depicted in Fig. 2. That is, it can be shown that there exist equivalent subsystems $H(\cdot)$'s that are $\mathcal{L}$ stable. As a result, the entire system is asymptotically stable. We introduce two following Lemmas to assist with the convergence proof.

Lemma 1. *A discrete-time system of $\boldsymbol{Z}_k$ with an input $\boldsymbol{U}_k$ and a sample time T:*

$$\boldsymbol{Z}_{k+1} = \boldsymbol{Z}_k - 2\Lambda\Gamma T\boldsymbol{A}_k^T\boldsymbol{A}_k Z_k + \boldsymbol{D}_k T + \boldsymbol{B}_k\boldsymbol{U}_k, \tag{17}$$

where $\boldsymbol{A}_k^T$ is full row rank, $\boldsymbol{D}_k$ is uniformly bounded, and Λ, Γ are constant symmetric positive definite matrices, has a globally asymptotically stable equilibrium point at $\boldsymbol{Z} = \boldsymbol{0}$ when $\Lambda \to \infty$ and $\Lambda\Gamma T \ll \mathbf{1}$. In the presence of a bounded input $\boldsymbol{U}_k$, if $\boldsymbol{B}_k$ is finite-gain $\mathcal{L}_p$ stable for $p \in [1, \infty]$, then Z_k is also finite-gain $\mathcal{L}_p$ stable.

Proof. See the Supplemental Materials.

Lemma 2. *A discrete-time system*

$$Z_{k+1} = Z_k - A_k Z_{k-1} + D_k T + \boldsymbol{B}_k\boldsymbol{U}_k, \tag{18}$$

in which $0 < B_k < \frac{2}{3}$, and D_k is uniformly bounded as $|D_k| < D_+$, is Lyapunov stable at the equilibrium point $Z = 0$. Moreover, the system becomes asymptotically stable as $T \to 0$. In the presence of a bounded input $\boldsymbol{U}_k$, if $\boldsymbol{B}_k$ is finite-gain $\mathcal{L}_p$ stable for $p \in [1, \infty]$, then Z_k is also finite-gain $\mathcal{L}_p$ stable.

Proof. See the Supplemental Materials.

1) Convergence of $\bar{\boldsymbol{I}}_k$: Suppose $\delta\boldsymbol{I}_k$ represents the measurement (image) noise such that $\boldsymbol{Y}_k = \boldsymbol{I}_k + \delta\boldsymbol{I}_k$, it can be deduced from equations (15) and (16) that

$$\hat{\boldsymbol{I}}_{k+1} = \hat{\boldsymbol{I}}_{k+1|k} + \mathbf{1}(\boldsymbol{Y}_{k+1} - \hat{\boldsymbol{I}}_{k+1|k}) = \boldsymbol{I}_{k+1} + \delta\boldsymbol{I}_{k+1}. \tag{19}$$

In other words, $\bar{\boldsymbol{I}}_k = \delta\boldsymbol{I}_k$ or the estimate of $\boldsymbol{I}$ immediately converges to the true value in the absence of the measurement noise regardless of the prediction $\hat{I}_{k+1|k}$. Correspondingly, H_I in Fig. 2 is $\mathbf{1}_{N\times N}$.

2) Convergence of $\bar{\boldsymbol{n}}_k$: The dynamics of $\hat{\boldsymbol{n}}_{k+1}$ through the update and correction steps from equations (11)-(16) are

$$\begin{aligned}\hat{\boldsymbol{n}}_{k+1} = \hat{\boldsymbol{n}}_k &+ [\hat{\boldsymbol{n}}_k]_\times\boldsymbol{\omega}_k T \\ &- 2\Lambda_n\Gamma_n f\hat{\boldsymbol{A}}_{n,k}(\boldsymbol{Y}_{k+1} - \hat{\boldsymbol{I}}_{k+1|k}),\end{aligned} \tag{20}$$

where

$$\hat{\boldsymbol{A}}_{n,k} = \left\{\boldsymbol{p}\hat{\boldsymbol{\vartheta}}_k^T[\boldsymbol{p}]_\times[\nabla I_k]_\times\boldsymbol{e}_3\right\}_N. \tag{21}$$

The term $\boldsymbol{Y}_{k+1} - \hat{\boldsymbol{I}}_{k+1|k}$ in equation (20) can be written using the form provided by equation (12) and the fact that $\boldsymbol{e}_3^T[\nabla I]_\times[\boldsymbol{p}]_\times\boldsymbol{\vartheta}\boldsymbol{n}^T\boldsymbol{p} = (\boldsymbol{p}\boldsymbol{\vartheta}^T[\boldsymbol{p}]_\times[\nabla I]_\times\boldsymbol{e}_3)^T\boldsymbol{n}$, as a result,

$$\begin{aligned}\boldsymbol{Y}_{k+1} - \hat{\boldsymbol{I}}_{k+1|k} &= (\boldsymbol{I}_{k+1} - \boldsymbol{I}_k) - \left(\hat{\boldsymbol{I}}_{k+1|k} - \hat{\boldsymbol{I}}_k\right) + \Delta\boldsymbol{I}_k \\ &= -\frac{T}{f}\left\{\boldsymbol{p}\boldsymbol{\vartheta}_k^T[\boldsymbol{p}]_\times[\nabla I_k]_\times\boldsymbol{e}_3\right\}_N^T\boldsymbol{n}_k \\ &\quad + \frac{T}{f}\left\{\boldsymbol{p}\hat{\boldsymbol{\vartheta}}_k^T[\boldsymbol{p}]_\times[\nabla I_k]_\times\boldsymbol{e}_3\right\}_N^T\hat{\boldsymbol{n}}_k + \Delta\boldsymbol{I}_k,\end{aligned} \tag{22}$$

where $\Delta\boldsymbol{I}_k = \bar{\boldsymbol{I}}_{k+1} - \bar{\boldsymbol{I}}_k$. To deal with the unknown $\boldsymbol{\vartheta}_k$, we define a scalar quantity $\Psi_k(x, y)$ for each individual pixel at $\boldsymbol{p} = [x, y, f]^T$ as $\Psi_k(x, y) = \boldsymbol{\vartheta}_k^T[\boldsymbol{p}]_\times[\nabla I_k]_\times^T\boldsymbol{e}_3 / \hat{\boldsymbol{\vartheta}}_k^T[\boldsymbol{p}]_\times[\nabla I_k]_\times^T\boldsymbol{e}_3$. Subsequently,

$$\boldsymbol{Y}_{k+1} - \hat{\boldsymbol{I}}_{k+1|k} = -\frac{T}{f}\hat{\boldsymbol{A}}_{n,k}^T\left(\overset{N}{\Pi}\Psi_k\boldsymbol{n}_k - \hat{\boldsymbol{n}}_k\right) + \Delta\boldsymbol{I}_k, \tag{23}$$

where $\overset{N}{\Pi}\Psi_k$ is the product of $\Psi_k(x, y)$ for all N pixels (the nominal value of $\overset{N}{\Pi}\Psi_k$ when $\bar{\vartheta}_k \to 0$ is unity). With some manipulation, substitution of equation (23) into (20) produces

$$\begin{aligned}\hat{\boldsymbol{n}}_{k+1} = \hat{\boldsymbol{n}}_k &+ [\hat{\boldsymbol{n}}_k]_\times\boldsymbol{\omega}_k T - 2\Lambda_n\Gamma_n f A_{n,k}\Delta\boldsymbol{I}_k \\ &+ 2\Lambda_n\Gamma_n T\hat{\boldsymbol{A}}_{n,k}\hat{\boldsymbol{A}}_{n,k}^T\left(\overset{N}{\Pi}\Psi_k\boldsymbol{n}_k - \hat{\boldsymbol{n}}_k\right).\end{aligned} \tag{24}$$

Let $\boldsymbol{\eta}_k$ denote $\overset{N}{\Pi}\Psi_k\boldsymbol{n}_k - \hat{\boldsymbol{n}}_k$, we pre-multiply the discrete-time version of equation (11): $\boldsymbol{n}_{k+1} = \boldsymbol{n}_k + [\boldsymbol{n}_k]_\times\boldsymbol{\omega}T$ by $\overset{N}{\Pi}\Psi_k$. The difference between the outcome and equation (24) is

$$\boldsymbol{\eta}_{k+1} = \boldsymbol{\eta}_k - 2\Lambda_n\Gamma_n T\hat{\boldsymbol{A}}_{n,k}\hat{\boldsymbol{A}}_{n,k}^T\boldsymbol{\eta}_k + \boldsymbol{D}_{n,k}T + \boldsymbol{B}_{n,k}\Delta\boldsymbol{I}_k \tag{25}$$

where

$$\begin{aligned}\boldsymbol{D}_{n,k} &= \overset{N}{\Pi}\Psi_k \left[\boldsymbol{n}_k\right]_\times \boldsymbol{\omega}_k - \left[\hat{\boldsymbol{n}}_k\right]_\times \boldsymbol{\omega}_k,\\ \boldsymbol{B}_{n,k} &= 2\Lambda_n \Gamma_n f A_{n,k}.\end{aligned} \tag{26}$$

According to Lemma 1, the asymptotic stability of $\boldsymbol{\eta}_k$ is attained in the absence of $\Delta\boldsymbol{I}_k$ with suitable gains ($\Lambda_n \to \infty$ and $\Lambda_n\Gamma_n T \ll \mathbf{1}$) by treating $\boldsymbol{D}_{n,k}$ as bounded disturbances. The condition requires $\hat{\boldsymbol{A}}_{n,k}$ to be full row rank. This places some limitations on the camera movement ($\hat{\boldsymbol{\vartheta}}_k$) and image gradients ($\partial I_{xy}/\partial x$, $\partial I_{xy}/\partial y \neq 0$) as suggested by equation (21). The condition on camera motion is easily achieved on a flying robot in practice. Furthermore, this rank constraint is relaxed when the persistence of excitation is taken into account (see the Supplemental Materials and [32]). In addition, with the measurement noise ΔI_k, the system remains stable as long as $\boldsymbol{B}_{n,k}$ or $\Lambda_n\Gamma_n A_{n,k}$ remains bounded. The $\mathcal{L}$ gain of $\boldsymbol{B}_{n,k}$ influences the stability properties of the equivalent subsystem H_n in Fig. 2 as outlined in the Supplemental Materials. The stability of $\overset{\eta_k}{N}$ warrants the convergence of $\bar{\boldsymbol{n}}_k$ up to the scale factor $\overset{N}{\Pi}\Psi_k$. Then, $\hat{\boldsymbol{n}}_k$ is obtained through normalization after an update. In practice, it is adequate to perform the normalization every ~ 100 steps.

3) Convergence of $\bar{\boldsymbol{\vartheta}}_k$: The stability of $\bar{\boldsymbol{\vartheta}}_k$ is achieved through a similar framework to that of $\bar{\boldsymbol{n}}_k$ without the need for normalization. From equations (10) and (16), the estimation error evolves according to

$$\bar{\boldsymbol{\vartheta}}_{k+1} = \bar{\boldsymbol{\vartheta}}_k + 2\Lambda_\vartheta \Gamma_\vartheta f \hat{\boldsymbol{A}}_{\vartheta,k}\left(\boldsymbol{Y}_{k+1} - \hat{\boldsymbol{I}}_{k+1|k}\right) + \boldsymbol{D}_{\vartheta,k}T, \tag{27}$$

where

$$\begin{aligned}\boldsymbol{D}_{\vartheta,k} &= \mathbf{a}_k\alpha_k + \boldsymbol{\vartheta}_k^T\boldsymbol{n}_k\boldsymbol{\vartheta}_k + \left[\boldsymbol{\vartheta}_k\right]_\times \boldsymbol{\omega}_k\\ &\quad - \mathbf{a}_k\hat{\alpha}_k - \hat{\boldsymbol{\vartheta}}_k^T\hat{\boldsymbol{n}}_k\hat{\boldsymbol{\vartheta}}_k - \left[\hat{\boldsymbol{\vartheta}}_k\right]_\times \boldsymbol{\omega}_k, \quad \text{and}\\ \hat{\boldsymbol{A}}_{\vartheta,k} &= \left\{\boldsymbol{p}^T\hat{\boldsymbol{n}}_k[\boldsymbol{p}]_\times[\nabla I_k]_\times^T\boldsymbol{e}_3\right\}_N.\end{aligned} \tag{28}$$

The term $\boldsymbol{Y}_{k+1} - \hat{\boldsymbol{I}}_{k+1|k}$ in equation (27) can be treated in a similar manner to equation (22), but with the fact that $\boldsymbol{e}_3^T[\nabla I]_\times[\boldsymbol{p}]_\times\boldsymbol{\vartheta}\boldsymbol{n}^T\boldsymbol{p} = (\boldsymbol{p}^T\boldsymbol{n}[\boldsymbol{p}]_\times[\nabla I]_\times^T\boldsymbol{e}_3)^T\boldsymbol{\vartheta}$:

$$\begin{aligned}\boldsymbol{Y}_{k+1} - \hat{\boldsymbol{I}}_{k+1|k} &= -\frac{T}{f}\left\{\boldsymbol{p}^T\boldsymbol{n}_k[\boldsymbol{p}]_\times[\nabla I_k]_\times^T\boldsymbol{e}_3\right\}_N^T\boldsymbol{\vartheta}_k\\ &\quad + \frac{T}{f}\left\{\boldsymbol{p}^T\hat{\boldsymbol{n}}_k[\boldsymbol{p}]_\times[\nabla I_k]_\times^T\boldsymbol{e}_3\right\}_N^T\hat{\boldsymbol{\vartheta}}_k + \Delta\boldsymbol{I}_k\\ &= -\frac{T}{f}\hat{\boldsymbol{A}}_{\vartheta,k}^T\bar{\boldsymbol{\vartheta}}_k - \frac{T}{f}\boldsymbol{A}_{n,k}^T\bar{\boldsymbol{n}}_k + \Delta\boldsymbol{I}_k,\end{aligned} \tag{29}$$

Combining equations (27) and (29) yields

$$\begin{aligned}\bar{\boldsymbol{\vartheta}}_{k+1} &= \bar{\boldsymbol{\vartheta}}_k - 2\Lambda_\vartheta\Gamma_\vartheta T\hat{\boldsymbol{A}}_{\vartheta,k}\hat{\boldsymbol{A}}_{\vartheta,k}^T\bar{\boldsymbol{\vartheta}}_k + \boldsymbol{D}_{\vartheta,k}T\\ &\quad + \boldsymbol{B}_{\vartheta,k}\begin{bmatrix}\bar{\boldsymbol{n}}_k^T & \Delta\boldsymbol{I}_k^T\end{bmatrix}^T,\end{aligned} \tag{30}$$

where

$$\boldsymbol{B}_{\vartheta,k} = 2\Lambda_\vartheta\Gamma_\vartheta\begin{bmatrix}-T\hat{\boldsymbol{A}}_{\vartheta,k}\boldsymbol{A}_{n,k}^T & f\hat{\boldsymbol{A}}_{\vartheta,k}\end{bmatrix}. \tag{31}$$

Applying Lemma 1, the estimation error $\bar{\boldsymbol{\vartheta}}_k$ is reduced in the absence of disturbance ($\Delta\boldsymbol{I}_k, \bar{\boldsymbol{n}}_k \to \mathbf{0}$) as long as $\boldsymbol{A}_{\vartheta,k}$ is full rank and the gain conditions are met. That is the convergence of $\bar{\boldsymbol{\vartheta}}_k$ is obtained under the similar conditions to that of $\bar{\boldsymbol{n}}_k$, but by treating both $\Delta\boldsymbol{I}_k$ and $\bar{\boldsymbol{n}}_k$ as inputs of the equivalent subsystem H_ϑ as shown in Fig. 2 and the Supplemental Materials.

4) Convergence of $\bar{\alpha}_k$: Again, we start with the incremental update of the estimation error $\alpha_k - \hat{\alpha}_k$ derived from equations (9), (15), and (29):

$$\begin{aligned}\bar{\alpha}_{k+1} &= \bar{\alpha}_k - 2\Lambda_\alpha\Gamma_\alpha T\mathbf{a}_{k-1}^T\hat{\boldsymbol{A}}_{\vartheta,k}\hat{\boldsymbol{A}}_{\vartheta,k}^T\bar{\boldsymbol{\vartheta}}_k + D_{\alpha,k}T\\ &\quad - 2\Lambda_\alpha\Gamma_\alpha\mathbf{a}_{k-1}^T\hat{\boldsymbol{A}}_{\vartheta,k}\left(T\boldsymbol{A}_{n,k}^T\bar{\boldsymbol{n}}_k - f\Delta\boldsymbol{I}_k\right)\end{aligned} \tag{32}$$

where

$$D_{\alpha,k} = \boldsymbol{\vartheta}_k^T\boldsymbol{n}_k\alpha_k - \hat{\boldsymbol{\vartheta}}_k^T\hat{\boldsymbol{n}}_k\hat{\alpha}_k. \tag{33}$$

Exploiting equation (10), the term $\bar{\boldsymbol{\vartheta}}_k$ in equation (32) can be re-written:

$$\begin{aligned}\bar{\boldsymbol{\vartheta}}_k &= (\boldsymbol{\vartheta}_k - \boldsymbol{\vartheta}_{k-1}) - \left(\hat{\boldsymbol{\vartheta}}_k - \hat{\boldsymbol{\vartheta}}_{k-1}\right) + \bar{\boldsymbol{\vartheta}}_{k-1}\\ &= \mathbf{a}_{k-1}\bar{\alpha}_{k-1}T + \hat{\boldsymbol{\vartheta}}_k\hat{\boldsymbol{\vartheta}}_k^T\bar{\boldsymbol{n}}_kT\\ &\quad + \left(\hat{\boldsymbol{\vartheta}}_k^T\hat{\boldsymbol{n}}_k - \left[\boldsymbol{\omega}_k\right]_\times\right)\bar{\boldsymbol{\vartheta}}_kT + \bar{\boldsymbol{\vartheta}}_{k-1}\end{aligned} \tag{34}$$

Substituting this to equation (32) yields

$$\begin{aligned}\bar{\alpha}_{k+1} &= \bar{\alpha}_k - 2\Lambda_\alpha\Gamma_\alpha T^2\mathbf{a}_{k-1}^T\boldsymbol{A}_{\vartheta,k}\boldsymbol{A}_{\vartheta,k}^T\mathbf{a}_{k-1}\bar{\alpha}_{k-1}\\ &\quad + D_{\alpha,k}T + \boldsymbol{B}_{\alpha,k}\begin{bmatrix}\bar{\boldsymbol{n}}_k^T & \bar{\boldsymbol{\vartheta}}_{k-1}^T & \bar{\boldsymbol{\vartheta}}_k^T & \Delta\boldsymbol{I}_k^T\end{bmatrix}^T,\end{aligned} \tag{35}$$

where

$$\begin{aligned}\boldsymbol{B}_{\alpha,k} = &-2\Lambda_\alpha\Gamma_\alpha\mathbf{a}_{k-1}^T\hat{\boldsymbol{A}}_{\vartheta,k}\\ &\begin{bmatrix}T\left(\boldsymbol{A}_{n,k}^T + T\hat{\boldsymbol{A}}_{\vartheta,k}^T\hat{\boldsymbol{\vartheta}}_k\hat{\boldsymbol{\vartheta}}_k^T\right)^T\\ T\hat{\boldsymbol{A}}_{\vartheta,k}\\ T^2\left(\hat{\boldsymbol{\vartheta}}_k^T\hat{\boldsymbol{n}}_k - \left[\boldsymbol{\omega}_k\right]_\times\right)^T\hat{\boldsymbol{A}}_{\vartheta,k}\\ -f\end{bmatrix}^T.\end{aligned} \tag{36}$$

The stability of $\bar{\alpha}_k$ is then attained according to Lemma 2 as long as $2\Lambda_\alpha\Gamma_\alpha T^2\mathbf{a}_{k-1}^T\boldsymbol{A}_{\vartheta,k}\boldsymbol{A}_{\vartheta,k}^T\mathbf{a}_{k-1} \in \left(0, \frac{2}{3}\right)$, and $D_{\alpha,k}$ is bounded. The first condition requires the gains $\Lambda_\alpha\Gamma_\alpha$ to be sufficiently small, $\boldsymbol{A}_{\vartheta,k}$ to be full rank as in section III-B3, and the acceleration to be non-zero. The restriction on the motion is similar and related to the rank condition of $\boldsymbol{A}_{n,k}$ considered earlier. The boundedness of $D_{\alpha,k}$ is readily achieved in physically feasible flights. Furthermore, as $T \to 0$, the system is asymptotically stable, or the system remains stable in the presence of non-zero $\bar{\boldsymbol{n}}_k$, $\bar{\boldsymbol{\vartheta}}_k$, and $\Delta\boldsymbol{I}_k$ when they are regarded as bounded inputs to the subsystem H_α as long as $\Lambda_\alpha\Gamma_\alpha$ is finite.

In summary, the stability of the proposed observer is achieved for each quantity in the state vector in the order shown in Fig. 2. If present, estimation errors accumulate over each step but the output remains bounded. Under suitable gains, the system is globally asymptotically stable as $T \to 0$. This renders the NLO more robust against disturbances and initial errors compared to EKF-based strategies that suffer from inaccuracies introduced by the linearization and restricted region of attraction.

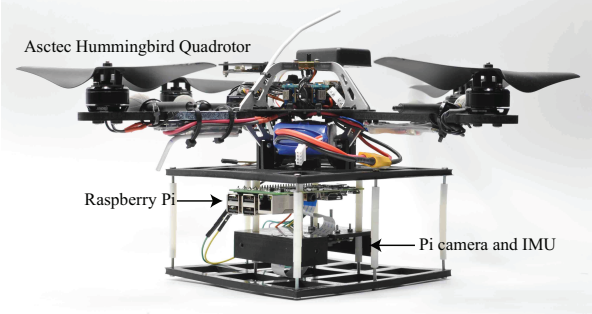


Fig. 3. A Hummingbird Quadrotor carrying the IMU, Raspberry Pi and camera for verification of the NLO. The IMU and camera are located next to each other such that their coordinate frames can be assumed coincident.

IV. Experimental Verification

In this section, we perform indoor flight experiments to validate the proposed nonlinear observer and compare the performance to that of an EKF and a state-of-the-art map-based approach. In-depth investigations on different flight patterns, textures, computational requirements, and the impact of the planar scene, are provided.

A. Experiment setup and implementation

For flight experiments, we employed a single-board computer (Raspberry Pi 3 Model B) and the Pi Camera v2 with the field of view of $24.4^\circ \times 19.0^\circ$ to record 640×480-px images at 90 Hz. An IMU (MPU9250) was used to provide the angular velocity ($\boldsymbol{\omega}$) and gravity integrated acceleration ($\mathbf{a}$) at the rate of 100 Hz, synchronized with the images using Python scripts. After the camera calibration, the camera-IMU unit was attached to an AscTec Hummingbird robot (Ascending Technologies) with the downlooking camera as shown in Fig. IV-A(a). The experiments were carried out in a $3.0 \times 3.0 \times 2.5$ m arena equipped with six motion capture cameras (OptiTrack Prime 13w) for tracking the position and orientation of the quadrotor for the ground-truth measurements of α, $\boldsymbol{\vartheta}$, and $\boldsymbol{n}$ and robot's position control as the estimation results were not used for feedback. All data are time aligned.

To validate the proposed estimator, experiments were carried out to demonstrate the estimation performance for various flight patterns, textures, plane inclinations, etc. Each flight contains 120 seconds of mid-air measurements suitable for the estimator. After the experiments, the proposed NLO was implemented in Python and applied to the collected images and IMU measurements for estimation of the state vector defined in equation (7) on the Raspberry Pi. The offline implementation allows results from different estimators to be directly compared. In the implementation, 640×480-px images were uniformly downsampled to 160×120-px images after the application of 5×5 averaging kernels. After that the spatial gradients ∇I in equation (12) were then obtained by applying an image convolution with normalized 3×3 Sobel kernels.

To provide benchmark comparisons, an EKF was employed to provide estimations of the state vector from the image luminosity measurements according to the nonlinear dynamics described by equations (7)-(13). Compared to the proposed NLO, EKF has two major drawbacks: (i) as a result of the linearization, the stability is no longer guaranteed and the convergence critically depends on the initial conditions; (ii) the evaluation of Kalman gain involves an inversion of an $N \times N$ matrix. This is impractically expensive even for downsampled images, rendering the method unsuitable for real-time operation, particular on platforms with limited power. In contrast, it can be seen that the complexity of the proposed method is dominated by simple matrix multiplications related to matrices of size $7 \times N$ and $N \times 1$ only.

B. Flight Experiments

1) Flights on horizontal plane: The first set of flight experiments were performed on a horizontal ground. We executed three flying patterns over four types of textures. Three flights were carried out for each combination of texture and flight pattern, producing 36 flights in total. Three hovering flights were performed at 0.4 m, 0.8 m, 1.2 m altitude for each texture, whereas for the other two flight patterns, all three flights for each texture are nominally identical.

As illustrated in Fig. 5, four textures chosen include checkerboard (CHB), ramp (RMP), sinusoid (SIN) and artificial leaves (LEAF). The first three textures were selected to emphasize the difference between feature-based and direct methods. The checkerboard, prevalent in computer vision owing to pronounced edges and corners for easy detection, features 6×6-cm tiles. The ramp and sinusoidal textures used are grayscale patterns with the spatial intensity varying according to 2D sinusoidal and ramp functions. The absence of apparent corners and edges from these textures is intended for evaluating one key benefit of the proposed gradient-based approach when compared to feature-based methods. The periods of the first three textures are 12 cm. For the leaf pattern, it mimics a more realistic scene to evaluate the performance of the NLO with less structured patterns.

Three flight patterns investigated here are hovering, vertical trajectory, and 3D circular trajectory. During hovering flights, the robot is anticipated to exhibit minimal movement, representing the scenario where the proposed estimator is near the unobservable condition. This is to demonstrate that slight oscillations and vibrations provide sufficient excitation for the estimator. For the vertical trajectory, the robot was commanded to follow a sinusoidal trajectory at the altitude between 0.45-0.95 m at 0.2 Hz. The altitude variation verifies the observer's ability to deal with changes in altitude and flow divergence in the absence of horizontal motion. For the 3D flight pattern, the robot followed a cyclic circular trajectory with the radius of 0.65 m at 0.2 Hz, spanning the approximate volume of $0.5 \times 0.5 \times 0.4$ m. The pattern, similar to the trajectory used in [22], verifies the effectiveness of the proposed method in practical situations in which the robot traverses in 3D space.

2) Flights on inclined plane: Since the proposed NLO is capable of estimating relative plane orientation ($\boldsymbol{n}$), six additional flights were performed on a 10° incline affixed with the CHB texture. The robot flew horizontally covering the distance of 1.5m at 0.8 m altitude in a cyclic manner at 0.2 Hz to assess the quality of the distance and other estimates when the robot traverses over a non-horizontal surface.

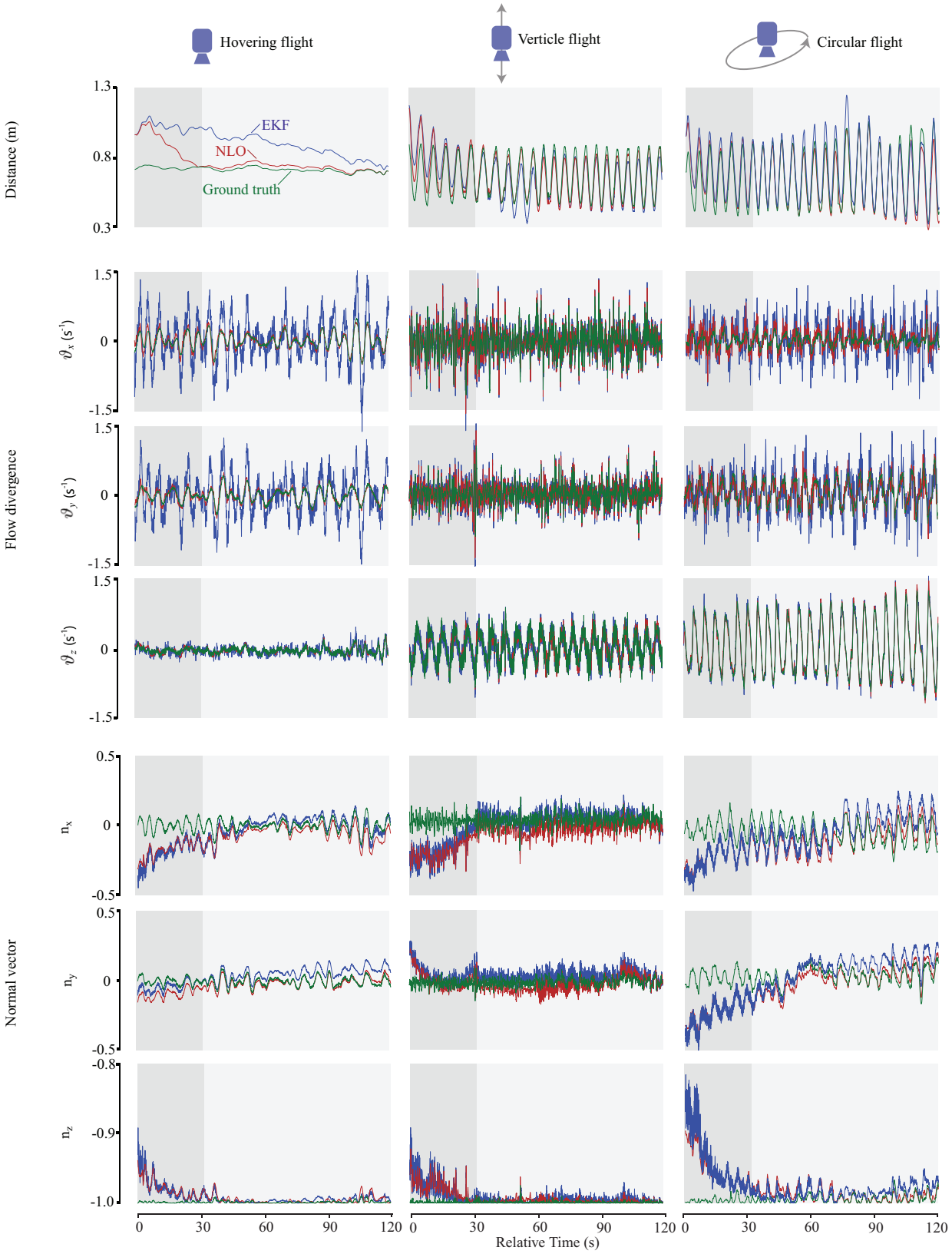


Fig. 4. Example estimation results corresponding to three flight patterns over SIN texture. Ground truth data are provided by the motion capture system. Both estimates from the NLO and EKF are plotted. The flight altitudes are in the range of 0.45-0.95 m, depending on the flight pattern. The darker region in each subplot indicates a 30-second period after the estimator initializes. The estimates are assumed to have converged thereafter. (Top) The plot shows the distance estimates and the ground truth. (Middle) The flow divergence in three directions. (Bottom) The estimate of the plane's normal with respect to the camera frame. A closed-up plot detailing the last five seconds of the vertical flight can be found in the Supplemental Materials.

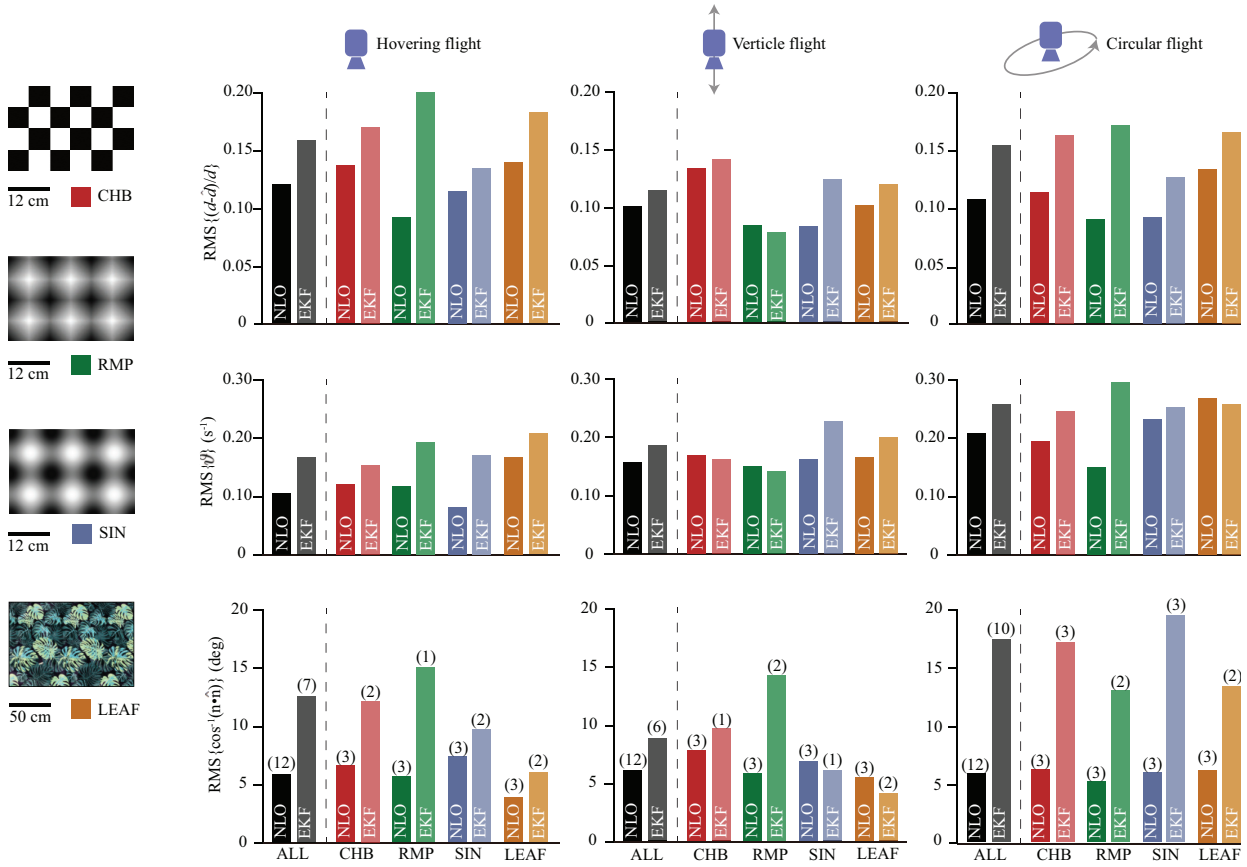


Fig. 5. Estimation errors from 36 flights with various flight patterns and ground textures. Three flights were conducted for each flight pattern and ground texture. The RMS errors from both NLO and EKF are computed from a 90-second duration, 30 s after the initialization of the filters for each dataset. As for the EKF, only 23 out of 36 datasets result in converged estimates that are eligible for the calculation of the RMS errors. The numbers of datasets with converged estimates for each combination of flight patterns and textures are labeled in brackets on top of each bars in the last row of the plots.

3) Flights over an occluded ground: The derivation of the NLO relies on the assumption of a planar scene. To assess the robustness and the performance of the NLO when such condition is violated, three additional flights were carried out over the CHB texture, partially occluded by a $25 \times 28 \times 30$-cm cardboard box. The robot flew vertically in a cyclic manner at the altitude 0.45 to 0.95 m at the frequency of 0.2 Hz.

C. Estimation Results

Fig. 4 shows representative results from three flight patterns over the SIN texture above a horizontal ground. The estimated state vectors are compared to ground truth measurements provided by the motion capture system to calculate the root mean square (RMS) errors. It can be seen that the NLO consistently produces better results than the EKF. For the distance estimates, the EKF displays larger steady-state errors and slower convergence speed. For the flow divergence, the EKF shows significant fluctuation while the NLO yields noticeably smoother results. For the normal vector, the steady-state errors from the EKF are also visibly larger. Among different flight patterns, hovering flights estimation yields a much slower convergence rate for the distance estimation.

1) Estimation results from flights over horizontal ground: In 13 out of the 36 flights, the EKF estimates diverge, failing to estimate the distance despite our best effort to adjust the noise covariances. In contrast, the NLO displays robustness with no divergence. The estimation errors are given in Fig. 5. The root-mean-square (RMS) errors are calculated from 90-s periods, 30 s after taking off to permit the estimates to converge. For the EKF, the RMS errors are computed from 23 flights with converged estimates, while all 36 flights are taken into consideration for the NLO. Overall, the RMS distance errors from the EKF for three flight patterns are 16.7, 11.7, and 15.3 cm, while for the NLO, the errors are 12.1, 10.2 and 10.7 cm. All data combined, the RMS error from the EKF is 28% larger than that of from the NLO.

For all four textures, Fig. 5 reveals that the NLO provides reliable estimates of the distance and flow divergence. Altogether, the RMS error of the estimated distance is approximately 10% at the 0.7 m average flight altitude and the RMS error of the flow divergence is 0.16 s^{-1}. The performance is comparable to that of the state-of-the-art feature-based method which yields the RMS distance error of 9.2 cm for a flight at 1 m altitude and the mean speed error of 0.1 ms^{-1} [22].

Among three flight patterns, the distance estimates from the NLO corresponding to vertical and circular flights show approximately 17% lower RMS errors than those from hovering flights. The observed trend is reversed for the estimates of the flow divergence, whereas the estimation errors of the normal vector from three flight patterns are nearly indistinguishable. This is likely because the quality of the distance

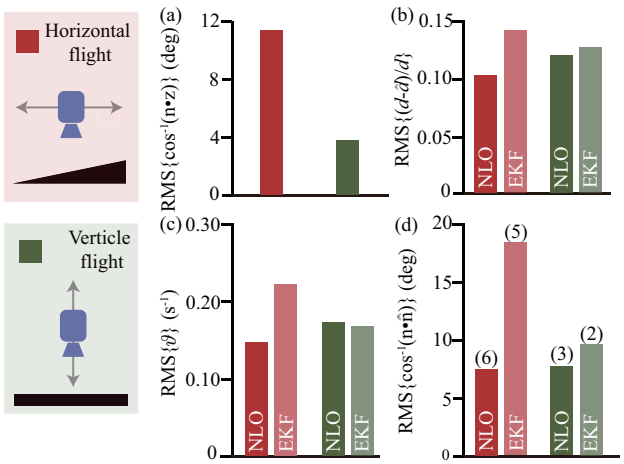


Fig. 6. Estimation results corresponding to horizontal flights on a tilted plane and vertical flights on a horizontal plane. (a) Mean angles between the camera-centric plane's normal and the optical axis ($\hat{z}$ in Fig. 1). These angles are nominally $10°$ and $0°$ in case the camera statically points downwards. (b) RMS errors of distance estimates. (c) RMS of ϑ. (d) RMS of angle between real normal vector and estimated normal vector.

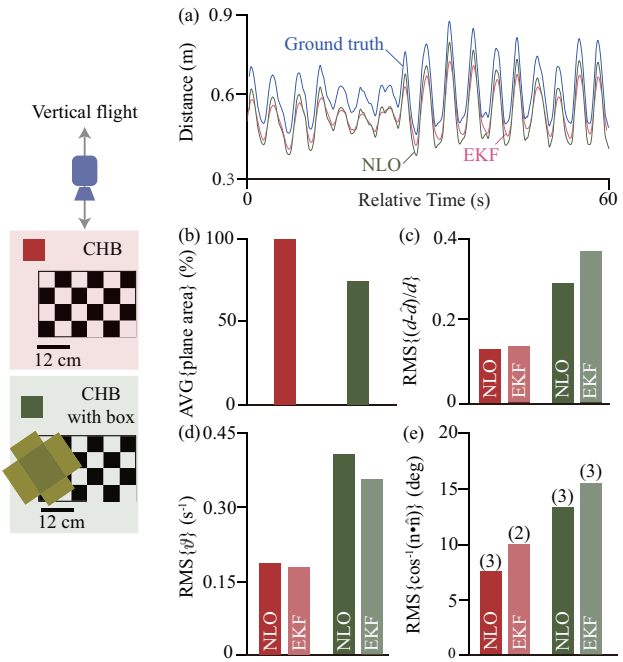


Fig. 7. Assessment of the estimation results when the when the single plane assumption is violated. (a) The altitude estimates from the NLO and EKF compared with the groundtruth. (b) Mean percentage of unoccluded region in the images in the benchmark dataset and occluded dataset. (c) RMS errors of the distance estimates. (d) RMS of the ϑ estimates. (e) The error between the actual and estimated plane's normal vector expressed in degree.

estimates benefits from motion or excitation. In contrast, the flow divergence is tightly and directly coupled with the measurements. Therefore, the estimates of the flow divergence are detrimentally affected by motion-induced disturbances.

2) Estimation results from flights over an incline: The results obtained from six flights above a $10°$ incline are shown in Fig. 6. During the horizontal translation, the camera perceived an alteration in the altitude, resembling the change induced by the vertical motion above a horizontal plane. For this reason, we benchmark the horizontal flight results against the outcomes belonging to vertical flights with the CHB from Section IV-B1. The major difference in two cases is highlighted in Fig. 6(a), which illustrates the RMS angle between the camera's optical axis and the true camera-centric plane's normal ($11.4°$ versus $3.4°$). The estimation results shown in Fig. 6(b)-(d) reveals that the errors from both scenarios are highly comparable. In case of the inclined surface, the NLO demonstrate no significant change in the error of the plane normal estimates, whereas the EKF evidently suffered with the RMS error of nearly $20°$. The results suggest that the NLO is more robust than the EKF when there exists an appreciable deviation between the camera axis and the plane's normal.

3) Estimation results from flights with occlusion: Fig. 7 shows the results when the one plane assumption is violated in comparison with the benchmark experiments. The benchmark results were taken from the previous experimental set. Fig. 7(a) shows the altitude estimates from the NLO and EKF , with respect to the groundtruth. It can be seen that that the estimates always underpredict the altitude. This is likely because the observers detect the combined distance to the plane and the occluding object. Fig. 7(b) highlights the average unoccluded area in terms of the percentage of the camera's field of view during three flights, demonstrating that $> 25\%$ of the view was obstructed by a box (refer to the Supplemental Materials for the method used for computing the segregation). The results in Fig. 7(c)-(e) reveal a noticeable degradation in the performance due to the provided reason. However, the proposed scheme was still able to provide converged estimates, demonstrating the robustness of the nonlinear approach.

D. Convergence rates among different flight patterns

According to section III-B, it has been concluded that the observability of the system lies in the non-zero values of $\boldsymbol{\vartheta}$ and the acceleration $\mathbf{a}$. In hovering flights, however, vibrations and fluctuations in velocity and acceleration turn out to provide sufficient excitation for the estimation. Among three flying patterns, both $\mathbf{a}$ and $\boldsymbol{\vartheta}$ associated with hovering flights are still smaller in magnitude than in circular and vertical flights. As a result, the distance converges at a slower rate as predicted by the persistency of excitation condition given in the Supplemental Materials.

E. Robustness to textures and image features

Unlike feature-based approaches that require prominent corners and edges, the NLO leverages the motion and non-zero image gradients, or ∇I (equation (5)), for the estimator to be observable. The NLO is anticipated to perform better with flights over RMP and SIN textures than over CHB texture owing to the absence of spatial gradients between edges in the CHB patttern. This is consistent with the results in Fig. 5, where the trend can be observed for all flight patterns.

The use of image gradients offers robustness when benchmarked against feature-based or semi-direct methods to scenes with scarce features. In evidence, we applied two popular feature detectors: FAST tracker [20] and Shi-Tomasi Corner Detector [19], on the obtained flight images. Both algorithms

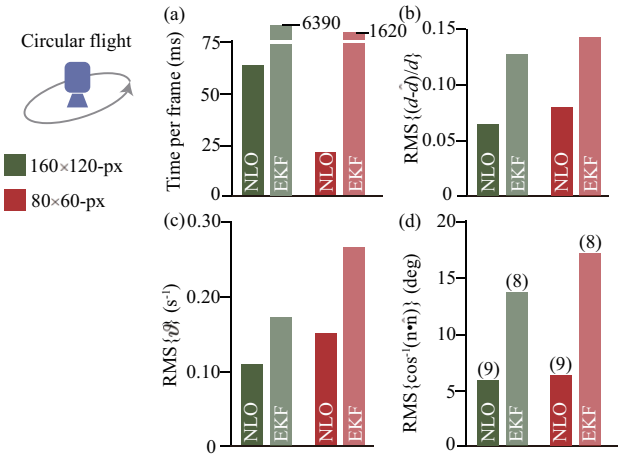


Fig. 8. Results from two image resolutions. (a) RMS of distance error. (b) RMS of ϑ error. (c) RMS of angle between real normal vector and estimated normal vector. (d) the mean time cost for each frame. Number of sets of data that converge for EKF are labelled on the top of the bars.

successfully identify multiple features in every image containing the CHB texture. Nevertheless, among images with RMP and SIN textures, the FAST detector fails to detect more than four corners—the amount minimally required for the Lucas-Kanade tracker [22], [33], in over 95% of the images (186,934/194,400). The Shi-Tomasi detector fails to detect more than four functional features in over 81% of the total images (138,132/194,400) (refer to the Supplemental Materials for example images and the description of functional features). Although it is possible to adjust some parameters to slightly enhance the detection in these images, the improvement does not generalize over wider scenarios. The unreliability of the feature detectors in these scenarios severely impair the performance of feature-based methods. The findings corroborates one primary advantage of the proposed direct approach over existing featured-based and semi-direct methods.

F. Impact of image resolutions on time costs and RMS errors

Intended for platforms with limited capability, low computational power is vital. Here, the time cost is inspected to assess the efficiency. Both Python scripts for NLO and EKF were implemented on a Raspberry Pi 3 Model B to process data collected from nine circular flights from Section IV-B1. Processing times were recorded. As presented in Fig. 8(a), for the 160×120-px images used to produce the results in Fig. 4-6, the average computing times per frame are 69 ms and 6390 ms for the NLO and EKF. The substantial reduction in the cost of the proposed NLO with respect to the EKF attests the need for a nonlinear observer as a Kalman-based estimator becomes inefficient when entire images are used.

Furthermore, to understand the trade-off between computational demand and accuracy, we repeated the comparison using the same orignal images that were uniformly downsampled to 80×60 px, instead of to 160×120 px, after applying 7×7 averaging kernels. As shown in Fig. 8(a), with the 4-fold decrease in image pixels, the time costs for both NLO and EKF reduce by a factor of 3 to 4. Nevertheless, the impact on the estimation errors is less pronounced as illustrated in Fig. 8(b)-(d). The decrease in image resolution magnifies the RMS errors of the distance and flow divergence by $\approx$20% and $\approx$35% and the influence on the normal vector is almost negligible. Depending on platform requirements, the outcomes suggest that low resolution images may provide sufficiently accurate estimates while demanding considerably less resources.

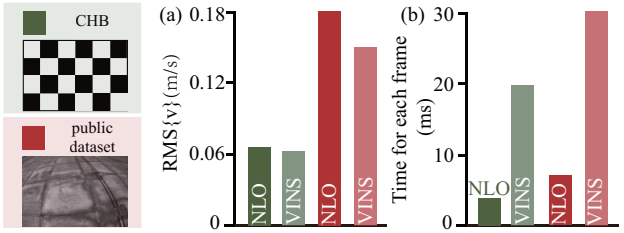


Fig. 9. Comparison between the NLO and VINS-mono using the CHB pattern and a public dataset from [34]. (a) RMS errors of the velocity estimates. (b) average processing time per frame.

G. Comparison to State-of-the-Art

To compare the performance of the proposed NLO against a state-of-the-art approach, we selected the flight data with the CJB pattern and a published dataset from [34] (with the top flight speed of 5 m/s). Both sets of flight data were pipelined to the NLO and the map-based VINS—VINS-mono [6]. Due to the nature of map-based VINS, the estimate of the flight altitude or distance to a flat terrain is not directly available for comparison. Instead, we resort to using the estimates of robo-centric flight velocity. For the NLO, flight velocity is a product of the flow divergence ϑ and the distance d.

As presented in Fig.9, for flight data with the checkerboard pattern, the RMS in velocity errors from the NLO and VINS-mono are highly comparable: 0.07 m/s and 0.06 m/s. For the published datasets, the flight speed reaches up to 5 m/s, rendering errors of the estimated velocity to be visibly larger. The RMS errors from the NLO and VINS-mono remain similar at 0.18 m/s and 0.15 m/s . In addition, for the CHB pattern, the average computational times per frame are 3.7 ms and 19.3 ms for the NLO and VINS-mono (both implemented on a laptop with Intel i5-4440 processor). For the dataset from [34], the average computing times are 6.7 ms and 31.3 ms. It can be seen that the omission of mapping and the use of the planar scene assumption renders the NLO is highly efficient and suitable for platforms with limited computational power.

V. Conclusion and Future Work

In this article, we have proposed a robust one-step ego-motion estimation strategy that integrate monocular vision and IMU measurements for feedback. When applied to aerial robots, the framework is capable of estimating the flight altitude and velocity for reactive navigation. The novelty lies in the use of image intensities as part of the state and output vectors. This eliminates the feature detection and tracking process, allowing the entire estimation to be achieved in a single step. As verified by several flight experiments, the featureless approach is robust to low-textured scenes and the proposed nonlinear observer offers superior efficiency when compared to an EKF-based implementation.

Similar to existing ego-motion estimation strategies for reactive navigation, the proposed framework is still limited for real-world deployment owing to the single-plane assumption. Possible extensions of this work include the consideration of multiple planes in the camera view. This would allow a flying robot to negotiate a corridor while maintaining safe distances to the ground and surrounding walls at the same time.

REFERENCES

[1] D. Floreano and R. J. Wood, "Science, technology and the future of small autonomous drones," *Nature*, vol. 521, no. 7553, p. 460, 2015.

[2] N. Jacobstein, "Autonomous vehicles: An imperfect path to saving millions of lives," *Science Robotics*, vol. 4, no. 28, 2019.

[3] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, 2015.

[4] M. Bloesch, M. Burri, S. Omari, M. Hutter, and R. Siegwart, "Iterated extended kalman filter based visual-inertial odometry using direct photometric feedback," *The International Journal of Robotics Research*, vol. 36, no. 10, pp. 1053–1072, 2017.

[5] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "Svo: Semidirect visual odometry for monocular and multicamera systems," *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 249–265, 2017.

[6] 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.

[7] Z. Huai and G. Huang, "Robocentric visual-inertial odometry," *The International Journal of Robotics Research*, vol. 0, no. 0, p. 0278364919853361, 2019.

[8] S. Weiss, M. W. Achtelik, S. Lynen, M. Chli, and R. Siegwart, "Real-time onboard visual-inertial state estimation and self-calibration of mavs in unknown environments," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 957–964.

[9] N. Mayer, E. Ilg, P. Fischer, C. Hazirbas, D. Cremers, A. Dosovitskiy, and T. Brox, "What makes good synthetic training data for learning disparity and optical flow estimation?" *International Journal of Computer Vision*, vol. 126, no. 9, pp. 942–960, 2018.

[10] K. McGuire, C. De Wagter, K. Tuyls, H. Kappen, and G. de Croon, "Minimal navigation solution for a swarm of tiny flying robots to explore an unknown environment," *Science Robotics*, vol. 4, no. 35, p. eaaw9710, 2019.

[11] Y. Chen, H. Zhao, J. Mao, P. Chirarattananon, E. F. Helbling, N.-s. P. Hyun, D. R. Clarke, and R. J. Wood, "Controlled flight of a microrobot powered by soft artificial muscles," *Nature*, vol. 575, no. 7782, pp. 324–329, 2019.

[12] Y. H. Hsiao and P. Chirarattananon, "Ceiling effects for hybrid aerial-surface locomotion of small rotorcraft," *IEEE/ASME Transactions on Mechatronics*, vol. 24, no. 5, pp. 2316–2327, Oct 2019.

[13] D. Scaramuzza, A. M. Lopez, A. Imiya, and T. Pajdla, "Application challenges from a bird's eye view," 2017.

[14] L. A. Schwarz, A. Mkhitaryan, D. Mateus, and N. Navab, "Human skeleton tracking from depth data using geodesic distances and optical flow," *Image and Vision Computing*, vol. 30, no. 3, pp. 217–226, 2012.

[15] D. Honegger, L. Meier, P. Tanskanen, and M. Pollefeys, "An open source and open hardware embedded metric optical flow cmos camera for indoor and outdoor applications," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1736–1741.

[16] H. W. Ho, G. C. de Croon, and Q. C. Chu, "Distance and velocity estimation using optical flow from a monocular camera," *International Journal of Micro Air Vehicles*, p. 1756829317695566, 2017.

[17] H. Wang, D. Zheng, J. Wang, W. Chen, and J. Yuan, "Ego-motion estimation of a quadrotor based on nonlinear observer," *IEEE/ASME Transactions on Mechatronics*, vol. 23, no. 3, pp. 1138–1147, 2018.

[18] M.-D. Hua, N. Manerikar, T. Hamel, and C. Samson, "Attitude, linear velocity and depth estimation of a camera observing a planar target using continuous homography and inertial data," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1429–1435.

[19] J. Shi *et al.*, "Good features to track," in *1994 Proceedings of IEEE conference on computer vision and pattern recognition*. IEEE, 1994, pp. 593–600.

[20] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European conference on computer vision*. Springer, 2006, pp. 430–443.

[21] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *International journal of computer vision*, vol. 56, no. 3, pp. 221–255, 2004.

[22] V. Grabe, H. H. Bülthoff, D. Scaramuzza, and P. R. Giordano, "Nonlinear ego-motion estimation from optical flow for online control of a quadrotor uav," *The International Journal of Robotics Research*, vol. 34, no. 8, pp. 1114–1135, 2015.

[23] S. Weiss, M. W. Achtelik, S. Lynen, M. C. Achtelik, L. Kneip, M. Chli, and R. Siegwart, "Monocular vision for long-term micro aerial vehicle state estimation: A compendium," *Journal of Field Robotics*, vol. 30, no. 5, pp. 803–831, 2013.

[24] H. Ho, G. de Croon, E. van Kampen, Q. Chu, and M. Mulder, "Adaptive gain control strategy for constant optical flow divergence landing," *IEEE Transactions on Robotics*, vol. 34, no. 2, pp. 508–516, 2018.

[25] B. K. Horn, Y. Fang, and I. Masaki, "Time to contact relative to a planar surface," in *Intelligent Vehicles Symposium, 2007 IEEE*. IEEE, 2007, pp. 68–74.

[26] L. Wang and B. K. Horn, "Time-to-contact control for safety and reliability of self-driving cars," in *2017 International Smart Cities Conference (ISC2)*. IEEE, 2017, pp. 1–4.

[27] H. D. Escobar-Alvarez, M. Ohradzansky, J. Keshavan, B. N. Ranganathan, and J. S. Humbert, "Bioinspired approaches for autonomous small-object detection and avoidance," *IEEE Transactions on Robotics*, vol. 35, no. 5, pp. 1220–1232, 2019.

[28] 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.

[29] H. Zhang and J. Zhao, "Bio-inspired vision based robot control using featureless estimations of time-to-contact," *Bioinspiration & biomimetics*, vol. 12, no. 2, p. 025001, 2017.

[30] P. Chirarattananon, "A direct optic flow-based strategy for inverse flight altitude estimation with monocular vision and imu measurements," *Bioinspiration & biomimetics*, vol. 13, no. 3, p. 036004, 2018.

[31] Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry, *An invitation to 3-d vision: from images to geometric models*. Springer Science & Business Media, 2012, vol. 26.

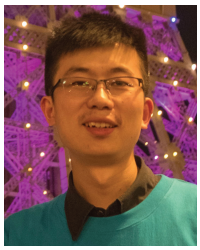
[32] P. R. Giordano, A. De Luca, and G. Oriolo, "3d structure identification from image moments," in *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*. IEEE, 2008, pp. 93–100.

[33] B. D. Lucas, T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision," 1981.

[34] K. Sun, K. Mohta, B. Pfrommer, M. Watterson, S. Liu, Y. Mulgaonkar, C. J. Taylor, and V. Kumar, "Robust stereo visual inertial odometry for fast autonomous flight," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 965–972, 2018.

[35] R. Spica and P. R. Giordano, "A framework for active estimation: Application to structure from motion," in *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*. IEEE, 2013, pp. 7647–7653.

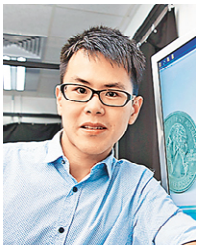
[36] H. K. Khalil, *Nonlinear systems*, 3rd ed. Upper Saddle River, N.J.: Prentice Hall, 2002.



Shixin Tan received a B.S. degree in Electronic Engineering from Fudan University, Shanghai, China, in 2013, and an M.S. degree in Electrical Engineering from University of Southern California, Los Angeles, CA, USA, in 2015. He is currently pursuing a Ph.D. degree with the Department of Biomedical Engineering, City University of Hong Kong. His research interests include micro air vehicles, vision based estimation and control.

Shangkun Zhong received the B.Eng. and M.S. from Harbin Institute of Technology (HIT), Harbin, China, in 2015 and 2017 respectively. He is currently working toward the PhD degree in robotics at City University of Hong Kong, Hong Kong SAR, China.

His interest includes computer vision, aerial vehicles, and robot control.



Pakpong Chirarattananon (S'12-M'15) received the B.A. degree in Natural Sciences from the University of Cambridge, U.K., in 2009, and the Ph.D. degree in Engineering Sciences from Harvard University, Cambridge, MA, USA, in 2014.

He is currently an Associate Professor with the Department of Biomedical Engineering, City University of Hong Kong, Hong Kong, China. His research interests include bio-inspired robots, micro air vehicles, and the applications of control and dynamics in robotic systems.

SUPPLEMENTAL MATERIALS FOR ONE-STEP VISUAL-INERTIAL EGO-MOTION ESTIMATIONUSING PHOTOMETRIC FEEDBACK

Shixin Tan, Shangkun Zhong, and Pakpong Chirarattananon

Department of Biomedical Engineering,
City University of Hong Kong,
Tat Chee Avenue, Hong Kong SAR, China.

A. Convergence Proofs

1) Proof of emma 1: Motivated by the method developed in [35], a radially unbounded discrete-time storage function can be created as:

$$L_k = \boldsymbol{Z}_k^T \Lambda^{-1} \boldsymbol{Z}_k > 0 \quad \forall \boldsymbol{Z}_k \neq \{\boldsymbol{0}\}. \tag{37}$$

It follows that, in the absence of the input $\boldsymbol{U}_k$,

$$\begin{aligned} L_{k+1} - L_k &= \boldsymbol{Z}_{k+1}^T \Lambda^{-1} \boldsymbol{Z}_{k+1} - \boldsymbol{Z}_k^T \Lambda^{-1} \boldsymbol{Z}_k \\ &= -4T \boldsymbol{Z}_k^T \boldsymbol{A}_k^T \boldsymbol{A}_k \Gamma \left(\boldsymbol{1} - T\Lambda\Gamma \boldsymbol{A}_k^T \boldsymbol{A}_k \right) \boldsymbol{Z}_k \\ &\quad + 2T \boldsymbol{Z}_k^T \Lambda^{-1} \left(\boldsymbol{1} - 2T\Lambda\Gamma \boldsymbol{A}_k^T \boldsymbol{A}_k \right) \boldsymbol{D}_k \\ &\quad + T^2 \boldsymbol{D}_k^T \Lambda^{-1} \boldsymbol{D}_k. \end{aligned} \tag{38}$$

Under the conditions (i) the disturbance $\boldsymbol{D}_k$ is bounded; (ii) the gain $\Lambda \to \infty$; and (iii) $T\Gamma \to 0$ or $T\Lambda\Gamma$ remains sufficiently small, the terms with $\boldsymbol{D}_k$ can be neglected. The criterion required for the global asymptotic stability of the system is $\boldsymbol{A}_k^T \boldsymbol{A}_k \Gamma \left(\boldsymbol{1} - T\Lambda\Gamma \boldsymbol{A}_k^T \boldsymbol{A}_k \right)$ must be positive definite. This is readily satisfied when the mentioned conditions are met and $\boldsymbol{A}_k$ is full rank. Moreover, with the consideration of the persistency of excitation [32], the requirement on the positive definiteness of $\boldsymbol{A}_k^T \boldsymbol{A}_k$ is relaxed to the existence of two positive numbers K and γ_A that renders the following condition satisfied:

$$\sum_{k}^{k+K} \boldsymbol{A}_k^T \boldsymbol{A}_k \Gamma \left(\boldsymbol{1} - T\Lambda\Gamma \boldsymbol{A}_k^T \boldsymbol{A}_k \right) \geq \gamma_A K \boldsymbol{1} > 0, \quad \forall k \geq 0. \tag{39}$$

In the presence of the input $\boldsymbol{U}_k$, we can further consider the system under the notion of input-output stability [36]. Based on the defined storage function L_k, as the sample time is sufficiently small or $T \to 0$, there exists a Lyapunov function $V(t, \boldsymbol{Z}, \boldsymbol{U} = 0)$ that satisfies $V = \boldsymbol{Z}^T \Lambda^{-1} \boldsymbol{Z} \geq 0$ and $\dot{V} \leq -4\gamma_A \boldsymbol{Z}^T \boldsymbol{Z}$. According to the theorem on $\mathcal{L}$ stability of of state models in [36], as $\dot{\boldsymbol{Z}} - \dot{\boldsymbol{Z}}|_{U=0} = \boldsymbol{BU}$, if there exists a non-negative constant γ_u such that $\|\boldsymbol{BU}\| \leq \gamma_u \|\boldsymbol{U}\|$,then the output $\boldsymbol{Z}$ satisfies

$$\|\boldsymbol{Z}\|_{\mathcal{L}_p} \leq \gamma \|\boldsymbol{U}\|_{\mathcal{L}_p} + \beta, \tag{40}$$

for some positive γ and β that can be determined [36]. That is, the system is finite-gain $\mathcal{L}_p$ stable for each $p \in [1, \infty]$. The condition stated by equation (40) is equivalent to the existstence of an $\mathcal{L}_p$ stable operator $\boldsymbol{H}$ that assigns each input signal $\boldsymbol{U}$ to the corresponding output $\boldsymbol{Z} = H(\boldsymbol{U})$.

2) Proof of Lemma 2: Define $\boldsymbol{Z}_k = \begin{bmatrix} Z_{k+1} & Z_k \end{bmatrix}^T$, the system described by equation (18) is equivalent to

$$\boldsymbol{Z}_{k+1} = \begin{bmatrix} 1 & -A_{k+1} \\ 1 & 0 \end{bmatrix} \boldsymbol{Z}_k + \begin{bmatrix} D_{k+1} \\ 0 \end{bmatrix} T + \begin{bmatrix} \boldsymbol{B}_{k+1} \boldsymbol{U}_{k+1} \\ 0 \end{bmatrix}. \tag{41}$$

Consider a storage function

$$L_k = \boldsymbol{Z}_k^T \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \boldsymbol{Z}_k \geq 0, \tag{42}$$

In the absence of the input $\boldsymbol{U}_k$, it follows that

$$L_{k+1} - L_k = \boldsymbol{Z}_k^T \boldsymbol{E}_2 \boldsymbol{Z}_k + \boldsymbol{E}_1 \boldsymbol{Z}_k T + E_0 T^2, \tag{43}$$

where

$$\begin{aligned} \boldsymbol{E}_2 &= \frac{1}{2} \begin{bmatrix} -1 & 1 - A_{k+1} \\ 1 - A_{k+1} & 2A_{k+1}^2 - 1 \end{bmatrix} \\ \boldsymbol{E}_1 &= \begin{bmatrix} D_{k+1} \\ -2A_{k+1} D_{k+1} \end{bmatrix}^T \\ E_0 &= D_{k+1}^2. \end{aligned} \tag{44}$$

Given that $A_{k+1} \in \left(0, \frac{2}{3}\right)$ and $|D_k| < D_+$, it can be shown that $\boldsymbol{E}_1 \boldsymbol{Z}_k < D_+ (Z_{k+1} + 2A_{k+1} Z_k) < D_+ (Z_{k+1} + \frac{4}{3} Z_k) \leq \frac{5}{3} D_+ |\boldsymbol{Z}_k|$ and $E_0 < D_+^2$. In addition, under these conditions, $\boldsymbol{E}_2$ is negative definite. Let λ_E denote the largest (negative) eigenvalue of $\boldsymbol{E}_2$: $\lambda_E = \frac{1}{2}(A_{k+1}^2 - 1 + \sqrt{A_{k+1}^4 + A_{k+1}^2 - 2A_{k+1} + 1})$, it can be verified that $\lambda_E \in (-0.1096, 0)$. Together, we obtain

$$\begin{aligned} L_{k+1} - L_k &< \lambda_E |\boldsymbol{Z}_k|^2 + \frac{5}{3} D_+ |\boldsymbol{Z}_k| T + D_+^2 T^2 \\ &< \lambda_E (|\boldsymbol{Z}_k| + \frac{5}{6\lambda_E} D_+ T)^2 \\ &\quad + \frac{1}{36\lambda_E} (36\lambda_E - 25) D_+^2 T^2. \end{aligned} \tag{45}$$

As a result, $L_{k+1} - L_k < 0$ as long as $|\boldsymbol{Z}_k| > -\frac{1}{6\lambda_E}(5 + \sqrt{25 - 36\lambda_E}) D_+ T$, or $\boldsymbol{Z}_k$ is stable in the sense of Lyapunov [36]. Moreover, in the limit $T \to 0$, the system described by equation (41) is globally asymptotically stable as $L_{k+1} - L_k < 0$.

In the continuous-time limit ($T \to 0$), we consider the Lyapunov function candidate analogous to the discrete-time storage function: $V(t, \boldsymbol{Z}, \boldsymbol{U} = 0) = \boldsymbol{Z}^T \Lambda^{-1} \boldsymbol{Z}$. This Lyapunov function candidate satisfies the conditions $V \geq 0$ and $\dot{V} \leq -\lambda_E \boldsymbol{Z}^T \boldsymbol{Z}$. The input-output stability of the system when $\boldsymbol{U} \neq 0$ can be analyzed similar to the proof of lemma 1 through the framework in [36]. Using the fact that $\left\| \dot{\boldsymbol{Z}} - \dot{\boldsymbol{Z}}|_{U=0} \right\| = \|\boldsymbol{BU}\|$, if there exists a non-negative constant γ_u such that $\|\boldsymbol{BU}\| \leq \gamma_u \|\boldsymbol{U}\|$,then the output $\boldsymbol{Z}$ satisfies

$$\|\boldsymbol{Z}\|_{\mathcal{L}_p} \leq \gamma \|\boldsymbol{U}\|_{\mathcal{L}_p} + \beta, \tag{46}$$

for some positive γ and β that can be determined [36]. That is, the system is finite-gain $\mathcal{L}_p$ stable for each $p \in [1, \infty]$. The condition stated by equation (46) is equivalent to the existstence of an $\mathcal{L}_p$ stable operator $\boldsymbol{H}$ that assigns each input signal $\boldsymbol{U}$ to the corresponding output $\boldsymbol{Z} = H(\boldsymbol{U})$.

B. Conventional Feature Detectors

Fig. 10 shows three different images taken during a flight over a SIN chosen to illustrate the shortcomings of the conventional feature detectors. The combination of poor image resolution, motion blur, and the lack of sharp edges make these images devoid of evident corners. In the majority of cases, both FAST [20] and Shi-Tomasi [19] detectors classify only point adjacent to the edges as features with highest qualities as demonstrated in Fig. 10(a). It can be seen that these detected features are unsuitable (not functional) for tracking purposes as they will likely be out of frame in subsequent images due to the motion. In addition, they do not correspond to actual features, rendering them susceptible to poor tracking performance when used with the Lucas- Kanade algorithm [33]. Fig. 10(b) shows a rare ($< 5\%$) example in cases which FAST identifies four or more functional features. Similarly, Fig. 10(c) provides another example in which Shi-Tomasi detector returns four functional points located at some distance from edges. This is only achieved in less than 20% of images with SIN or RMP textures. These examples highlight possible advantages of the featureless methods. While the outcomes shown here do not necessarily reflect real-world scenarios, where there are likely more prominent features visible in images, these example images are reasonable representatives of images from low-cost cameras or images degraded by motion blur.

C. Closed-up Estimation Results

In order to clearly differentiate the estimation performance of the NLO and EKF in Fig. 4, Fig. 11 provides closed-up plots of the flight data in Fig. 4, focusing on the last 30 seconds. It can be seen that the distance estimation from the NLO is superior to that of EKF. Other plots suggest that that the NLO produces smoother estimates of the flow divergence with less chattering compared to the EKF.

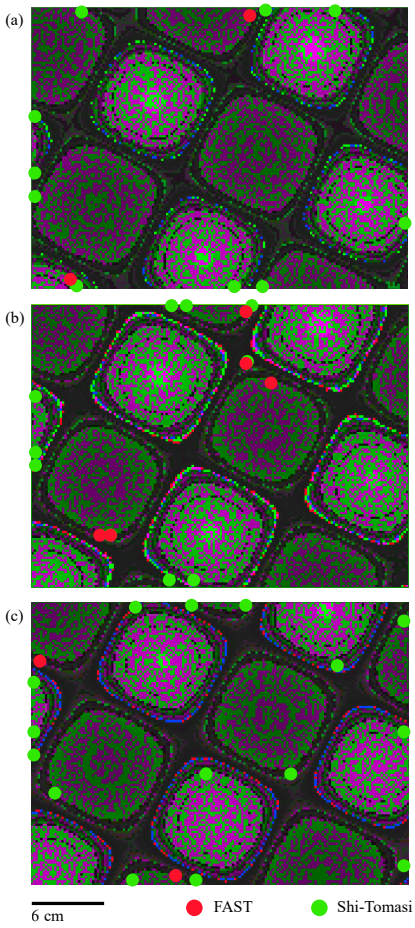


Fig. 10. Result of Shi-tomasi corner detector and FAST corner detector. (a) a representative case when both detectors find no functional feature away from the image edges. (b) another case that FAST corner detector detects feasible features however for Shi-tomasi corner detector, most of the features are on the edge of the image. (c) shows another case when FAST method only captures two features but Shi-tomasi corner detector works out fine. Each figure has a resolution of 160×120.

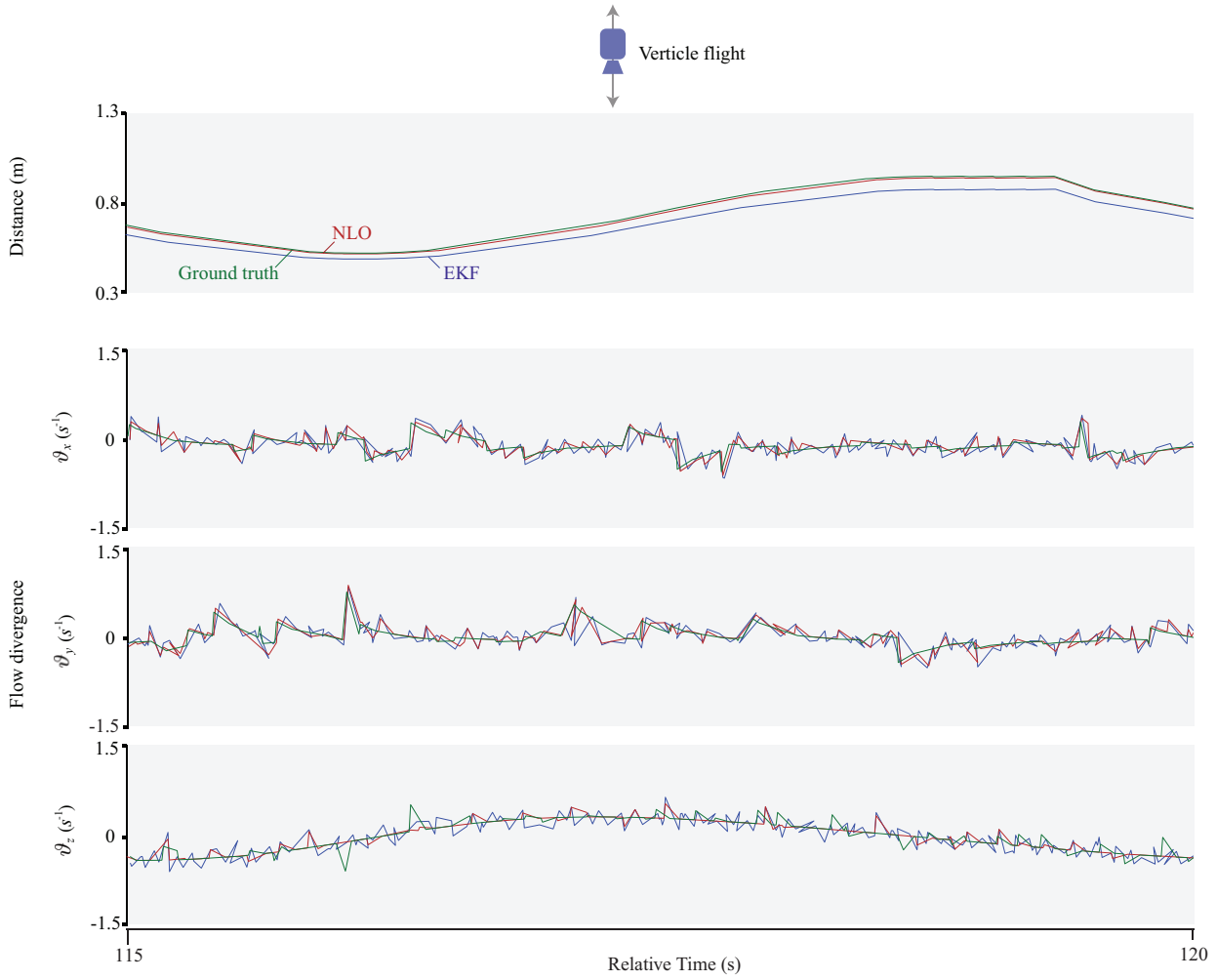


Fig. 11. Closed up plots of the estimation results (vertical flight) from Fig. 4.

D. Calculation of Occluded Region

To compute the percentage of an occluded area in the image, we post-processed the images by the application of the Hough Transform to the entire images. This results in both the boundaries of the boxes and the checkerboard pattern. To segment the box boundaries, a thresholding process was applied to only retain the shape of the box by considering the color of nearby pixels. Only the edges belonging to the edges of the brown cardboard box were kept. Thereafter, is straightforward to calculate the percentage of the occluded region in the image. Example results from each step are shown in Fig. 12.

E. Derivation of the observer gain G_k

Herein, we provide some explanations on the form of G_k to provide readers some intuition on its origin. The outline here applies to the convergence of $\bar{\boldsymbol{n}}_k$, but the convergence of other estimates follow the same strategy in the gain design.

We focus on the gain used for updating $\hat{\boldsymbol{n}}_k$, the third row of $\boldsymbol{G}_k$ or $\boldsymbol{G}_{k,3}$. To begin, observe that the time evolution of the photometric values from equation (12) is

$$\boldsymbol{I}_{k+1} - \boldsymbol{I}_k = -(1/f)\left\{(\boldsymbol{e}_3^T[\nabla I_k]_\times[\boldsymbol{p}]_\times \vartheta_k \boldsymbol{n}_k^T \boldsymbol{p})^T\right\}_N^T,$$

Recall that the operator $\{\mathbf{z}\}_N$ stands for a horizontal stacking operation of $\mathbf{z}_i \in R^3$ such that $\{\mathbf{z}\}_N = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N] \in \mathbb{R}^{3\times N}$. It can be seen that $\boldsymbol{I}_{k+1} - \boldsymbol{I}_k$ is directly related to $\boldsymbol{n}_k$. To make use of this, the portion $\left\{(\boldsymbol{e}_3^T[\nabla I_k]_\times[\boldsymbol{p}]_\times \vartheta_k \boldsymbol{n}_k^T \boldsymbol{p})^T\right\}_N^T$ is re-arragned to $\left\{\boldsymbol{p}\vartheta_k^T[\boldsymbol{p}]_\times[\nabla I_k]_\times \boldsymbol{e}_3\right\}_N^T \boldsymbol{n}_k$. Hence, the innovation term $\left(\boldsymbol{Y}_{k+1} - C\hat{\boldsymbol{X}}_{k+1|k}\right)$ is approximately proportional to

$$\left(\boldsymbol{Y}_{k+1} - C\hat{\boldsymbol{X}}_{k+1|k}\right) \approx \left\{\boldsymbol{p}\vartheta_k^T[\boldsymbol{p}]_\times[\nabla I_k]_\times \boldsymbol{e}_3\right\}_N^T \bar{\boldsymbol{n}}_k.$$

The intention is to ultimately obtain $\bar{\boldsymbol{n}}_{k+1} \propto -\bar{\boldsymbol{n}}_{k+1}$. This is approximately achieved by designing the third row of the observer's gain $\boldsymbol{G}_{k,3}$ to be $\approx -\left\{\boldsymbol{p}\vartheta_k^T[\boldsymbol{p}]_\times[\nabla I_k]_\times \boldsymbol{e}_3\right\}_N$, such that the condition

$$\boldsymbol{G}_{k,3}\left(\boldsymbol{Y}_{k+1} - C\hat{\boldsymbol{X}}_{k+1|k}\right) \approx -\left\{\boldsymbol{p}\vartheta_k^T[\boldsymbol{p}]_\times[\nabla I_k]_\times \boldsymbol{e}_3\right\}_N \left\{\boldsymbol{p}\vartheta_k^T[\boldsymbol{p}]_\times[\nabla I_k]_\times \boldsymbol{e}_3\right\}_N^T \bar{\boldsymbol{n}}_k$$

or $\boldsymbol{G}_{k,3}\left(\boldsymbol{Y}_{k+1} - C\hat{\boldsymbol{X}}_{k+1|k}\right) \propto -\bar{\boldsymbol{n}}_k$ is approximately met as $\left\{\boldsymbol{p}\vartheta_k^T[\boldsymbol{p}]_\times[\nabla I_k]_\times \boldsymbol{e}_3\right\}_N \left\{\boldsymbol{p}\vartheta_k^T[\boldsymbol{p}]_\times[\nabla I_k]_\times \boldsymbol{e}_3\right\}_N^T$ is positive definite. In the actual proof and implementation, there are extra terms emerging from the image noise and the fact that only the estimate of ϑ_k is available. However, they can be taken care of as detailed in the proof and the input-output stability is achieved.

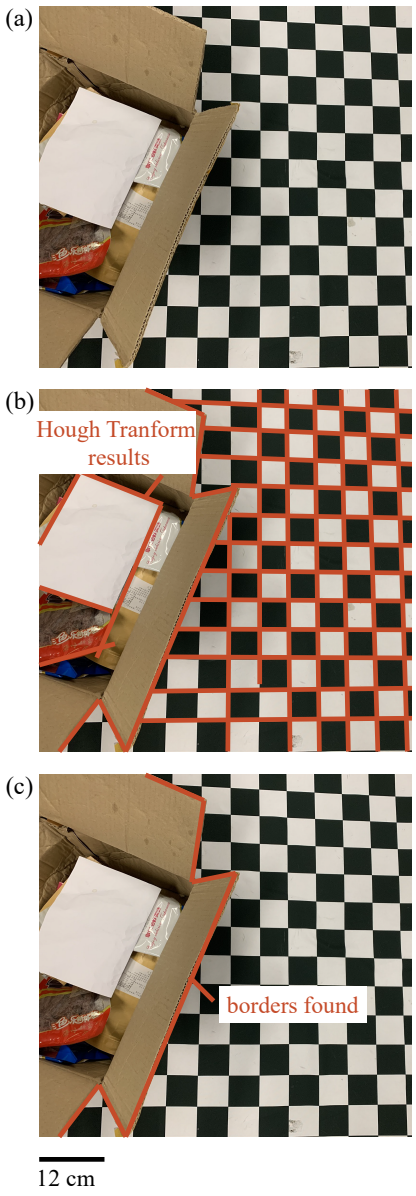


Fig. 12. Results of plane detector. (a) original image. (b) plane borders detected by Hough Tranform. (c) plane borders detected by plane detector.