Article

International Journal of Micro Air Vehicles

International Journal of Micro Air Vehicles
0(0) 1–14

DOI: 10.1177/1756829317705325
journals.sagepub.com/home/mav

SAGE

Vision-based SLAM system for MAVs in GPS-denied environments

Sarquis Urzua[1], Rodrigo Munguía[1] and Antoni Grau[2]

Abstract

Using a camera, a micro aerial vehicle (MAV) can perform visual-based navigation in periods or circumstances when GPS is not available, or when it is partially available. In this context, the monocular simultaneous localization and mapping (SLAM) methods represent an excellent alternative, due to several limitations regarding to the design of the platform, mobility and payload capacity that impose considerable restrictions on the available computational and sensing resources of the MAV. However, the use of monocular vision introduces some technical difficulties as the impossibility of directly recovering the metric scale of the world. In this work, a novel monocular SLAM system with application to MAVs is proposed. The sensory input is taken from a monocular downward facing camera, an ultrasonic range finder and a barometer. The proposed method is based on the theoretical findings obtained from an observability analysis. Experimental results with real data confirm those theoretical findings and show that the proposed method is capable of providing good results with low-cost hardware.

Keywords

Micro aerial vehicles, monocular SLAM, visual-based navigation, GPS-denied, state estimation

Date received: 11 July 2016; accepted: 13 March 2017

Introduction

The state estimation of vehicle position is a fundamental necessity for any application involving an autonomous micro aerial vehicle (MAV). In outdoor environments, this problem is seemingly solved with on-board global positioning system (GPS) and inertial measurement units (IMU), or with their integrated version, the Inertial Navigation Systems (INS). Contrary, unknown, cluttered and GPS-denied environments still pose a considerable challenge.

While available attitude and heading systems (AHRS) handle well the estimation of the orientation of MAVs, GPS-based position estimation has some drawbacks. Especially, GPS is not a reliable service as its availability can be limited in urban canyons and is completely unavailable in indoor environments. Moreover, even when the GPS signal is available, the problem of position estimation could be not solved for several scenarios. For instance, in Munguia et al.,[1] it is shown that the precision of standard GPS could not be enough in order to perform precision maneuvers.

Simultaneous localization and mapping (SLAM) methods can be used for addressing the problem of the state estimation of MAVs. In this case, MAV operates in a priori unknown environment using only on-board sensors to simultaneously build a map of its surroundings which it used to track its position. This means that no external infrastructure (i.e. GPS) is needed in order to localize the vehicle. Many different kinds of sensors can be used for implementing SLAM systems, for instance, laser[2,3] sonar,[4,5] sound sensors,[6] RFID[7] or computer vision.[8–10] The selection of such a sensor technology has a great impact on the algorithm used in SLAM and, depending on the application and other factors, each technology has some strong and weak points.

[1]Department of Computer Science, CUCEI, University of Guadalajara, Guadalajara, México
[2]Department of Automatic Control, Technical University of Catalonia, Barcelona, Spain

Corresponding author:
Rodrigo Munguía, Departamento de Ciencias Computacionales, CUCEI, UdeG, Blvd. Marcelino García Barragán 1421, C.P. 44430 Guadalajara, Jalisco, México.
Email: rodrigo.munguia@upc.edu

In the case of MAVs, there exist several limitations regarding to the design of the platform, mobility and payload capacity that impose considerable restrictions on the available computational and sensing resources. The aforementioned issues have motivated that recent works move towards the use of cameras as the primary sensor. Cameras provide a lot of information and are well adapted for embedded systems because they are light, cheap and power-saving.

Using a camera, an MAV can perform visual-based navigation in periods or circumstances when the position sensor is not available, when it is partially available, or when a local navigation application requires high precision. In particular, and compared to another kind of visual configurations (e.g. stereo vision), the use of monocular vision has some advantages in terms of weight, space, power-saving, or scalability. On the other hand, it introduces some technical difficulties as the challenge of directly recovering the metric scale of the world.[9]

Related work

In the case of monocular vision with application to aerial vehicles, different approaches have been followed for estimating the state of the vehicle without the aid of a GPS system. In Mirzaei and Roumeliotis,[11] a monocular SLAM system is proposed where the scale factor is retrieved from a feature pattern with known dimensions. In the case of monocular SLAM systems proposed in Weiss et al.[12] and Foster et al.,[13] the map is initially set by hand, by aligning the first estimates with the ground-truth in order to get the scale of the environment. In Celik and Somani,[14] the problem of the scale recovering is addressed for environments formed by corridors, like those commonly found in office buildings. In this case, several assumptions are made about the structure of the environment, like the flatness of the floor. Also, it is assumed that the relative altitude of the MAV respect the floor is known by the use of an ultrasonic range sensor, as well as the distance from the MAV to the wall of the corridor. Another approach for recovering the metric scale consists in integrating inertial measurements from an IMU (accelerometer and gyroscopes). In particular, in Nutzi et al.,[15] the scale is explicitly considered in the system state, and it is estimated through an Extended Kalman Filter (EKF). The filter makes use of an innovation error defined by the difference between the unscaled acceleration (obtained from the monocular vision), and the measured acceleration in the vertical axis (obtained from the IMU). In Wang et al.,[16] the same approach is also followed. The potential problem with this kind of approach has to do with the fact that the acceleration obtained from an IMU has a dynamic bias which is difficult to estimate. This bias introduces at the same time a bias in the estimated scale. Moreover, in this kind of set-up, it is required a precise calibration for the alignment of the camera and the IMU. In Chowdhary et al.,[17] an EKF-based method is proposed in order to perform visual odometry with an unmanned aircraft. This method makes use of inertial sensors, a monocular downward facing camera, and a range sensor (sonar altimeter). Unlike vision-based SLAM, there is no mapping process in visual odometry approaches.

Objectives and contributions

In this work, a novel monocular SLAM system with application to MAVs is proposed. The method is intended to be useful for performing visual-based navigation in fully GPS-denied environments or as a backup system in periods where GPS-signal is not available. In order to estimate the state of the vehicle and a map of the environment, the proposed system makes use of: (i) visual information captured from a monocular downward facing camera, (ii) range measurements obtained from an ultrasonic range finder, and (iii) measurements of atmospheric pressure obtained from a barometer.

An observability test is carried out in order to analyze the problem of the recovery of the metric scale. The theoretical findings obtained from this test are used as a basis for developing the proposed system.

Unlike the approaches,[15–17] the proposed method does not make use of inertial sensors and thus, there is no need of an extensive pre-calibration routine for aligning the IMU and the camera. In Chowdhary et al.,[17] a range finder is used as an altimeter, and therefore it is assumed that all the landmarks lie on a plane (flat terrain assumption). In this work, the range finder is used for computing an approximation of the relative depth of features. Thus, the assumption of a completely flat terrain is considerably relaxed by the assumption of a terrain with soft but continuous changes in altitude.

Moreover, in this work, a barometer is used for incorporating altitude information into the system in order to improve the observability of the metric scale. In certain conditions, the use of the barometer may be sufficient for recovering the metric scale when the ultrasonic range finder is out of its operation range. For instance, in Celik and Somani[14] or in Chowdhary et al.,[17] accelerometers are used for recovering the metric scale, but the dynamic error bias of an accelerometer is larger than the error bias of a barometer.[18] Moreover, the barometer is commonly used as augmentation sensor for limiting the error in inertial navigation systems.

Preliminaries

The problem to be addressed will be introduced using a simplified 2-DOF model. It is important to note that

this model is representative of the main aspects of the full problem. Later, the proposal will be extended in order to be applied to a three-dimensional context.

Consider the following unconstrained model $\dot{x}_c = f(x,u)$ of a camera attached to an MAV (see Figure 1)

$$\begin{aligned} &\dot{x}_c = v_x \quad \dot{z}_c = v_z \quad \dot{\theta}_c = \omega_c \quad \dot{v}_x = V_x \\ &\dot{v}_z = V_z \quad \dot{\omega}_c = \Omega \end{aligned} \tag{1}$$

Let $x_c = [x_c, z_c, \theta_c, v_x, v_z, \omega_c]^T$ be the vector state of camera C_s. Let $[x_c, z_c, \theta_c]$ represent the position and orientation of the camera, and $[v_x, v_z, \omega_c]$ their first derivatives. In this model, it is assumed an unknown input $u = [V_x, V_z, \Omega]^T$ of linear and angular accelerations with zero-mean and known-covariance Gaussian processes. Also it is assumed that the camera C_s is capable of detecting and tracking 2D feature points. The measurement process is modeled by equations of the form

$$y_i = h_\theta i(x) = \arctan 2\left(\frac{z_c - z_i}{x_c - x_i}\right) - \theta_c \tag{2}$$

Let $[x_i, z_i]$ be the Euclidean position of a ith feature coded by its inverse form

$$\begin{aligned} x_i &= (1/\rho_i)\cos(\theta_{0_i}) + x_{0_i} \\ z_i &= (1/\rho_i)\sin(\theta_{0_i}) + z_{0_i} \end{aligned} \tag{3}$$

The state of a ith feature y_i is defined by $y_i = [x_{0_i}, z_{0_i}, \theta_{0_i}, \rho_i]^T$, let $[x_{0_i}, z_{0_i}]$ be the position of the camera C_s when the feature was first detected, let θ_{0_i} be the first bearing measurement, and let $\rho_i = 1/d_i$ be the inverse of the feature depth d_i, (see Figure 1).

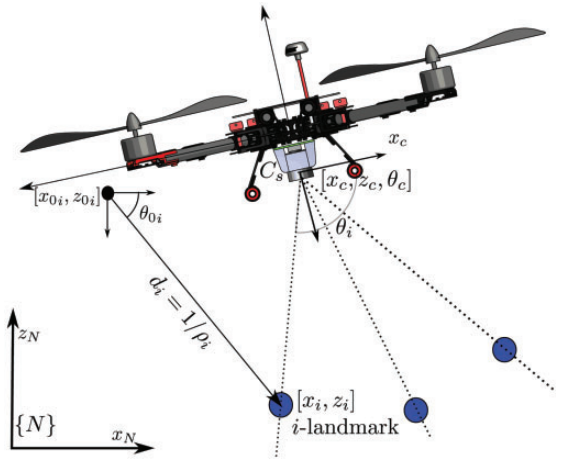

Figure 1. System parametrization.

The system state x to be estimated is composed by the camera state x_c, and it is augmented with the state y_i of every feature contained in the map

$$x = [x_c, y_1, y_2, \ldots, y_n]^T \tag{4}$$

When a feature is detected for the first time from a monocular camera (bearing sensor), only information about the ray can be retrieved. In the case of the well-known undelayed inverse depth (UID) EKF-SLAM method proposed in Montiel et al.,[19] new features are incorporated into the system state by assuming an hypothetical initial inverse depth Gaussian prior on $\rho_i \sim N(\rho_0, \sigma_{\rho_0})$, which is applied in order to cover with a probability of 95% the range of depths from the closest one to infinity.

Observability of metric scale

The use of monocular vision introduces some technical challenges. First, depth information is difficult to retrieve in a single frame and hence, robust techniques for recovering the depth of the features are required.[20,21] Another challenging aspect of working with monocular sensors has to do with the difficulty of directly recovering the metric scale of the world. Davison et al.[9] stated that if no additional information is used and a single camera is used as the solely source of data to the system, then the map and trajectory can be recovered without metric information.

In this section, the problem of the observability of monocular SLAM systems applied to MAVs is studied. The theoretical results obtained here will be used later as a basis for developing a monocular-based SLAM system capable of recovering the metric scale without need of absolute position measurements as those provided by GPS.

In Civera et al.,[22] it is shown that a monocular SLAM system can be initialized with no prior knowledge of scene objects within the context of a dimensionless parametrization of the problem. It is also described how the monocular SLAM state vector can be partitioned into two parts: a dimensionless part, representing up-to-scale scene and camera motion geometry, and an extra metric parameter representing scale.

In our work, in order to make explicit the relevance of the metric scale in monocular SLAM problem, the approach proposed in Civera et al.[22] is followed. In this case, the state vector defined in equation (4) is split into a metric parameter s, unobservable when only angular measurements are available, and a dimensionless map and camera part.

$$x_s = [s,\ \Pi_{x_c},\ \Pi_{z_c},\ \theta_c,\ \Pi_{v_x},\ \Pi_{v_z},\ \Pi_{\omega_c},\ \Pi_{y_1} \ldots, \Pi_{y_n}]^T \tag{5}$$

Camera measurements will reduce scene geometry uncertainty, but not the uncertainty in the metric parameter s. The mapping process from the state vector x_s to the metric geometry is a non-linear computation process involving the dimensionless geometry and the parameter s

$$\begin{cases} x_c = s\Pi_{x_c} \quad z_c = s\Pi_{z_c} \quad v_x = s\Pi_{v_x}\Delta t \\ v_z = s\Pi_{v_z}\Delta t \quad \omega_c = s\Pi_{\omega_c}\Delta t \\ y_i = [s\Pi_{x_0 i},\ s\Pi_{z_0 i},\ \theta_i,\ \Pi_{\rho_i}/s] \end{cases} \tag{6}$$

In order to define the system dynamics in terms of the metric parameter s and the dimensionless parameters, equation (6) is substituted into equations (1) to (3), and the system state is augmented with s. Hence, the system dynamics becomes

$$\begin{aligned} &\dot{s} = 0 \quad \dot{x}_c = s\Pi_{v_x}\Delta t \quad \dot{z}_c = s\Pi_{v_z}\Delta t \quad \dot{\theta}_c = s\Pi_{\omega_c}\Delta t \\ &\dot{v}_x = 0 \quad \dot{v}_z = 0 \quad \dot{\omega}_c = 0 \quad \dot{\rho}_i = 0 \end{aligned} \tag{7}$$

In the system represented by equation (7), a constant-acceleration camera model is assumed. Also it assumed a rigid scene (map features remain static) and a constant metric scale. The system output equation is

$$y_i = h_{\theta_i}(x) = \arctan2\left(\frac{s\Pi_{z_c} - z_i}{s\Pi_{x_c} - x_i}\right) - \theta_c \tag{8}$$

where

$$\begin{aligned} x_i &= (s/\Pi_{\rho_i})\cos(\theta_i) + x_{0_i} \\ z_i &= (s/\Pi_{\rho_i})\sin(\theta_i) + z_{0_i} \end{aligned} \tag{9}$$

Remark 1: In applications like aerial vehicles, the attitude and heading (roll, pitch and yaw) estimation is well handled by available systems.[23,24] In particular, in this work, it is assumed that the orientation of the camera always points toward the ground. In practice, the foregoing assumption can be easily addressed with the use of a servo-controlled camera gimbal. Note that in this downward pointing camera configuration, the roll of the camera is aligned with the heading of the MAV.

Considering the above aspects, the system state can be simplified by removing the variables related to attitude and heading (which are provided by the AHRS). Therefore, the problem will be focused on the position estimation of the MAV.

Remark 2: From section *Preliminaries*, let recall that the state of a ith feature y_i is defined by $y_i = [x_{0_i}, z_{0_i}, \theta_i, \rho_i]$ where $[x_{0_i}, z_{0_i}]$ is the position of the camera C_s when the feature was first detected, θ_i is the first bearing measurement, and $\rho_i = 1/d_i$ is the inverse of the feature depth d_i. Because $[x_{0_i}, z_{0_i}, \theta_i]$ is directly given when the ith feature is initialized, the observability analysis will be focused on the state of the camera C_s and the inverse depth of the features.

Considering the above remarks the system state x_s becomes

$$x_s = \left[s,\ \Pi_{x_c},\ \Pi_{z_c},\ \Pi_{v_x},\ \Pi_{v_z},\ \Pi_{\rho_1},\ \Pi_{\rho_2}, \ldots, \Pi_{\rho_n}\right]^T \tag{10}$$

If n landmarks are measured by the camera, the system output is defined as $y = [h_{\theta_1}, \ldots, h_{\theta_n}]^T$.

A system is defined as observable if the initial state x_0, at any initial time t_0, can be determined given the state transition and observation models of the system and observations $y[t_0, t]$ from time t_0 to a finite time t. When a system is fully observable, the lower bound of the error in the estimations of its state will only depend on the noise parameters of the system and will not be reliant on initial information about the state. This fact has important consequences in the context of SLAM.

In Hermann and Krener,[25] it is demonstrated that a non-linear system is *locally weakly observable* if the observability rank condition $\text{rank}(\mathcal{O}) = \dim(x)$ is verified. The observability matrix $\mathcal{O}$ is computed as

$$\mathcal{O} = \left[\frac{\mathcal{L}_f^0(h_{\theta_1})}{\partial x}^T \frac{\mathcal{L}_f^1(h_{\theta_1})}{\partial x}^T \cdots \frac{\mathcal{L}_f^0(h_{\theta_n})}{\partial x}^T \frac{\mathcal{L}_f^1(h_{\theta_n})}{\partial x}^T\right]^T \tag{11}$$

where $\mathcal{L}_f^i(h)$ is the ith order Lie Derivative[26] of the scalar field of the measurement h with respect to the vector field f. Note that in equation (11), the zero-order and first-order Lie Derivatives are used for each bearing measurement $y_i = h_\theta i(x)$. The observability matrix $\mathcal{O}$ was computed using the MATLAB symbolic toolbox.

The following result was obtained for the system conformed by the state of equation (10):

- The maximum degree of observability was obtained with four landmarks. In this case, $\dim(x_s) = 9$, $\text{rank}(\mathcal{O}) = 8$. Adding more landmarks the observability does not increase, and therefore, the system is partially observable, in this case with one non-observable mode. Based on Civera et al.,[22] the non-observable mode should correspond to the metric parameter s.

Now let consider that the measurements of altitude of the MAV camera are becoming available.

The additional system output equation y_a is

$$y_a = h_{z_c}(x) = z_c = s\Pi_{z_c} \tag{12}$$

Hence, if n landmarks are measured by the camera, the system output is now defined as $y = [h_{z_c}, h_{\theta_1}, \ldots, h_{\theta_n}]^T$. The observability matrix $\mathcal{O}$ is now computed from

$$\mathcal{O} = \begin{bmatrix} \frac{\mathcal{L}_f^0(h_{z_c})^T}{\partial x} & \frac{\mathcal{L}_f^1(h_{z_c})^T}{\partial x} & \frac{\mathcal{L}_f^0(h_{\theta_1})^T}{\partial x} & \frac{\mathcal{L}_f^1(h_{\theta_1})^T}{\partial x} \\ \cdots & \frac{\mathcal{L}_f^0(h_{\theta_n})^T}{\partial x} & \frac{\mathcal{L}_f^1(h_{\theta_n})^T}{\partial x} \end{bmatrix}^T \tag{13}$$

By considering the availability of measurements of altitude, the following results were obtained:

- The maximum degree of observability was obtained with only three landmarks, but in this case, the system becomes observable, that is, $\dim(x_s) = 8$, $\text{rank}(\mathcal{O}) = 8$.
- The movement of the vehicle on the vertical axis ($\Pi_{v_z} \neq 0$) is a sufficient condition of full observability.

Now let consider that instead of altitude readings, the measurements of range are available for a subset of features. The range measurement of ith feature is modeled by an equation of the form

$$y_d = h_{\rho_i}(x) = 1/\rho_i = s/\Pi_{\rho_i} \tag{14}$$

Hence, if n landmarks are measured by the camera C_s, and m range measurements ($m \leq n$) are available, the system output is defined as $y = [h_{\theta_1}, \ldots, h_{\theta_n}, h_{\rho_1}, \ldots, h_{\rho_m}]^T$. The observability matrix $\mathcal{O}$ is computed as

$$\mathcal{O} = \begin{bmatrix} \frac{\mathcal{L}_f^0(h_{\theta_1})^T}{\partial x} & \frac{\mathcal{L}_f^1(h_{\theta_1})^T}{\partial x} & \cdots & \frac{\mathcal{L}_f^0(h_{\theta_n})^T}{\partial x} & \frac{\mathcal{L}_f^1(h_{\theta_n})^T}{\partial x} \\ \times & \frac{\mathcal{L}_f^0(h_{\rho_1})^T}{\partial x} & \cdots & \frac{\mathcal{L}_f^0(h_{\rho_m})^T}{\partial x} \end{bmatrix}^T, \tag{15}$$

Note that in equation (15) only the zero-order Lie Derivative is used for range measurements. Considering the availability of such range measurements the following result was obtained:

- Independently of the number of features included into the system state, full observability, $\dim(x_s) = \text{rank}(\mathcal{O})$, can be reached by including a single measurement of range h_{ρ_i}.

It is important to note that the purpose of the previous analysis over a simplified 2-DOF model is not to prove (by extension) the full observability of the whole SLAM system, but to show that the metric parameter s representing scale could become observable if measurements of absolute altitude or measurements of range for a subset of features are available. A full observability analysis in 6-DOF should be of great interest for future work.

Method description

Sensor fusion approach

As it has previously seen, the metric scale is difficult to be recovered using monocular vision. On the other hand, it was found that the metric scale can become observable if altitude information is included into the system. As a consequence of this result, altitude measurements obtained from a barometer are incorporated into the system. However, it was also found that if this approach is followed, the observability depends on the movement of the vehicle. These facts mean that the effectiveness of this approach can be compromised in periods where the MAV performs only short and small movements.

Another interesting theoretical result has to do with the fact that the system becomes observable with the inclusion of a single measurement of depth of a map feature. In order to take advantage of this result for improving the robustness of the proposed method, a technique for incorporating approximate information about the depth of features is proposed. For the foregoing purpose, an ultrasonic range finder is also incorporated into the system.

Aerial platform

As shown in Figure 2, the platform considered in this work is a quadrotor moving freely in any direction in $\mathbb{R}^3 \times SO(3)$. However, it is important to highlight that the proposed monocular SLAM method could be applied to other kinds of aerial platforms.

The quadrotor is equipped with a barometer for measuring atmospheric pressure. The monocular camera is mounted over a servo-controlled gimbal. The gimbal is configured in order to counteract the changes in attitude of the quadcopter, and therefore it stabilizes the orientation of the camera toward the ground. Besides the camera, an ultrasonic range finder is also mounted over the servo-controlled gimbal. The ultrasonic sensor is aligned in parallel to the optical axis of the camera (see Figure 2). In this work, it is assumed that a possible misalignment of the camera respect to the ultrasonic sensor is negligible

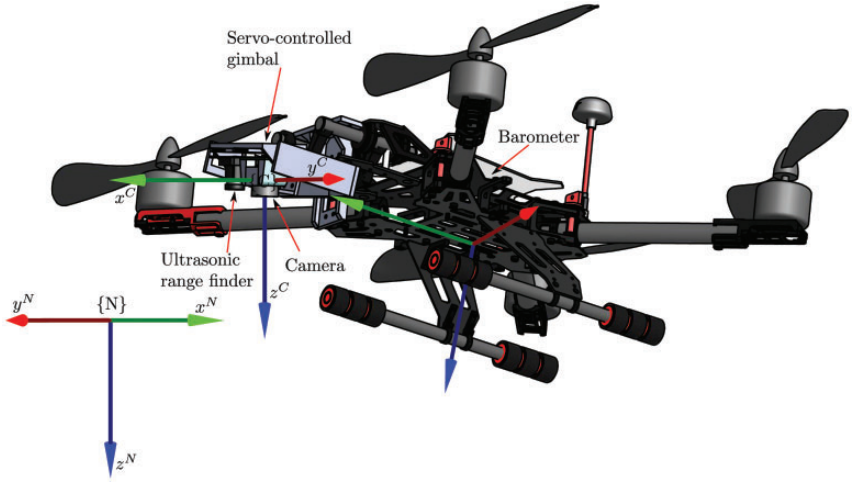

Figure 2. Aerial platform and coordinate systems.

if it is compared with other inherent constraints of the ultrasonic sensor. In this case, it is assumed that the lack of a fine calibration is taken into account by means of an uncertainty parameter used for modeling the error in sensor measurements.

The proposed method is mainly intended for local autonomous vehicle navigation. In this case, the local tangent frame is used as the navigation reference frame. Thus, the initial position of the vehicle defines the origin of the navigation coordinates frame. The navigation system follows the NED (North, East, Down) convention. The magnitudes expressed in the navigation and camera frame are denoted respectively by the superscripts ^{N}, and ^{C}. All the coordinate systems are right-handed defined. In this work, it is assumed that the location of the origin of the camera frame respect to other elements of the quadcopter (e.g. barometer) is known and determined. Hence, the position of the origin of the vehicle can be computed from the estimated location of the camera.

Remark 3: In section *Observability of metric scale*, the system state was split into a metric parameter s and a dimensionless map and camera part, in order to explicitly show that the metric scale of the system can become observable if altitude or range measurements are included into the system. By knowing this fact, hereinafter, the metric parameter will be considered implicit into the system variables.

Sensor measurement models

Visual measurements. A standard monocular camera is considered to be mounted aboard the quadrotor. In this case, a central-projection camera model is assumed. The image plane is located in front of the camera's origin where a non-inverted image is formed. The camera frame ^{C} is right-handed with the z-axis pointing to the field of view.

The $\mathbb{R}^3 \Rightarrow \mathbb{R}^2$ projection $p = (u, v)$ of a 3D point, located at $p^N = (x, y, z)^T$, to the image plane is defined by

$$u = \frac{x'}{z'} \qquad v = \frac{y'}{z'} \tag{16}$$

Let u and v be the coordinates of the image point p expressed in pixel units, and

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{bmatrix} p^C \tag{17}$$

Let p^C be the same 3D point p^N, but expressed in the camera frame ^{C} by $p^C = R^{NC} p^N$. Let R^{NC} be the rotation matrix that allows to transform from the navigation frame ^{N} to the camera frame ^{C}. Also, it is fulfilled that $R^{NC} = (R^{CN})^T$, and R^{CN} is known by the use of the gimbal.

Inversely, a directional vector $h^C = [h_x^C, h_y^C, h_z^C]^T$ can be computed from the image point coordinates u and v as

$$h^C = \left[\frac{u_0 - u}{f}, \frac{v_0 - v}{f}, 1 \right]^T \tag{18}$$

Vector h^C points from the camera optical center position to the 3D point location. The vector h^C can be expressed in the navigation frame by $h^N = R^{CN} h^C$. Note that for the $\mathbb{R}^2 \Rightarrow \mathbb{R}^3$ mapping case, defined in equation (18), depth information is lost.

The distortion caused by the camera lens is considered through the model described in Bouguet.[27] In this case, radial and tangential distortions are considered. Using the former model (and its inverse form), undistorted pixel coordinates $(u,\ v)$ can be obtained from (u_d, v_d) and conversely. In this work, it is assumed that the intrinsic parameters of the camera are already known: focal length f, principal point $(u_0,\ v_0)$, and radial lens distortion $k_1, \ldots, k_n$.

Range measurements. An ultrasonic range finder is used, whenever is possible, in order to obtain approximate information about the range (depth) of features. The idea is to define an image region where visual features lying inside could be associated with a range provided by the ultrasonic sensor. The image region is a function of the range measured by the sensor as well as its beam pattern.

Every time that a range reading, provided by the ultrasonic sensor, is available, it will associated with the next camera frame. Due to the operating frequency of ultrasonic range sensors, typically lower (3–4 Hz) than the frame rate of cameras, only a subset of the frames will have an associated range.

The beam pattern of the ultrasonic range finder is modeled by an elliptic paraboloid (see Figure 3) satisfying: $z = ax^2 + ay^2$. The parameter a is chosen in order to adjust the paraboloid (as far as possible) to the actual beam pattern of the sensor (see Figure 3, right).

First, let z_r be the range measured by the ultrasonic sensor. Then, a circle $z_r = ax^2 + ay^2$ is defined by the intersection of the paraboloid and the plane $z = z_r$ (see Figure 3, left). The region defined by this circle represents the ground surface from which depth of features can be inferred from the ultrasonic sensor. In order to define a circular region over the image plane, a 3D point belonging to the circle (e.g. $p^N = (\sqrt{z_r/a}, 0, z_r)$) is projected to the image plane using equation (16). The radius r_c of the circular image region (in pixels) is computed by $r_c = \|[u, v]^T - [u_0, v_0]^T\|$. Note that the radius $r_c = f(z_r, a)$ is a function of the range z_r and the parameter a, which defines the paraboloid representing the beam pattern of the sensor. The size of the circular image region increases as the vehicle flies closer to the ground (see Figure 4).

A 3D point located at $p^N = (x, y, z)^T$ with projection $(u_i,\ v_i)$ in the image plane, which lies inside the circle with radius r_c and center at $(u_0,\ v_0)$, is assumed to have an approximate depth d computed by

$$d = \frac{z_r \|h^N\|}{h_z^N} \tag{19}$$

Let h^N be the directional vector pointing from the camera to the location of the 3D point (see the related equation (18)), and let h_z^N be the z component of h^N.

In other related methods, the range measured by the telemeter is associated (directly or indirectly) with all the visual features seen by the camera.[17] In the proposed method, the range is associated only with the visual features lying inside of the region covered by the beam pattern of the ultrasonic sensor. Hence, it is assumed that only the portion of the ground detected by the ultrasonic sensor is approximately planar.

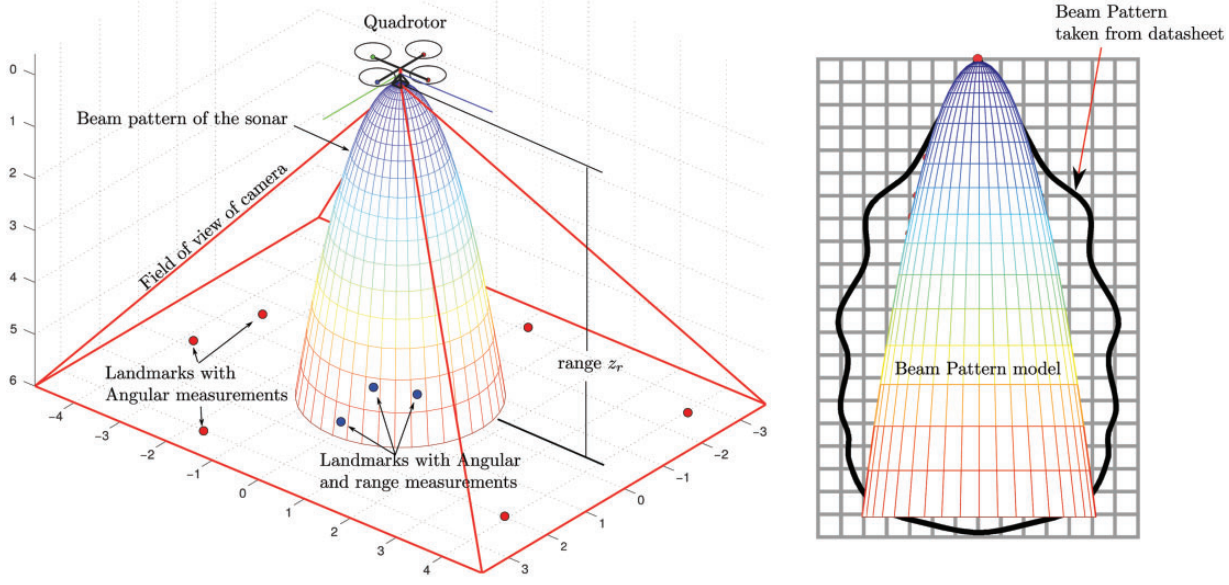

Figure 3. The ultrasonic range finder is used for estimating the approximate depth of visual features lying in its beam pattern (left). An elliptic paraboloid is used as a simple model of the actual beam pattern of the ultrasonic range finder (right).

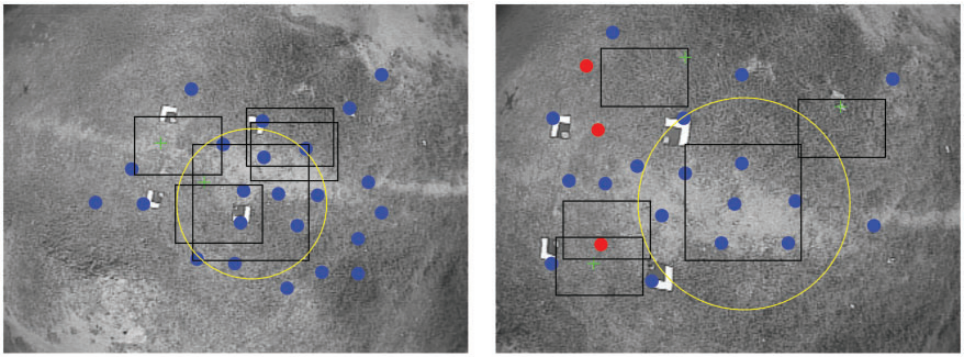

Figure 4. The size of the terrain region detected by the ultrasonic range finder is a function of the beam pattern of the sensor and the flight altitude of the MAV. Frame captured at 7 m of altitude (left). Frame captured at 3.9 m of altitude (right).

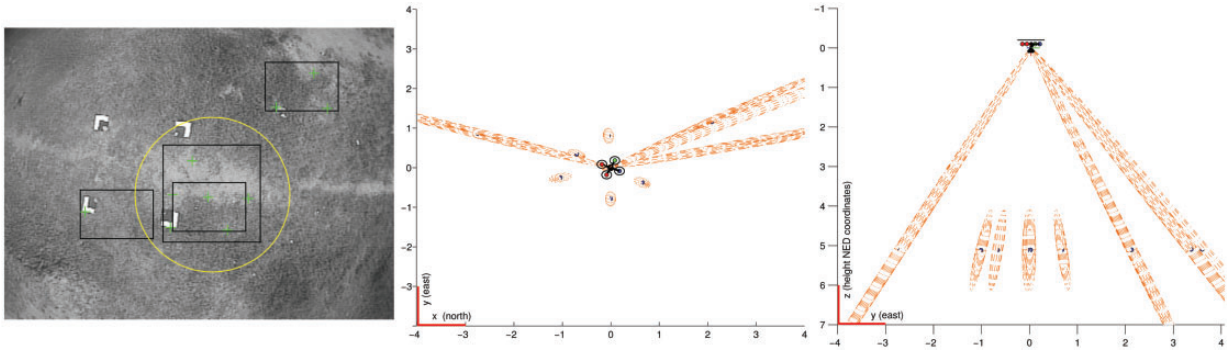

Figure 5. Initialization of new map features: Camera frame with an associated range measurement (left plot). Zenithal view (middle plot) and sectional view (right plot) of the estimates. The ellipses indicate the initial uncertainty region of a new feature 3D location. Note that the visual features lying inside of the circular image region, determining the portion of the terrain detected by the ultrasonic range finder, are initialized as new map features with a smaller initial uncertainty in depth.

This is important because the assumption of a quasi-flat terrain is locally restricted to a circular central image region and not the global image itself (see Figure 5). In this manner, the flat-terrain assumption could be relaxed if the whole terrain is supposed to be composed of several connected sub-regions with small but continuous changes in altitude between them.

Altitude measurements. Measurements of altitude can be inferred from measurements of atmospheric pressure. The proposed method is mainly intended for local autonomous vehicle navigation. Hence, the altitude or height of the MAV above a local ground location is computed from the change in pressure between the ground and the altitude of interest. The following formula can be used for computing the local altitude z_a from a barometer

$$z_a = \left(1 - \left(\frac{P}{P_g}\right)^{\frac{RL_0}{Mg}}\right)\frac{T}{L_0} \tag{20}$$

where P is the current barometric pressure measurement; P_g is the barometric pressure at the initial position (home position); $R = 8.31432$ N-m/(mol-K) is the universal gas constant for air; $L_0 = -.0065$ K/m is the rate of temperature decrease in the lower atmosphere; $M = 0.0289644$ kg/mol is the standard molar mass of atmospheric air; $g = 9.80665\,\mathrm{m/s^2}$ is the gravitational constant and T is the temperature at flight location in Kelvin degrees.

EKF-SLAM

The system state to be estimated is

$$x = \left[r^N, v^N, y_1, y_2, \ldots, y_n\right]^T \tag{21}$$

Let $r^N = [x_c, y_c, z_c]$ represent the position of the vehicle (camera) expressed in the navigation frame, and let $v^N = [v_x, v_y, v_z]$ denote the linear velocity of the vehicle expressed in the navigation frame.

Map features are defined by

$$y_i = [r_i, \theta_i, \phi_i, \rho_i]^T \tag{22}$$

Let $r_i = [x_{0_i}, y_{0_i}, z_{0_i}]$ be the coordinates of the center of the camera when the feature was observed for the very first time; let θ_i and ϕ_i be azimuth and elevation, respectively, and let $\rho_i = 1/d$ be the inverse of the depth d, and

$$\theta_i = \text{atan2}\left(h_y^N, h_x^N\right) \quad \phi_i = \text{acos}\left(\frac{h_z^N}{\sqrt{\left(h_x^N\right)^2 + \left(h_y^N\right)^2 + \left(h_z^N\right)^2}}\right) \tag{23}$$

Let $h^N = [h_x^N, h_y^N, h_z^N]^T$ be computed from equation (18). The architecture of the system is defined by the typical loop of prediction updates of the standard EKF-SLAM, where the EKF propagates the vehicle state as well as the feature estimates. Interested readers are referred to literature[28,29] for an extensive review on the EKF-SLAM methodology.

System prediction. The system state x is taken a step forward by the following discrete model

$$\begin{cases} r_{k+1}^N = r_k^N + v_k^N \Delta t \\ v_{k+1}^N = v_k^N + V^N \\ y_{1[k+1]} = y_{1[k]} \\ \vdots \\ y_{n[k+1]} = y_{n[k]} \end{cases} \tag{24}$$

At every step, it is assumed that there is an unknown linear velocity with acceleration zero-mean and known-covariance Gaussian processes σ_a, producing an impulse of linear velocity: $V^N = \sigma_a^2 \Delta t$.

Note that in this work, for simplicity, a Gaussian random process is used for propagating the velocity of the vehicle. However, a feasible alternative could be the use of the dynamical model of the aircraft instead of the Gaussian random process. However, this approach commonly requires having a considerable knowledge about the specific physics of each aerial vehicle where the proposed method could be applied.

Initialization of map features. The initialization process of new map features is carried out using frames with an associated range z_r (see subsection *Range measurements*). A random search is conducted over the image in order to detect new visual features in regions without them.

In this work, the detector proposed in Rosten and Drummond[30] is used for detecting new visual features, but any other detector with good performance could be used instead.

The coordinates of a ith new visual feature are defined by (u_i, v_i).

Any new feature $y_{new} = J(x, u_i, v_i, d_i)$ to be included in the map is defined by

$$y_{new} = [r_0, \theta_0, \phi_0, \rho_0]^T \tag{25}$$

with $[r_0, \theta_0, \phi_0]$ calculated as in equation (22); and $\rho_0 = 1/d_0$, being d_0 the hypotheses of depth computed with equation (19) from the range measurement z_r. Therefore, the system state x is augmented by: $x_{new} = [r^N, v^N, y_1, y_2, \ldots, y_n, y_{new}]^T$.

The new covariance matrix P_{new} is computed by

$$P_{new} = \nabla J \begin{bmatrix} P_{old} & 0 & 0 \\ 0 & R & 0 \\ 0 & 0 & \sigma_\rho^2 \end{bmatrix} \nabla J^T \tag{26}$$

where ∇J is the Jacobian for the initialization function; R is the measurement noise covariance matrix for (u_i, v_i); and σ_ρ is chosen according to the probability of being measured by the range sensor (see Figure 5):

- σ_ρ is chosen with a small value ($\sigma_\rho = \rho_0/10$) if the new visual feature (u_i, v_i) lies inside the circular image region, with radius r_c, computed as it has been defined in subsection *Range measurements*.
- σ_ρ is chosen with a big value ($\sigma_\rho = \rho_0/2$), as proposed in the UID method (See section *Preliminaries*), if the new visual feature lies outside the circular image region, in order to cover a big depth uncertainty region.

Measurement of map features. In order to update the filter with the re-observation of map features the following approach is followed:

Case 1: For frames without an associated range z_r, the classical visual measurement model $(u, v) = h_i(x)$, which is related to features parametrized in their inverse depth form, is used.

Each feature y_i does model a 3D point p^N located at

$$p^N = r_i + \frac{1}{\rho_i} m(\theta_i, \phi_i) \tag{27}$$

where $m(\theta_i, \phi_i) = (\cos\theta \sin\phi, \sin\theta \sin\phi, \cos\phi)^T$ is the unit vector defined by the pair of azimuth-elevation angles. Then, p^N is expressed in the camera frame and projected to the image plane by equations (16) and (17) (See subsection *Range measurements*).

Case 2: For frames with an associated range z_r, the measurement model used for updating the filter depends on whether the measured location of the visual feature lies inside the circular image region, with radius r_c, which is computed as defined in subsection *Range measurements*:

- If the visual feature lies outside the circle, it is assumed that the landmark is out of range of the ultrasonic range finder and hence, the visual measurement model $(u, v) = h_i(x)$, which is defined in Case 1, is used.
- On the other hand, if the visual feature lies inside the circle, it is assumed that the landmark is on the range of the ultrasonic range finder. Hence, depth information is incorporated through the measurement model $(u, v, d) = h_i(x)$, where (u, v) is computed by the visual measurement model defined in Case 1, and d is the feature depth $(d = 1/\rho_i)$. In this case, the actual measurement of depth is computed from the range z_r using equation (19).

Altitude updates. Whenever is possible, information about the relative altitude of the MAV is incorporated to the filter using the standard update equations.

The measurement model of altitude $z_c = h_i(x)$ is simply the estimated altitude of the vehicle taken from the current state vector. The actual measurement of altitude is computed from equation (20), (see subsection *Altitude measurements*).

Experimental results

To perform the experiments with real equipment, a custom-built quadrotor is used (Figure 6). The vehicle is equipped with an Ardupilot unit as flight controller,[31] a Radio Telemetry 3DR 915 MHz, a DX201 DPS camera with wide angle lens, a 5.8-GHz video transmitter and an ultrasonic range finder XL-MaxSonar-EZ0. The flight controller has a built-in barometer. The camera is mounted over a very low-cost gimbal which is servo-controlled by standard servomotors.

Figure 6. A custom-built quadrotor was used for performing the experiments with real data.

In experiments, the quadrotor has been manually radio-controlled. A custom-built C++ application running over a laptop has been used for capturing data from the vehicle received via MAVLINK protocol,[32] as well as capturing the digitalized video signal transmitted from the vehicle. The data obtained from the sensors (barometer and ultrasonic sensor) as well as the frames captured by the camera were synchronized and stored in a dataset. The frames with a resolution of 320×240 pixels, in gray scale, were captured at 26 *fps*. A MATLAB® implementation of the proposed method was executed offline over the dataset, in order to estimate the flight trajectory and the map of the environment.

In experiments, in order to evaluate the performance of the proposed method, an independent trajectory computed by a perspective on 4-point (P4P) technique was used. For computing the P4P trajectory, four marks were placed in the floor, forming a square of known dimensions (see Figure 4). Then, a perspective on 4-point (P4P) technique[33] was applied to each frame in order to compute the relative position of the camera with respect to this known reference. It is important to note that the trajectory obtained by means of this technique should not be considered as a perfect reference of ground-truth. However, this approach was very helpful to have a fully independent reference of flight for evaluation purposes.

Two different flight trajectories (*a* and *b*) were performed over two different outdoor test fields consisting in urban parks. Figure 7 shows the trajectory and map for both cases, estimated by means of the proposed method. The upper plots show a 3D view of maps and estimated trajectories. The middle plots show the zenithal (x-y) view of maps and estimated trajectories. The lower plots show the sectional (z-x/y) view of maps and estimated trajectories. In this case, a good concordance between the P4P trajectory and the one computed with the proposed method was found.

A comparative study has been performed in order to gain more insight about the performance of the proposed method. For this purpose, the same flight trajectories were also computed under other two different conditions: (i) using only visual information, (ii) using visual and altitude information.

In the first case (i), the undelayed inverse depth method (UID)[19] was used for initializing the visual features. Because no extra aid of any kind of sensory information is used, the metric scale of the estimates depends only on the initial hypotheses of inverse depth ρ_0, which is a parameter to be manually set. Considering that the flights of the MAV were performed at low altitude,

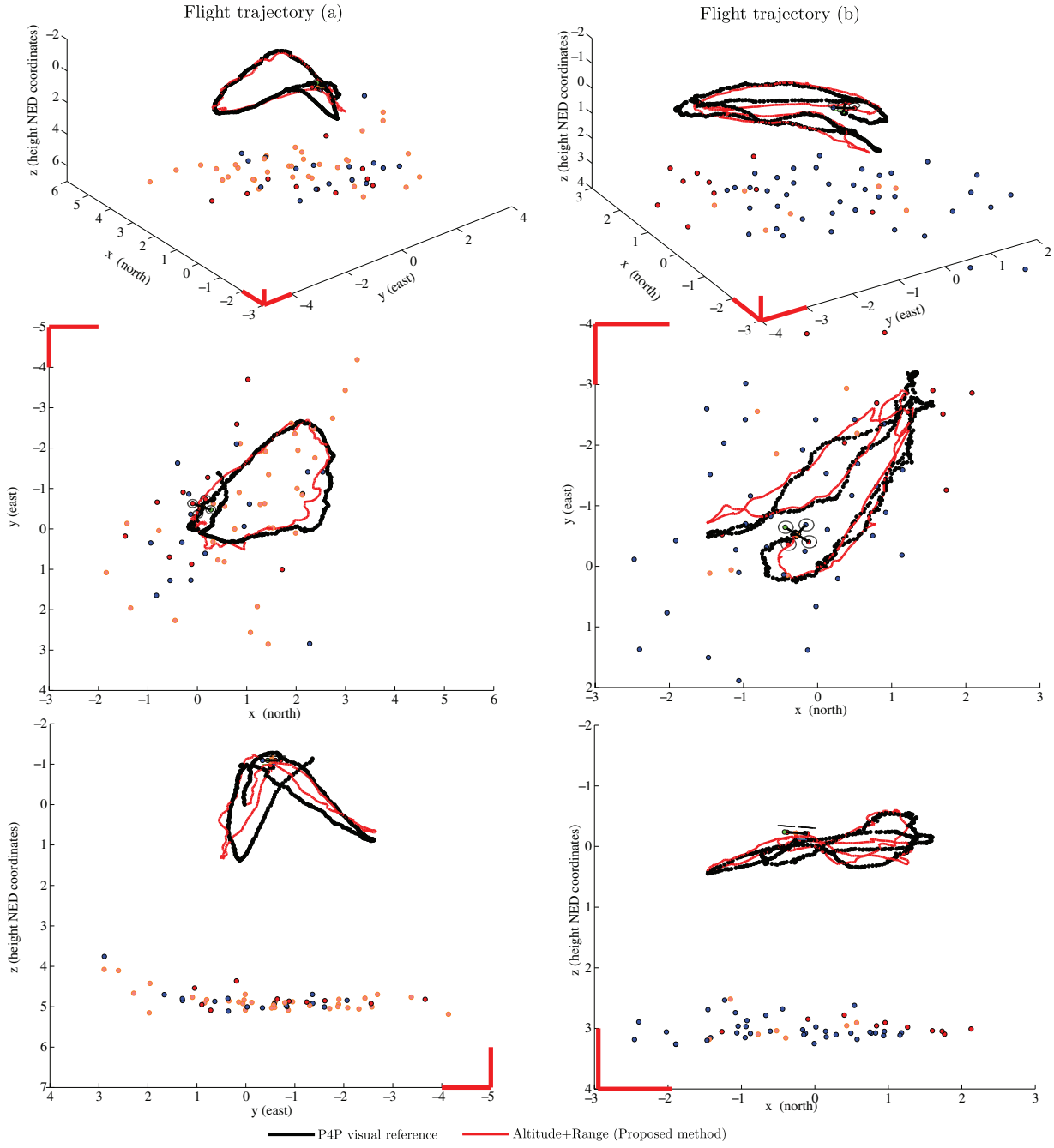

Figure 7. Flight trajectory and map estimated with the proposed method. For comparison purposes the flight trajectory was also computed in an independent manner using a P4P technique.

a value of initial depth $\rho_0 = 1/2$ m was used in experiments. The usefulness of the UID variant is to have a scaled estimation of the flight trajectories, and use it as a reference in order to evaluate the benefits obtained from incorporating information from other sources such as the barometer and the ultrasonic sensor. In this sense, recalling from section *Observability of metric scale*, the metric scale can become observable if altitude or range information is incorporated into the system. In the second case (ii), the same conditions of the UID method are used but also measurements of altitude obtained from the barometer are incorporated into the system (see subsection *Altitude updates*).

Figure 8 shows the progression over time for each estimated trajectory by: (i) the proposed method (altitude + range), (ii) UID method, (iii) altitude + UID, and (iv) P4P visual reference. A separate plot for each coordinate North, East, and Down (x, y, z) is presented.

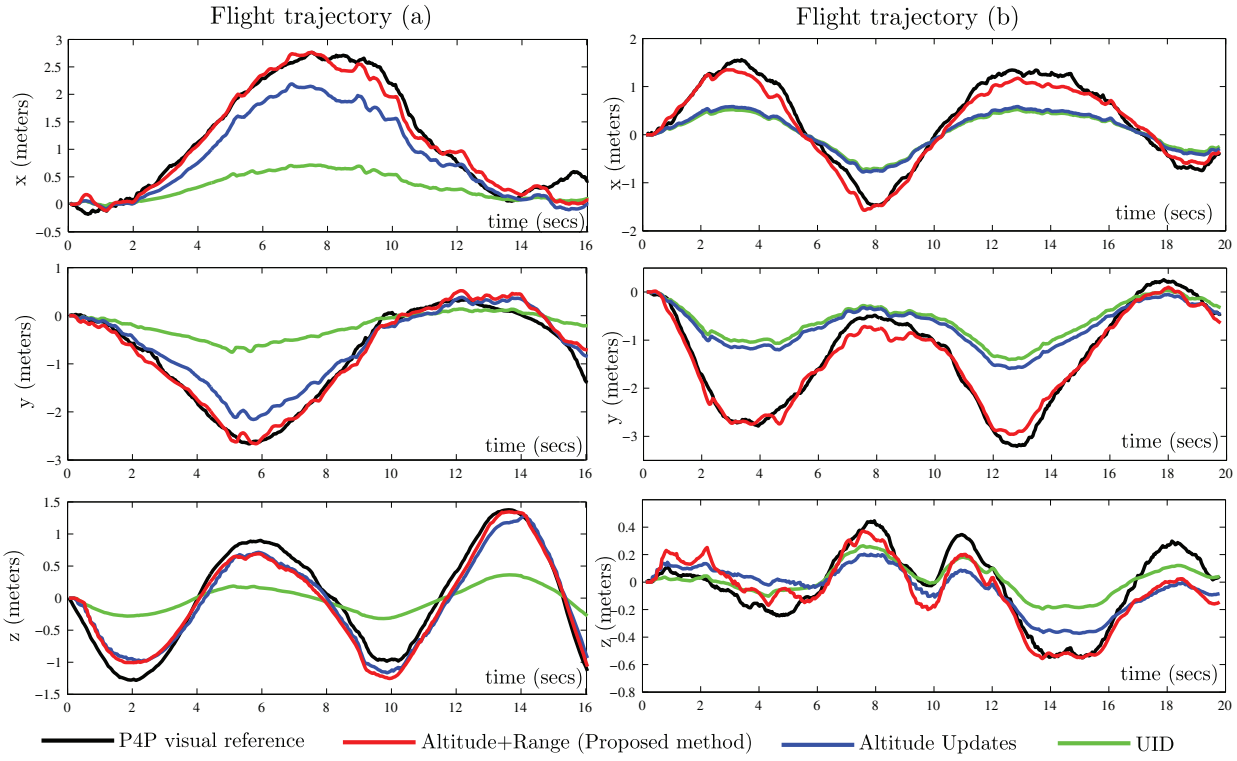

Figure 8. Estimated positions expressed in each coordinate north, east, and down (x, y, z).

Table 1. Results for flight trajectories 'a' and 'b'.

Method	Flight	NIF	NDF	FPF	TTE (s)	aMAE (m)
A + R	a	$126 \pm 2.5\sigma$	$100 \pm 7.3\sigma$	$22.9 \pm 4.2\sigma$	$268 \pm 16\sigma$	$.28 \pm .12\sigma$
A + UID	a	$178 \pm 8.6\sigma$	$148 \pm 8.7\sigma$	$26.1 \pm 4.6\sigma$	$336 \pm 19\sigma$	$.51 \pm .20\sigma$
UID	a	$154 \pm 8.3\sigma$	$121 \pm 10\sigma$	$28.0 \pm 4.5\sigma$	$323 \pm 9\sigma$	$.50 \pm .36\sigma$
A + R	b	$76.8 \pm 5.6\sigma$	$27.6 \pm 1.8\sigma$	$35.2 \pm 10.9\sigma$	$170 \pm 21\sigma$	$.25 \pm .09\sigma$
A + UID	b	$89.6 \pm 4.2\sigma$	$32.4 \pm 3.5\sigma$	$43.2 \pm 11.3\sigma$	$197 \pm 11\sigma$	$.87 \pm .54\sigma$
UID	b	$84.2 \pm 4.9\sigma$	$27.6 \pm 4.0\sigma$	$43.5 \pm 11.7\sigma$	$192 \pm 9\sigma$	$1.43 \pm .73\sigma$

UID: undelayed inverse depth; NIF: number of the features initialized into the system state; NDF: number of features deleted from the system state; FPF: number of features been tracked at each frame; TTE: total time of execution; aMAE: average mean absolute error.

In this comparison study, the results were obtained averaging 10 executions of each method. It is important to note that those averages are computed because the methods are not deterministic since the search and detection of new visual features points is conducted in a random manner over the images.

Table 1 summarizes the results obtained in the foregoing experiment. In the table, the proposed method is indicated by (A + R), the UID + Altitude variant is indicated by (A + UID), and the undelayed inverse depth method by (UID). The following results have been computed for each method: (i) number of the features initialized into the system state (NIF), (ii) number of features deleted from the system state (NDF), (iii) number of features been tracked at each frame (FPF), (iv) total time of execution (TTE) and (v) average mean absolute error (aMAE) of the vehicle position. For computing the aMAE, the P4P trajectory has been used as an independent reference of the vehicle position.

From these results it can be concluded that

- For flight (a): Taking the UID trajectory as an unscaled reference, it can be appreciated that the inclusion of altitude measurements (A + UID) can be useful by itself in order to recover the metric

scale of the estimates. Note that for this flight, there is a considerable variation in altitude along the trajectory. Even though an acceptable result was obtained with the (A + UID) method, the result obtained with the proposed method (A + R) was considerable better.

- For flight (b): The solely inclusion of altitude measurements (A + UID) contributes little in terms of improvements of the recovery of the metric scale. This result can be explained because for this flight there is little variation in altitude along the trajectory, (especially at the beginning). On the other hand, with the proposed method (A + R), the actual trajectory was also successfully recovered.
- In both flights, with the proposed method (A + R) fewer features were initialized into the system state than the other approaches, because features are initialized only in frames where range measurements are available. It is advantageous if the system performs well with few features because a smaller computational cost is need.

It is interesting to note that the theoretical findings presented in section *Observability of metric scale* are well supported by the empirical results obtained with real data. Also, it is shown that the proposed method is capable of working with the data obtained from low-cost sensors, in order to estimate the flight trajectory of an MAV.

Conclusions

Using monocular SLAM, an MAV can navigate relying on visual information in environments where GPS is not available. Compared with other approaches, the monocular vision has advantages in terms of weight, space, power-saving, and scalability, but it introduces some technical difficulties as the impossibility of directly recovering the metric scale of the world.

In this work, a theoretical analysis has been carried out in order to study the observability of a monocular SLAM system applied to an MAV. According to the theoretical findings, under certain conditions the metric scale can become observable by the inclusion of altitude measurements into the system. Also, it was found that depth measurement of even a single landmark can improve the observability of the system.

Based on the theoretical findings, a novel monocular SLAM system is proposed. In this case, for recovering the metric scale of the world, two extra sources of sensory information have been considered: (i) an ultrasonic range finder is used for computing the approximate depth of features whenever is possible and (ii) a barometer is used for updating the system with information about the local altitude of the MAV.

In order to validate the proposal, experimental results have been obtained using real data acquired by sensors onboard of a custom-built quadrotor. Based on these results, it is shown that by means of the proposed system, the MAV is able to recover its flight trajectory as well as a map of the environment using only those sensors. The proposed method is robust enough to be implemented with low-cost hardware.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research has been funded by Spanish Science ministry project DPI2016-78957-R (ColRobTransp).

References

1. Munguia R, Urzua S, Bolea Y, et al. Vision-based SLAM system for unmanned aerial vehicles. *Sensors* 2016; 16: 372.
2. Zhao H, Chiba M, Shibasaki R, et al. SLAM in a dynamic large outdoor environment using a laser scanner. In: *IEEE international conference on robotics and automation*, Pasadena California, 19–23 May 2008, pp. 1455–1462.
3. Bosse M and Roberts J. Histogram matching and global initialization for laser-only SLAM in large unstructured environments. In: *IEEE international conference on robotics and automation* 2007. pp. 4820–4826.
4. Fallon MF, Folkesson J, McClelland H, et al. Relocating underwater features autonomously using sonar-based SLAM. *IEEE J Oceanic Eng* 2013; 38: 500–513.
5. Yap TN and Shelton CR. SLAM in large indoor environments with low-cost, noisy, and sparse sonars. In: *IEEE international conference on robotics and automation*; 2009. pp. 1395–1401.
6. Luo RC, Huang CH and Huang CY. Search and track power charge docking station based on sound source for autonomous mobile robot applications. In: *IEEE/RSJ international conference on intelligent robots and systems* 2010. pp.1347–1352.
7. Kleiner A, Dornhege C and Dali S. Mapping disaster areas jointly: RFID-Coordinated SLAM by Hurnans and Robots. In: *IEEE international workshop on safety, security and rescue robotics*; 2007. pp.1–6.
8. Lemaire T, Berger C, Jung IK, et al. Vision-based SLAM: stereo and monocular approaches. *Int J Comput Vision* 2007; 74: 343–364.
9. Davison AJ, Reid ID, Molton ND, et al. MonoSLAM: real-time single camera SLAM. *IEEE Trans Pattern Anal Mach Intell* 2007; 29: 1052–1067.

10. Strasdat H, Montiel JMM and Davison AJ. Real-time monocular SLAM: why filter? In: *IEEE international conference on robotics and automation*; 2010. pp.2657–2664.
11. Mirzaei FM and Roumeliotis SI. A Kalman filter-based algorithm for IMU-camera calibration: observability analysis and performance evaluation. *IEEE Trans Rob* 2008; 24: 1143–1156.
12. Weiss S, Scaramuzza D and Siegwart R. Monocular SLAM based navigation for autonomous micro helicopters in GPS-denied environments. *J Field Rob* 2011; 28: 854–874.
13. Forster C, Lynen S, Kneip L, et al. Collaborative monocular SLAM with multiple micro aerial vehicles. In: *IROS*. IEEE; 2013. pp.3962–3970.
14. Celik K and Somani AK. Monocular vision SLAM for indoor aerial vehicles. *J Electr Comput Eng* 2013; 15. Article ID 374165. DOI: 10.1155/2013/374165.
15. Nutzi G, Weiss S, Scaramuzza D, et al. Fusion of IMU and vision for absolute scale estimation in monocular SLAM. *J Intell Rob Syst* 2011; 61: 287–299.
16. Wang CL, Wang TM, Liang JH, et al. Bearing-only visual SLAM for small unmanned aerial vehicles in GPS-denied environments. *Int J Autom Comput* 2014; 10: 387–396.
17. Chowdhary G, Johnson EN, Magree D, et al. GPS-denied indoor and outdoor monocular vision aided navigation and control of unmanned aircraft. *J Field Rob* 2013; 30: 415–438.
18. Hopkins RE, Barbour NM, Gustafson DE, et al. Miniature inertial and augmentation sensors for integrated inertial/GPS based navigation applications. DTIC Document: ADA581022, 2010.
19. Montiel JMM, Civera J and Davison A. Unified inverse depth parametrization for monocular SLAM. In: *Proceedings of the robotics: science and systems conference*, Philadelphia Pennsylvania, 16–19 August 2006.
20. Munguia R, Castillo-Toledo B and Grau A. A robust approach for a filter-based monocular simultaneous localization and mapping (SLAM) system. *Sensors* 2013; 13: 8501–8522.
21. Civera J, Davison AJ and Montiel J. Inverse depth parametrization for monocular SLAM. *EEE Trans Rob* 2008; 24: 932–945.
22. Civera J, Davison AJ and Montiel JMM. Dimensionless monocular SLAM. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2007, pp.412–419.
23. Munguia R and Grau A. A practical method for implementing an attitude and heading reference system. *Int J Adv Rob Syst* 2014; 11. DOI: https://doi.org/10.5772/58463.
24. Euston M, Coote P, Mahony R, et al. A complementary filter for attitude estimation of a fixed-wing UAV. In: *IEEE/RSJ international conference on intelligent robots and systems*; 2008. pp.340–345.
25. Hermann R and Krener AJ. Nonlinear controllability and observability. *IEEE Trans Autom Control* 1977; 22: 728–740.
26. Slotine JE and Li W. *Applied nonlinear control*. Englewood Cliffs: Prentice-Hall, 1991.
27. Bouguet JY. Camera calibration toolbox for Matlab. www.vision.caltech.edu/bouguetj/calib_doc (2008, accessed 17 May 2017).
28. Durrant-Whyte H and Bailey T. Simultaneous localization and mapping: part I. *IEEE Rob Autom Mag* 2006; 13: 99–110.
29. Bailey T and Durrant-Whyte H. Simultaneous localization and mapping (SLAM): part II. *IEEE Rob Autom Mag* 2006; 13: 108–117.
30. Rosten E and Drummond T. Fusing points and lines for high performance tracking. *IEEE Int Conf Comput Vision* 2005; 2: 1508–1511.
31. Community OS. Ardupilot, www.ardupilot.com (2015, accessed 17 May 2017).
32. Community OS. Ardupilot, http://qgroundcontrol.org/mavlink (2015, accessed 17 May 2017).
33. Chatterjee C and Roychowdhury VP. Algorithms for coplanar camera calibration. *Mach Vision Appl* 2000; 12: 84–97.