Vision-aided Inertial Navigation Using Planar Terrain Features

Ghazaleh Panahandeh and Magnus Jansson
KTH Royal Institute of Technology, ACCESS Linnaeus Center
Stockholm, Sweden
{ghazaleh.panahandeh, magnus.jansson}@ee.kth.se

Abstract—The idea is to implement a vision-aided inertial navigation system (INS) for estimating inertial measurement unit (IMU)-camera ego-motion. The system consists of a ground facing monocular camera mounted on an IMU that is observing ground plane feature points. The motion estimation procedure is through tracking detected corresponding feature points between two successive image frames. The main contribution of this paper is a novel closed-form measurement model based on the image data and IMU output signals. In contrast to existing methods, our algorithm is independent of the underlying vision algorithm such as image motion estimation or optical flow algorithms for camera motion estimation. Additionally, unlike the visual-SLAM based methods, our approach is not based on data association. The algorithm has been implemented using an Extended Kalman filter (EKF), which propagates the current and the last state of the system updated in the previous measurement state. Simulation results show that the introduced method is persistent to the level of the noise and works well even with few numbers of features.

Keywords—vision-aided INS; ego-motion; augmented EKF;

I. INTRODUCTION

Numerous approaches towards aiding INS with applied computer vision algorithms for mobile robots exist. They rely on tracking distinguishable features across multiple images for motion estimation. For instance, in [1] a mobile robot in rough outdoor terrain is localized using the bundle adjustment method over multiple frames from stereo cameras. Using a short-term visual memory of past views of a monocular camera, in [2] a camera-aided INS is implemented based on the epipolar points constraints between the current and past views. Similarly in [3], imposing the vanishing points constraints between a pair of images, 6-DoF pose change of a monocular camera was estimated for mobile navigation. The above methods are based on using a forward-looking camera for capturing salient image features such as edges, markers, and lines in the scenes. However, there are feature poor regions where they cannot be used such as cluttered environments with complex backgrounds or dynamic objects. Ground-facing cameras avoid dynamic backgrounds, which provide difficulties for forward looking vision systems. For instance in [4], the velocity of mobile robots is measured by the Kalman filter (KF) integrated optical flow method for a downward-looking camera. In [5], an image based motion estimation algorithm together with IMU measurement is used for relative pose estimation in an enhanced KF framework. Likewise in [6], an algorithm for onboard motion estimation is presented which is based on automatic feature tracking between a pair of descent camera images followed by two frame motion estimation and scale recovery using laser altimetry data. In [7], the integration of GPS with an IMU-camera in hand held pedestrian navigation is introduced with the camera pointing approximately towards the ground. Using a computer vision algorithm, 3-D camera frame translation is derived that provides the velocity of the camera in the measurement model; although, the transition is scaled by the height of the camera above the ground, the effect of scaling is reflected in the measurement model as well as the state parameters. A planar ego-motion estimation method together with ground layer detection has been studied in [8]. Besides, they virtually rotate the forward-looking camera to the downward-looking pose in order to exploit the fact that the vehicle motion is roughly constrained to be planar motion on the ground. In [9], a feature point matching strategy and motion recovery has been described for vehicle navigation. To keep the motion of the vehicle parallel to ground plane, image plane are transformed to linearly estimate the ground floor feature points position.

In this paper, a novel closed-form model for vision-aided inertial navigation is derived which, unlike the existing methods, is not based on the underlying vision algorithm for image motion estimation; additionally, the problem of scale estimation in the 3-D camera translation is solved. The proposed algorithm is based on tracking feature points on the image plane taken by a monocular camera whose optical axis is assumed to be parallel with the gravity field. A modified EKF is used to estimate the motion parameters of the system.

The paper is organized as follows. The process and measurement models, which lead to the state-space equations of the system, are derived in Section II. The structure of the used EKF algorithm is presented in Section III. In Section V, the performance of the proposed method and simulation results in different scenarios are examined. Finally, the conclusion of the study is summarized in Section VI.

In the following sections scalars are denoted by lowercase letters (s), vectors by bold letters (f), and matrices by bold capitals (K).
ground floor, which are detected using the SURF method, are connected by colored lines between two image planes.

II. SYSTEM DESCRIPTION

The goal of our proposed algorithm is to estimate the position and orientation of the body frame \( \{ b \} \) in the navigation frame \( \{ n \} \), where the IMU is rigidly mounted to a camera. In order to simplify the treatment with different coordinate frames, we assume the IMU coordinate frame \( \{ i \} \) is located in the center of the \( \{ b \} \) frame and the translation between the camera coordinate frame \( \{ c \} \) and the IMU coordinate frame are known [10], [11]. In fact, the IMU is aided with the camera while the camera is capturing feature points located in the ground plane; however, no prior knowledge about the height of the camera is assumed (Figure 1). In order to estimate parameters in the EKF, we first describe the total state vector as

\[
x^{\text{ins}} = [p_b^\top, v_b^\top, \mathbf{q}_b^\top, \mathbf{f}_b^\top, \mathbf{w}_b^\top]^\top
\]

where the position and velocity of the IMU in the navigation frame are denoted by \( p_b^n \) and \( v_b^n \), respectively, \( \mathbf{q}_b^n \) is the unit quaternion representing the rotation from body frame to navigation frame, and finally \( \mathbf{f}_b \) and \( \mathbf{w}_b \) are the bias vectors affecting the accelerometer and gyroscope measurements, respectively. The time evolution of the INS state can be described by

\[
\begin{align*}
\dot{p}_b^n(t) &= \frac{1}{2} \Omega(\omega(t)) \mathbf{q}_b^n(t) \\
\dot{v}_b^n(t) &= a^n(t) \\
\dot{\mathbf{q}}_b(t) &= n_{\delta \omega}(t) \\
\dot{n}_{\delta \theta}(t) &= n_{\delta \phi}(t)
\end{align*}
\]

where \( \Omega(\omega) \) is the skew-symmetric matrix representation of the cross product operation [12]. The IMU and gyroscope bias increments, \( n_{\delta \phi} \) and \( n_{\delta \omega} \), are modeled as white Gaussian noises. The output measurement signals of the accelerometer \( \mathbf{f}_m \) and the gyroscope \( n_m \) are modeled as

\[
\begin{align*}
\mathbf{f}_m(t) &= R_k^b(a^n(t) - g^n) + b_a(t) + n_t(t) \\
n_m(t) &= \omega(t) + b_\omega(t) + n_\omega(t)
\end{align*}
\]

where \( R_k^b \) is the direction-cosine matrix [12] which can be used alternatively for attitude representation, \( n_t \) and \( n_\omega \) are temporally uncorrelated zero-mean noise process, respectively, and \( g^n \) is the gravitational acceleration expressed in the navigation frame. The discrete-time error state space model is described by

\[
\begin{align*}
\delta \dot{p}_{b,k+1}^n &= \delta p_{b,k}^n + dt \delta v_{b,k}^n \\
\delta v_{b,k+1}^n &= \delta v_{b,k}^n + dt (\dot{\mathbf{R}}_b^c \mathbf{f}_{m,k}^c) - \delta \theta_k + dt \dot{\mathbf{R}}_b^c (\delta \mathbf{f}_k^n + \mathbf{n}_k^n) \\
\delta \theta_{k+1} &= \delta \theta_k - dt \mathbf{R}_b^c R_{f,k}^e (\delta \omega_k^n + \mathbf{n}_{\delta \omega,k}^n) \\
\delta \mathbf{f}_{k+1}^c &= \delta \mathbf{f}_k^c + dt \mathbf{n}_{\delta f,k} \\
\delta \omega_{k+1}^n &= \delta \omega_{k}^n + dt \mathbf{n}_{\delta \omega,k}
\end{align*}
\]

where \( \mathbf{R}_b^n \) is the estimated rotation matrix. The equations in (3) have been derived based on the standard additive error definition for the position, velocity, and biases (\( \omega \approx x + \delta x \)) and quaternion error for the rotational angles \( \theta \) (\( \delta \mathbf{q} \approx \begin{bmatrix} 1 & -2\delta x^2 & 0 \\ 2\delta x & 1 & -2\delta y \\ 0 & 2\delta y & 1 \end{bmatrix} \)). The advantage of the quaternion error definition is the direct use of error angle vectors \( \delta \theta \) instead of \( \delta \mathbf{q}^n \).

Accordingly, the error state vector can be written as

\[
\delta x^{\text{ins}} = [\delta p_b^n, \delta v_b^n, \delta \theta, \delta \mathbf{f}_b, \mathbf{w}_b]^\top
\]

Hence, the state-space model of the discrete time process is

\[
\delta x_{k+1} = f_k(\delta x_k, \mathbf{n}_k) \in \mathbb{R}^{15}
\]

where the process noise \( \mathbf{n}_k = [n_{\delta f,k}^c, n_{\delta \omega,k}^c, n_{\delta \theta,k}^c, \mathbf{n}_{\delta \omega,k}^n] \) is assumed to be time invariant, with the covariance matrix \( Q \).

A. Projection From 3-D to 2-D

Given a static point in the navigation frame (\( \pi^n \in \mathbb{R}^3 \)), its position relative to the camera frame can be represented as

\[
\begin{bmatrix}
\pi_{x}^c \\
\pi_{y}^c \\
\pi_{z}^c
\end{bmatrix} = \mathbf{R}_b^c (\begin{bmatrix}
\pi_{x}^n \\
\pi_{y}^n \\
\pi_{z}^n
\end{bmatrix} - \mathbf{p}_c^b)
\]

where \( \mathbf{R}_b^c \) and \( \mathbf{p}_c^b \) are the camera to body rotation and translation, respectively. The transformation of \( \pi^c \), while the IMU-camera is moving over the ground plane, between two time instances \( t_{k-m} \) and \( t_k \) can be described by

\[
\pi_{k}^c = \mathbf{R}_{k-m}^c \pi_{k-m}^c + \mathbf{t}_{k-m}^c
\]

where \( \mathbf{R}_{k-m}^c \) and \( \mathbf{t}_{k-m}^c \) are the relative camera rotation and translation, respectively. In fact, this transformation is a function of state variables as

\[
\begin{align*}
\mathbf{R}_{k-m}^c &= \mathbf{R}_c^b \mathbf{R}_{k-n}^c (\mathbf{R}_c^b \mathbf{R}_{n-k-m}^c)^\top \\
\mathbf{t}_{k-m}^c &= \mathbf{R}_c^b \mathbf{p}_{n,k-m}^c + \mathbf{R}_n^c \mathbf{p}_b^c - \mathbf{p}_{b,k}^n - \mathbf{R}_c^b \mathbf{p}_c^b
\end{align*}
\]

Using the pinhole camera model [13], the projection of \( \pi_{k-m}^c \) in the image plane is

\[
\mathbf{z}_{k-m} = \begin{bmatrix}
u_{k-m} \\
v_{k-m}
\end{bmatrix} + \mathbf{w}_{k-m} = \begin{bmatrix}
u_{k-m} \\
w_{k-m}
\end{bmatrix} = \begin{bmatrix}
u_{k-m} \\
w_{k-m}
\end{bmatrix}
\]

where \( \mathbf{w}_{k-m} \) is the feature-measurement noise with covariance matrix \( \mathbf{R}_{k-m} = \mathbf{K}_{k-m} \mathbf{K}_{k-m}^T + \mathbf{K}_{k-m} \mathbf{K}_{k-m}^T \), and \( \mathbf{K} \) is the camera intrinsic matrix, assumed to be known. In this study a simplified
model for the camera intrinsic matrix is considered as \( \mathbf{K} = \text{diag}(f, f, 1) \), where \( f \) is the focal length of the camera. Based on (8), \( \pi^c_{k-m} \) can be written as a function of its position in the image plane [13] as

\[
\pi^c_{k-m} = \frac{\pi^c_{z-k-m}}{f} \left[ \begin{array}{c} u_{k-m} \\ v_{k-m} \end{array} \right].
\] (9)

Replacing (9) in (6), \( \pi^c_k \) can be represented mainly as a function of state variables by

\[
\pi^c_k = \frac{\pi^c_{z-k-m}}{f} \left[ \begin{array}{c} u_{k-m} \\ v_{k-m} \end{array} \right] + t^c_{k-m} \] (10)

where \( \pi^c_{z-k-m} \) can be referred as the metric scale. Determining this parameter is not possible except using an extra sensor such as laser scanner, altimeter, or an extra camera. However, when the camera rotation relative to the navigation frame is planar the metric scale can be written as a function of the state variables. For the sake of simplicity, we consider the case when the camera optical axis is parallel to the gravity field as shown in Figure 1; so, it can be claimed that \( \pi^c_{z-k-m} = p^n_{c} \pi^c_m \) where \( \mathbf{p}^n = \mathbf{p}^n_0 + \mathbf{R}^n_{k-m} \mathbf{p}^b_0 \). Inserting this equality together with equation (7) in (10), \( \pi^c_k \) can be written as

\[
\pi^c_k = \frac{1}{f} R^c_{b,k-m} \left[ \begin{array}{c} u_{k-m} \\ v_{k-m} \end{array} \right] + t^c_{k-m}
\] (11)

\[
\left[ \begin{array}{c} 0 & 0 & 1 \end{array} \right] \left( \mathbf{p}^n_{b,k-m} + \mathbf{R}^n_{b,k-m} \mathbf{p}^b_0 \right) + R^c_{b,k-m} \mathbf{p}^b_0 \mathbf{p}^b_0 - R^c_{b} \mathbf{p}^c_0.
\]

Although, the requirement of such model is the downward-looking camera, the movement of the camera along the \( z \)-axis (height of the camera) is not limited to be fixed relative to the navigation frame. However, the limitation of using the downward-looking camera can be removed using a forward-looking camera which is virtually transformed [8].

B. Measurement model

When the camera along with the IMU is moving, the body frame angular velocity and specific force are measured by the high rate IMU. Meanwhile, the camera records images with a lower rate (every \( m \)th time instance); Therefore, the measurement model of the system based on equation (8) and (11) is

\[
\mathbf{z}_k = h_k(\mathbf{x}^\text{ins}_{k-m}, \mathbf{u}_{k-m}, \mathbf{v}_{k-m}, \mathbf{w}_k)
\] (12)

where \( \mathbf{x}^\text{ins}_{k-m} = \mathbf{Tx}^\text{ins}_{k-m} = [\mathbf{p}^n_{k-m,b} \mathbf{q}^n_{k-m,b} \mathbf{T}^\text{ins}_{k-m}]^\top \) is the truncated state of the system updated in the previous measurement. Such a structure in \( \mathbf{z}_k \) force the state vector of the system to be augmented each time that a measurement from the camera is received with a part of the current state.

III. Extended Kalman filter

The nonlinearity of the measurement model with respect to the state variables is the reason to use a nonlinear filter for estimation of the states. In this paper the EKF framework is used to linearize the measurement equation about the estimates for the variables [14]. Using the standard additive error \( \delta \mathbf{x} \sim \mathbf{x} + \delta \mathbf{x} \) for the positions, velocity, and biases; and \( \delta \mathbf{R}^n = (\delta \mathbf{R}^n_{b}) \mathbf{R}^n_{b} \) for the rotational Euler angles, the measurement Jacobian matrix is \( \mathbf{H}_k = [\mathbf{H}^\text{ins}_k \mathbf{H}_f] \). The elements in the Jacobian matrix \( \mathbf{H}_k \) are given in Appendix A.

For \( M \) matched features between two successive frames, the measurement vector and the Jacobian matrix should be augmented as \( \mathbf{z}_k = [\mathbf{z}^i_1, \ldots, \mathbf{z}^i_M]^\top \) and \( \mathbf{H}_k = [\mathbf{H}^\text{ins}_1, \ldots, \mathbf{H}^\text{ins}_M]^\top \) and \( \mathbf{H}_f = \text{diag}(\mathbf{H}_{f1}, \ldots, \mathbf{H}_{fM}) \). Therefore, the final representation of the residual for \( M \) detected feature points in two successive frames is

\[
\mathbf{r} = \mathbf{z}_k - \mathbf{z} \sim \mathbf{H}_k \mathbf{x}_k + \mathbf{w}
\] (13)

where \( \mathbf{x}_k = [\mathbf{x}^\text{ins}_{k-m}, \mathbf{u}^*_{k-m}, \mathbf{v}^*_{k-m}, \ldots, \mathbf{u}^*_{M}, \mathbf{v}^*_{M}]^\top \) is the augmented state vector of the system for \( M \) feature points. Assuming the measurement noise to be mutually uncorrelated, the covariance matrix of \( \mathbf{w} \) will be \( \mathbf{R} = \text{diag}(\mathbf{R}_i) \), \( i = 1, \ldots, M \). Based on (11), the measurement equation of the system (12) not only depends on the current state of the system (1) but also depends on the state and the feature points position in the previous measurement. This structure in the current measurement equation is the reason why an augmented structure in the state vector needs to be considered. Hence, similar to the implementation in [5], an augmented EKF is used to estimate the current state of the system (1). Let's consider \( \mathbf{x}^\text{ins}_{k-m} \) as the current state estimate at time \( k-m \), when the first image has been recorded, so the state augmented vector with a second copy of position and the rotational angles states is

\[
\delta \mathbf{x}^\text{ins}_{k-m} = \left[ \delta \mathbf{x}^\text{ins}_{k-m}^\top \delta \mathbf{z}^\text{ins}_{k-m}^\top \delta \mathbf{z}^\text{img}_{k-m}^\top \right]^\top
\] (14)

where \( \delta \mathbf{x}^\text{ins}_{k-m} \in \mathbb{R}^{2M} \) and contains the current feature-measurement noise (\( \mathbf{w}_k \)) in the shape of \( \delta \mathbf{z}^\text{ins}_{k-m} \). Since, at \( k-m \), both \( \delta \mathbf{z}^\text{ins}_{k-m} \) and \( \delta \mathbf{z}^\text{img}_{k-m} \) contain the same uncertainty in the positions and rotational angles, the covariance matrix for the augmented system will be

\[
\mathbf{P}_{k-m} = \left[ \begin{array}{cc} \mathbf{P}^\text{ins}_{k-m} & \mathbf{P}^\text{ins}_{k-m} \mathbf{P}^\text{ins}_{k-m}^\top \\ 0 & \mathbf{R}_{k-m} \end{array} \right]
\] (15)

where \( \mathbf{P}^\text{ins}_{k-m} \) is the covariance matrix for the error state at time \( k-m \). If the next measurement arrives at \( k \), during this interval only \( \delta \mathbf{z}^\text{ins}_{k-m} \) is propagated while the rest of the states remain stationary:

\[
\delta \mathbf{x}^\text{ins}_{k-m+1} = \left[ \begin{array}{c} \mathbf{F}^\text{ins}_{k-m+1} \mathbf{0} \\ \mathbf{0} \end{array} \right] \delta \mathbf{x}^\text{ins}_{k-m} + \left[ \begin{array}{c} \mathbf{G}^\text{ins}_{k-m+1} \mathbf{0} \\ \mathbf{0} \end{array} \right] \mathbf{n}_{k-m}
\]
where $F_{k-m+1}$ and $G_{k-m+1}$ are discrete time state and system noise propagation matrices, respectively [14]. Equivalently, the covariance of the augmented system after $m$ steps is
\[
P_{k|m} = \begin{bmatrix}
P_{k|m-k} & P_{m-k|m}F^T & 0 \\
P_{m-k|m}F & T & 0 \\
0 & 0 & R_{k-m}
\end{bmatrix}
\]
(16)

where $\tilde{F} = \prod_{i=m}^{k-1} F_{k-i}$ and $P_{k|m}$ is the propagated covariance of the evolving state at time $t_k$. An overview of the proposed algorithm is given in Algorithm 1.

**Algorithm 1** Proposed algorithm in the Extended Kalman filter framework, $N = 15 + 6 + 2M$

1: Initialize $x_0$ and $P_0$
2: for $k = 0$ end do
3: update the INS equations
4: if the measurement from the camera is ready then
5: $z_k = [z_k^1, z_k^2, \ldots, z_M]^T$
6: $R_k = \text{diag}(R_i) \% i \in 1, \ldots, M$
7: Compute $z_k$ as a function of the current states
8: Compute the Jacobian matrix $H_k$
9: $r_k = z_k - \hat{z}_k$
10: Form the augmented covariance matrix $P_{k|m}$
11: $K_k = P_{k|m}H^T(HP_{k|m}H^T + R_k)^{-1}$
12: $\delta z_k = K_k r_k$
13: $P_{k|m} = P_{k|m} - K_k HP_{k|m}$
14: $P_{imp} = P_{k}(1:15, 1:15)$
15: else
16: $P_{imp} = P_k$
17: end if
18: $P_{k+1} = FP_{imp}F^T + GQG^T$
19: Compute $\tilde{F}$
20: end for

**IV. Feature Extraction and Matching**

The image processing module of our algorithm contains feature selection and matching that is identifying distinguishable features between two successive image frames. To get closer to real time, the speeded up robust features (SURF) [15], which is invariance to image scale, viewing angle, and viewing condition, has been used to extract image descriptors from the query image and search for the best match with the previous image on the ground plane. The Matlab OpenSURF Toolbox v0.1c (Dirk-Jan Kroon) was used in our implementation. An initial test has been done in the Q-building on the KTH university campus. Despite of poor lighting situation in corridors and feature poor texture of plastic parquet flooring, the algorithm was capable to identify corresponding features in two successive images that were taken from the height of 1[m] (Figure 2).

**V. Simulation Results**

The estimated trajectory along with the ground truth is plotted in Figure 3. The figure shows that the vision-aided inertial system is able to follow the ground truth trajectory along the $x$, $y$, and $z$ well. Zero-mean white Gaussian noise with standard deviation of $\sigma_w = 2$ [pixel] is added to the projected image where the maximum number of tracked feature points between two successive images is 15 over the whole simulation time. Furthermore, the estimated errors of the position and velocity along with the ±σ levels are plotted in Figure 4. The σ-levels are computed from the corresponding diagonal elements of the filter’s error covariance matrix that provides a representation of its estimate uncertainty.

**VI. Conclusion**

An IMU-camera sensor fusion approach has been used to construct a vision-aided inertial navigation system by
tracking salient features of planar terrain. In the proposed algorithm a new measurement model has been derived for a downward-looking monocular camera. It can be alternatively used for a forward-looking camera that is calibrated relative to the ground plane or by virtually transforming a downward-looking camera. The proposed algorithm is not only capable of velocity estimation compared to optical flow methods but also accurate pose estimation. The problem of nonlinearity of the state space model of the system has been handled by using modified EKF to estimate the relative position of the system in the navigation frame.

REFERENCES


APPENDIX A.

Elements in the Jacobian matrix $\hat{H}_k$:

$$H_{\text{mx}} = J_{\theta_k} J_{p_k} 0_{3 \times 3} J_{\theta_k} 0_{1 \times 6} J_{p_{k-m}} J_{\theta_{k-m}}$$

where

$$J_{\theta_k} = \frac{f}{\pi_{c,k}} \begin{bmatrix} 1 & 0 & -\pi_{c,k} \hat{R}^b_h \hat{R}^b_{n,k-m}^\top \left[ z_{k-m} \right]^T \left[ f \right] \end{bmatrix}^T \times$$

$$J_{p_k} = \hat{R}^b_h \hat{R}^b_{n,k-m}^\top \left[ z_{k-m} \right]^T \left[ f \right] \times$$

$$J_{\theta_{k-m}} = -\frac{\pi_{c,k-m}}{f} \hat{R}^b_h \hat{R}^b_{n,k-m}^\top \left[ z_{k-m} \right]^T \left[ f \right] \times$$

$$-\frac{1}{f} \hat{R}^b_h \hat{R}^b_{n,k-m}^\top \left[ z_{k-m} \right]^T \left[ f \right] \times$$

$$-\hat{R}^b_h \hat{R}^b_{n,k-m}^\top \left[ f \right] \times$$

$$J_{p_{k-m}} = \frac{\pi_{c,k-m}}{f} \hat{R}^b_h \hat{R}^b_{n,k-m}^\top \left[ f \right] \times$$

$$J_{u_{k-m} \cdot v_{k-m}} = \frac{\pi_{c,k-m}}{f} \hat{R}^b_h \hat{R}^b_{n,k-m}^\top \left[ I_2 \right]$$

with

$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$