Sensors **2013**, *13*, 1919-1941; doi:10.3390/s130201919

OPEN ACCESS

sensors

ISSN 1424-8220
www.mdpi.com/journal/sensors

Article

Extended Kalman Filter-Based Methods for Pose Estimation Using Visual, Inertial and Magnetic Sensors: Comparative Analysis and Performance Evaluation

Gabriele Ligorio * and Angelo Maria Sabatini

The Institute of BioRobotics, Scuola Superiore Sant'Anna, Piazza Martiri della Libertà 33, 56124 Pisa, Italy; E-Mail: angelo.sabatini@sssup.it

* Author to whom correspondence should be addressed; E-Mail: g.ligorio@sssup.it;
Tel.: +39-050-883-415; Fax: +39-050-883-101.

Received: 12 December 2012; in revised form: 24 January 2013 / Accepted: 26 January 2013 / Published: 4 February 2013

Abstract: In this paper measurements from a monocular vision system are fused with inertial/magnetic measurements from an Inertial Measurement Unit (IMU) rigidly connected to the camera. Two Extended Kalman filters (EKFs) were developed to estimate the pose of the IMU/camera sensor moving relative to a rigid scene (ego-motion), based on a set of fiducials. The two filters were identical as for the state equation and the measurement equations of the inertial/magnetic sensors. The DLT-based EKF exploited visual estimates of the ego-motion using a variant of the Direct Linear Transformation (DLT) method; the error-driven EKF exploited pseudo-measurements based on the projection errors from measured two-dimensional point features to the corresponding three-dimensional fiducials. The two filters were off-line analyzed in different experimental conditions and compared to a purely IMU-based EKF used for estimating the orientation of the IMU/camera sensor. The DLT-based EKF was more accurate than the error-driven EKF, less robust against loss of visual features, and equivalent in terms of computational complexity. Orientation root mean square errors (RMSEs) of 1° (1.5°), and position RMSEs of 3.5 mm (10 mm) were achieved in our experiments by the DLT-based EKF (error-driven EKF); by contrast, orientation RMSEs of 1.6° were achieved by the purely IMU-based EKF.

Keywords: sensor fusion; extended Kalman filtering; inertial/magnetic sensing; monocular vision; ego-motion

1. Introduction

Sensor fusion methods combine data from disparate sources of information in a way that should ideally give better performance than that achieved when each source of information is used alone. The design of systems based on sensor fusion methods requires the availability of complementary sensors in order that the disadvantages of each sensor are overcome by the advantages of the others. An interesting application niche for sensor fusion—the one dealt with in this paper—is motion tracking. None of the several existing sensor technologies, taken alone, can meet the desired performance specifications, especially when motion is to be tracked without restrictions in space and time [1]. Vision and inertial/magnetic sensors are considered in this regard a particularly useful combination for developing a sense of position (localization) and motion, which is critically important in several technical fields, including augmented reality [2,3], robotics [4–7] and human machine interfaces [8].

Vision-based tracking systems can accurately track the relative motion between the camera and objects within its field of view (FOV) by measuring the frame-by-frame displacements of selected features, such as points or lines [9]. The camera pose relative to the scene can be estimated in all six degrees of freedom (DOFs) by using a stereo-camera system or by incorporating some a priori knowledge of the scene when a monocular system is used. The information provided by finding and associating image points of interest through a monocular video stream (monocular visual tracking) can be used to estimate the camera orientation relative to an absolute reference frame. The concurrent estimation of environment structure and motion allows to recover the perception of depth, otherwise lost from a single perspective view, using multiple images taken from different viewpoints [9]. The main shortcoming of vision-based tracking systems is the slow acquisition rate, which is due to both the physics of the image acquisition process and the computational workload of the computer-vision algorithms, especially those used to detect the visual features in each image frame. The consequence is that vision-based tracking systems lack robustness against fast motion dynamics, which may easily lead to loss of visual features. Another difficulty with vision-based tracking systems is that the line of sight between the camera and objects within its FOV must be preserved as much as possible, in other words vision-based tracking systems are severely prone to problems of occlusions.

Inertial-based tracking systems integrate Inertial Measurement Units (IMUs) that incorporate accelerometers and gyroscopes for measuring translational accelerations and angular velocities of the objects they are affixed to with high sampling rates; this feature makes them ideally suited to capture fast motion dynamics. Being internally referenced and immune to shadowing and occlusions, inertial sensors can track body motion, in principle, without restrictions in space. Unfortunately, measurements of linear accelerations and angular velocities are affected by time-varying bias and wideband measurement noise of inertial sensors. Accurate estimates of body orientation in the three-dimensional (3D) space can be produced using quite complex filtering algorithms, sometimes with the addition of magnetic sensors that sense the Earth's magnetic field to help producing drift-free heading estimates [10]; conversely, the 3D body position can be accurately estimated in tracking systems operating in a single IMU configuration only within temporally limited intervals of time, unless specific motion constraints are known and exploited to mitigate the double-time integration errors of gravity-compensated measured accelerations. The latter approach has been successfully implemented in strap-down inertial navigation systems (INS) for applications of pedestrian navigation [11,12].

Fusing visual and inertial/magnetic measurements can therefore yield, in principle, a tracking system for pose estimation in all six DOFs that retains, at the same time, the long-term stability and the accuracy of a vision-based tracking system with the short-term robustness and promptness of response typical of an INS [13]. Two main approaches have been tried to exploit the complementary properties of visual and inertial sensors, namely the loosely coupled approach and the tightly coupled approach [13]. In the loosely coupled approach [14–16], the vision-based tracking system and the INS exchange information each other, while the sensor data processing takes place in separate modules. The information delivered by the IMU can be used to speed up the tracking task of the features by predicting their locations within the next frame; in turn, data from the visual sensor allows updating the calibration parameters of inertial sensors. Conversely, in the tightly coupled approach all measurements, either visual or inertial, are combined and processed using a statistical filtering framework. In particular, Kalman filter-based methods are the preferred tool to perform sensor fusion [2,17,18].

In this paper the problem of estimating the ego-motion of a hand-held IMU-camera system is addressed. The presented development stems from our ongoing research on tracking position and orientation of human body segments for applications in telerehabilitation. While orientation tracking can be successfully performed using EKF-based sensor fusion methods based on inertial/magnetic measurements [10,19,20], position tracking requires some form of aiding [21].

A tightly coupled approach was adopted to the design of a system in which pose estimates were derived from observations of fiducials. Two EKF-based sensor fusion methods were developed that built somewhat upon the approaches investigated in [2,18], respectively. They were called DLT-based EKF (DLT: Direct Linear Transformation) and error-driven EKF. Their names were intended to denote the different use made of visual information available from fiducials: the visually estimated pose produced by the DLT method was directly delivered to the DLT-based EKF, while in the error-driven EKF the visual measurements were the difference between the measured and predicted location of the fiducials in the image plane. In each filter 2D frame-to-frame correspondences were established by a process of model-based visual feature tracking: a feature was searched within a size-variable window around its predicted location, based on 3D known coordinates of fiducials and the *a priori* state estimate delivered by the EKF. Moreover, the visual measurement equations were stacked to the measurement equations for the IMU sensors (accelerometer and magnetic sensor), and paired to the state transition equation, where the state vector included quaternion of rotation, position and velocity of the body frame relative to the navigation frame.

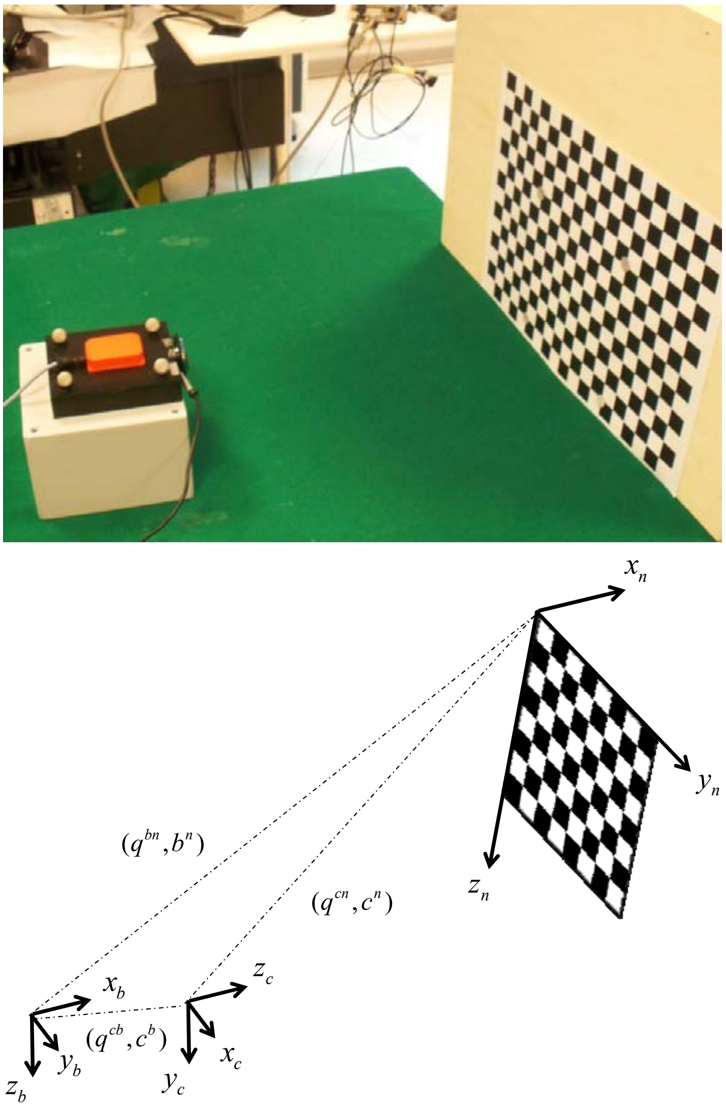
The main contributions of this paper are: (a) the comparative analysis and performance evaluation of the two different forms of visual aiding—the study was extended to the case when visual and inertial/magnetic measurements were used alone; (b) the investigation of the role played by magnetic sensors and related measurements of the Earth's magnetic field for heading stabilization, never attempted before in research on visuo-inertial integration (to the best of our knowledge). This paper is organized as follows: Section 2 reports the description of our experimental setup and a detailed mathematical analysis of the filtering methods. Main results achieved so far are presented in Section 3 and then discussed in Section 4. Finally, we offer concluding remarks and perspectives for our future work in Section 5.

2. Methods

We introduce the reference frames that are used in the experimental setup shown in Figure 1:

- Navigation frame {**n**}—this is the frame in which the coordinates of the corner points of a chessboard are known and the Earth's gravity and magnetic fields are assumed known, or measurable. The goal of the sensor fusion methods is to estimate the pose of the IMU case, namely the body pose, in {**n**}.
- Body frame {**b**}—this frame is attached to the IMU case, and the inertial and magnetic measurements delivered by the IMU are resolved in {**b**}.
- Camera frame {**c**}—this frame is attached to the camera, with its origin located in the camera optical center and the Z-axis pointing along the optical axis; although the camera is rigidly connected with the IMU, {**c**} is different from {**b**}.
- Image frame {**i**}—the 2D coordinate frame of the camera images; it is located in the image plane, which is perpendicular to the optical axis.

Figure 1. The camera is rigidly attached to the same support where the IMU sensor case is also attached. The axes of the three frames {**n**}, {**c**} and {**b**} are also drawn.



The following notation is used to express the relation between two frames, for instance $\{\mathbf{c}\}$ and $\{\mathbf{b}\}$: $\boldsymbol{R}^{cb}$ and $\overline{\boldsymbol{q}}^{cb} = [(\boldsymbol{q}^{cb})^T\ q_4^{cb}]^T$ denote, respectively, the rotation matrix and the quaternion from $\{\mathbf{b}\}$ to $\{\mathbf{c}\}$ ($\boldsymbol{q}^{cb}$ is the vector part and q_4^{cb} is the scalar part of $\overline{\boldsymbol{q}}^{cb}$, [22]); $\boldsymbol{b}^c$ represents the position of $\{\mathbf{b}\}$ relative to $\{\mathbf{c}\}$.

Figure 1 shows the sensor unit assembly and the chessboard. The sensor unit assembly contains one web-cam and one IMU; they are housed in a plastic box and are rigidly connected to each other. The visual sensor is a Microsoft web-cam with resolution 640 × 480 that acquires black-and-white visual images at approximately 30 fps; the images are transferred to the host computer via a USB port. The time elapsed between the time instant when the acquisition process starts and the time instant when a new image frame is available is returned together with the visual data.

The IMU is an MTx orientation tracker (Xsens Technologies B.V., Enschede, The Netherlands) equipped with one tri-axial accelerometer, one tri-axial gyroscope and one tri-axial magnetic sensor, with mutually orthogonal sensitive axes; the raw sensory data are delivered to the host computer at 100 Hz via another USB port. Both the camera and the IMU are electrically synchronized to an optical motion analysis system Vicon 460 equipped with six infrared (IR) cameras running at 100 Hz. The 3D coordinates of eight IR-reflective markers are acquired. Four markers (diameter: 15 mm) are located at the corners of the plastic box housing the sensor unit assembly, and four markers of the same diameter are located on the chessboard plane, where they are used for capturing the 3D coordinates of four black-and-white extreme corners of the chessboard. Since the size of the chessboard printed on an A3 sheet of paper is known, the 3D coordinates resolved in $\{\mathbf{n}\}$ of each black-and-white corner of the chessboard are easily determined.

The ancillary laboratory frame where the 3D positions of the markers are given is used to compute the transformation from $\{\mathbf{b}\}$ to $\{\mathbf{n}\}$, yielding the reference data $\mathbf{R}_{\mathrm{ref}}^{nb}$ or $\overline{\boldsymbol{q}}_{\mathrm{ref}}^{nb}$, and $\boldsymbol{b}_{\mathrm{ref}}^{n}$ that are needed for assessing the performance of the proposed sensor fusion methods. As for the IMU-camera relative pose calibration problem, namely the estimation of the rigid body transformation from $\{\mathbf{c}\}$ to $\{\mathbf{b}\}$, $\mathbf{R}^{cb}$or $\overline{\boldsymbol{q}}^{cb}$ are determined using the method proposed in [23]; the translation vector $\boldsymbol{b}^c$ is determined using a ruler, since accurate knowledge of this quantity is not critically important, especially when tracking slow motions [2].

2.1. Purely IMU-Based Method of Orientation Estimation

The purely IMU-based method for determining the IMU orientation relative to $\{\mathbf{n}\}$ revolves around the EKF developed in [10]. The major difference is that neither gyro bias nor magnetic distortions are included in the state vector for self-compensation purposes: the state vector $\boldsymbol{x}_{\mathrm{R}\,k} = \boldsymbol{x}_{\mathrm{R}}(t_k)$ is simply composed of the quaternion $\overline{\boldsymbol{q}}^{nb}$ sampled at the time instants t_k. The suffix R stands for rotation, to indicate the components of the state vector that describe the rotational behavior of the IMU-camera sensor unit assembly relative to $\{\mathbf{n}\}$, see below. The angular velocity $\boldsymbol{\omega} = [p\ q\ r]^T$ measured by the gyroscopes is used to update the state vector according to the state-transition model:

$$\boldsymbol{x}_{\mathrm{R}\,k} = \mathbf{F}_{\mathrm{R}\,k-1}\,\boldsymbol{x}_{\mathrm{R}\,k-1} + \boldsymbol{w}_{\mathrm{R}\,k-1} \tag{1}$$

The rotational state transition matrix $\mathbf{F}_{\mathrm{R}\,k-1}$ is related to $\boldsymbol{\omega}$ as follows:

$$\mathbf{F}_{\mathrm{R}\,k-1} = \exp[\boldsymbol{\Omega}(\boldsymbol{\omega}_{k-1})\,\Delta_k] \tag{2}$$

where $\Delta_k = t_k - t_{k-1}$ is the sampling interval and $\boldsymbol{\Omega}(\boldsymbol{\omega})$ is the operator:

$$\boldsymbol{\Omega}(\boldsymbol{\omega}) = \begin{bmatrix} [\boldsymbol{\omega}\times] & \boldsymbol{\omega} \\ -\boldsymbol{\omega}^T & 0 \end{bmatrix} \tag{3}$$

and $[\boldsymbol{\omega}\times]$ is the skew-symmetric operator, [22]:

$$[\boldsymbol{\omega}\times] = \begin{bmatrix} 0 & -r & q \\ r & 0 & -p \\ -q & p & 0 \end{bmatrix} \tag{4}$$

The process noise vector $\boldsymbol{w}_{\mathrm{R}\,k-1}$ is related to the noise in the angular velocity measurements as follows:

$$\boldsymbol{w}_{\mathrm{R}\,k-1} = \frac{\Delta_k}{2}\cdot\begin{bmatrix} [\boldsymbol{q}_{k-1}^{nb}\times] + \mathbf{I}_3\cdot q_{4,k-1}^{nb} \\ -\boldsymbol{q}_{k-1}^{nb} \end{bmatrix} {}^{g}\boldsymbol{v}_{k-1} = \frac{\Delta_k}{2}\cdot\Xi(\overline{\boldsymbol{q}}_{k-1}^{nb})\,{}^{g}\boldsymbol{v}_{k-1} \tag{5}$$

where ${}^{g}\boldsymbol{v}_{k-1}$ is the gyroscope measurement noise, which is assumed white Gaussian with zero mean and covariance matrix $\boldsymbol{\Sigma}_g = \mathbf{I}_3\cdot\sigma_g^2$ ($\mathbf{I}_n$ is the $n \times n$ identity matrix). The process noise covariance matrix can be shown to have the following expression [10]:

$$\mathbf{Q}_{\mathrm{R}\,k-1} = \left(\frac{\Delta_k}{2}\right)^2\cdot\Xi(\overline{\boldsymbol{q}}_{k-1}^{nb})\,\boldsymbol{\Sigma}_g\,\Xi(\overline{\boldsymbol{q}}_{k-1}^{nb})^T \tag{6}$$

When tracked motions are relatively slow, as it is assumed in this paper, the sensed acceleration is simply taken as the projection of the gravity $\boldsymbol{g}^n$ along the sensitivity axes of the tri-axial accelerometer.

Since no heading information is available when the gravity vector is sensed, the measurement of the Earth's magnetic field $\boldsymbol{h}^n$ by the magnetic sensor may help producing drift-free heading estimates. The measurement equations are written as:

$$\begin{aligned} \boldsymbol{h}_k^b &= \mathbf{R}_k^{bn}\,\boldsymbol{h}^n + {}^{h}\boldsymbol{v}_k = \left\{\left(\overline{\boldsymbol{q}}_k^{nb}\right)^{-1}\otimes\overline{\boldsymbol{h}}^n\otimes\overline{\boldsymbol{q}}_k^{nb}\right\}_V + {}^{h}\boldsymbol{v}_k \\ \boldsymbol{a}_k^b &= \mathbf{R}_k^{bn}\,\boldsymbol{g}^n + {}^{a}\boldsymbol{v}_k = \left\{\left(\overline{\boldsymbol{q}}_k^{nb}\right)^{-1}\otimes\overline{\boldsymbol{g}}^n\otimes\overline{\boldsymbol{q}}_k^{nb}\right\}_V + {}^{a}\boldsymbol{v}_k \end{aligned} \tag{7}$$

where ${}^{h}\boldsymbol{v}_k$ and ${}^{a}\boldsymbol{v}_k$ are the measurement noises superimposed to the output of the accelerometer and the magnetic sensor, respectively; they are assumed white Gaussian with zero mean and covariance matrices $\boldsymbol{\Sigma}_h = \mathbf{I}_3\cdot\sigma_h^2$ and $\boldsymbol{\Sigma}_a = \mathbf{I}_3\cdot\sigma_a^2$, respectively. The operator $\otimes$ in Equation (7) is the quaternion product, $\overline{\boldsymbol{q}}^{-1}$ denotes the quaternion inverse, and $\overline{\boldsymbol{h}}^n$ and $\overline{\boldsymbol{g}}^n$ are quaternions with zero scalar part and vector part $\boldsymbol{h}^n$ and $\boldsymbol{g}^n$, respectively. The operator $\{\overline{\boldsymbol{q}}\}_V$ denotes the vector part of the quaternion $\overline{\boldsymbol{q}}$.

The EKF linearization requires the computation of the Jacobian matrices of the measurement Equation (7):

$$\mathbf{H}_k^{mag} = \mathbf{\Psi}\left(\bar{\boldsymbol{q}}_k^{nb}, \boldsymbol{h}^n\right) = \frac{\partial\left(\left(\bar{\boldsymbol{q}}_k^{nb}\right)^{-1} \otimes \bar{\boldsymbol{h}}^n \otimes \bar{\boldsymbol{q}}_k^{nb}\right)}{\partial \bar{\boldsymbol{q}}_k^{nb}}$$

$$\mathbf{H}_k^{acc} = \mathbf{\Psi}\left(\bar{\boldsymbol{q}}_k^{nb}, \boldsymbol{g}^n\right) = \frac{\partial\left(\left(\bar{\boldsymbol{q}}_k^{nb}\right)^{-1} \otimes \bar{\boldsymbol{g}}^n \otimes \bar{\boldsymbol{q}}_k^{nb}\right)}{\partial \bar{\boldsymbol{q}}_k^{nb}} \tag{8}$$

The operator $\mathbf{\Psi}(\bar{\boldsymbol{q}}, \boldsymbol{p})$ can be written for the quaternion $\bar{\boldsymbol{q}} = [q_1 \ q_2 \ q_3 \ q_4]^T$ and a quaternion $\bar{\boldsymbol{p}}$ with vector part $\boldsymbol{p}$ and zero scalar part as follows [22]:

$$\mathbf{\Psi}(\bar{\boldsymbol{q}}, \boldsymbol{p}) = \{\bar{\boldsymbol{q}}\}_\mathrm{R} \{\bar{\boldsymbol{p}}\}_\mathrm{R}^T + \{\bar{\boldsymbol{q}}\}_\mathrm{L}^T \{\bar{\boldsymbol{p}}\}_\mathrm{L} \begin{bmatrix} -\mathbf{I}_3 & 0 \\ 0 & 1 \end{bmatrix} \tag{9}$$

where:

$$\{\bar{\boldsymbol{q}}\}_\mathrm{L} = \begin{bmatrix} q_4 & q_3 & -q_2 & q_1 \\ -q_3 & q_4 & q_1 & q_2 \\ q_2 & -q_1 & q_4 & q_3 \\ -q_1 & -q_2 & -q_3 & q_4 \end{bmatrix} \{\bar{\boldsymbol{q}}\}_\mathrm{R} = \begin{bmatrix} q_4 & -q_3 & q_2 & q_1 \\ q_3 & q_4 & -q_1 & q_2 \\ -q_2 & q_1 & q_4 & q_3 \\ -q_1 & -q_2 & -q_3 & q_4 \end{bmatrix} \tag{10}$$

The measurement noise covariance matrix is given by:

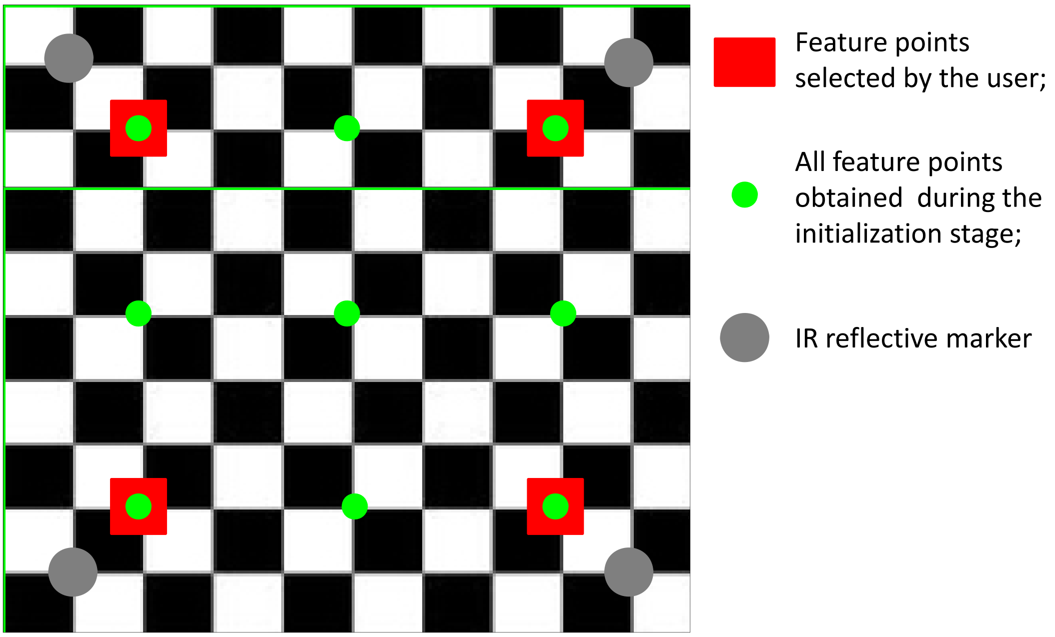
$$\mathbf{R} = \begin{bmatrix} \mathbf{\Sigma}_h & \mathbf{0}_3 \\ \mathbf{0}_3 & \mathbf{\Sigma}_a \end{bmatrix} \tag{11}$$

where $\mathbf{0}_n$ is the $n \times n$ null matrix. In order to guard against the effect of spurious magnetic measurements, which can be produced especially in indoor environments where magnetic fields are far from being homogeneous, the vector selection technique proposed in [24] is implemented: the strength of the sensed magnetic field h_norm and the dip angle θ_dip, namely the angle that is formed between the sensed magnetic field and the sensed gravity acceleration, are compared to their nominal values using suitably chosen threshold values, λ_dip and λ_h, respectively. Whenever either difference exceeds the corresponding threshold value, the magnetic measurement is considered invalid and therefore it is discarded from the filtering process by setting the matrix $\mathbf{H}_k^{mag}$ to zero. A similar vector selection technique is implemented by comparing the norm of the measured acceleration vector a_norm with the value of gravity ($1g = 9.81$ m/s^2) [19]: the acceleration measurement vector is assimilated by the EKF only when the absolute difference between a_norm and g is less than a threshold value λ_g, otherwise $\mathbf{H}_k^{acc}$ is set to zero.

2.2. Purely Vision-Based Method of Pose Estimation

We assume that the visual features are the projections into the image plane of N_f chessboard corners ($N_f = 9$) which represents our fiducial markers (Figure 2). Initially the user is asked to click on the four extreme corners of the chessboard in the first image frame, starting from the upper-left corner and proceeding counterclockwise; five additional 3D/2D correspondences are established by projecting the 3D chessboard model available in {**n**} back to the image plane in {**i**} based on the homography estimated using the four features selected by the user. The nine image point features we choose to identify in the first frame are then tracked using the pyramidal implementation of the Kanade-Lucas tracker (KLT) [25–27]. Henceforth, the squared area whose vertices are the four extreme corners of the chessboard is called the chessboard area.

Figure 2. The nine feature points constructed during the initialization stage are shown. Red squares: the four feature points manually selected by the user; green circles: the nine feature points constructed during the initialization stage.


Feature points selected by the user;
All feature points obtained during the initialization stage;
IR reflective marker

The image point features are fed to a least-squares estimation algorithm to calculate the transformation from $\{\mathbf{n}\}$ to $\{\mathbf{c}\}$ [28]. This algorithm is a variant of the DLT method [29], suited for tracking plane surfaces like the chessboard. The covariance matrix of the estimated pose is computed at each iteration step by analyzing the projection errors of the feature image points as suggested in [9].

2.3. EKF-Based Sensor Fusion Methods of Pose Estimation

The EKF-based sensor fusion method of body pose estimation requires that the rotational state vector $\boldsymbol{x}_{\mathrm{R}} = \bar{\boldsymbol{q}}^{nb}$ is extended using the components of the translational state vector $\boldsymbol{x}_{\mathrm{T}} = \begin{bmatrix} b_x^n & \dot{b}_x^n & b_y^n & \dot{b}_y^n & b_z^n & \dot{b}_z^n \end{bmatrix}^T$ which includes the position $\boldsymbol{b}^n = [b_x^n \quad b_y^n \quad b_z^n]^T$ and velocity $\dot{\boldsymbol{b}}^n = \begin{bmatrix} \dot{b}_x^n & \dot{b}_y^n & \dot{b}_z^n \end{bmatrix}^T$ of the IMU case in $\{\mathbf{n}\}$. The state-transition model equation is given by:

$$\boldsymbol{x}_{\mathrm{T}\,k} = \mathbf{F}_{\mathrm{T}\,k-1}\,\boldsymbol{x}_{\mathrm{T}\,k-1} + \boldsymbol{w}_{\mathrm{T}\,k-1} \tag{12}$$

In our approach accelerometers are used for stabilizing the IMU-camera attitude with respect to gravity (roll and pitch angles), as prescribed by the measurement Equation (7), under the assumption that the magnitude of the gravity vector is large enough to dominate the body acceleration, which is modeled as noise:

$$\ddot{\boldsymbol{b}}^{\boldsymbol{n}} = \boldsymbol{w} \tag{13}$$

where $\boldsymbol{w}$ is white Gaussian noise, with zero mean and covariance matrix ${}^{a}\boldsymbol{\Sigma} = \mathbf{I}_3 \cdot {}^{a}\sigma^2$, where the variance ${}^{a}\sigma^2$ is also called the strength of the driving noise [30].

The state transition matrix can be written as:

$$\mathbf{F}_{\mathrm{T}\,k} = \begin{bmatrix} \begin{bmatrix} 1 & \Delta_k \\ 0 & 1 \end{bmatrix} & \mathbf{0}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & \begin{bmatrix} 1 & \Delta_k \\ 0 & 1 \end{bmatrix} & \mathbf{0}_2 \\ \mathbf{0}_2 & \mathbf{0}_2 & \begin{bmatrix} 1 & \Delta_k \\ 0 & 1 \end{bmatrix} \end{bmatrix} \tag{14}$$

where Δ_k is the time interval elapsed between successive measurements, regardless of which sensors produce them.

The noise covariance matrix of the process noise $\boldsymbol{w}_{\mathrm{T}\,k-1}$ can be written as:

$$\mathbf{Q}_{\mathrm{T}\,k} = \begin{bmatrix} \begin{bmatrix} \Delta_k^4/4 & \Delta_k^3/2 \\ \Delta_k^3/2 & \Delta_k^2 \end{bmatrix} & \mathbf{0}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & \begin{bmatrix} \Delta_k^4/4 & \Delta_k^3/2 \\ \Delta_k^3/2 & \Delta_k^2 \end{bmatrix} & \mathbf{0}_2 \\ \mathbf{0}_2 & \mathbf{0}_2 & \begin{bmatrix} \Delta_k^4/4 & \Delta_k^3/2 \\ \Delta_k^3/2 & \Delta_k^2 \end{bmatrix} \end{bmatrix} {}^{a}\sigma^2 \tag{15}$$

The simplifying assumption that the translational and rotational components of the body motion are uncoupled is then made in writing the state transition model of the overall state vector $\boldsymbol{x}_k = [\boldsymbol{x}_{\mathrm{R}\,k}^T \quad \boldsymbol{x}_{\mathrm{T}\,k}^T]^T$ as follows:

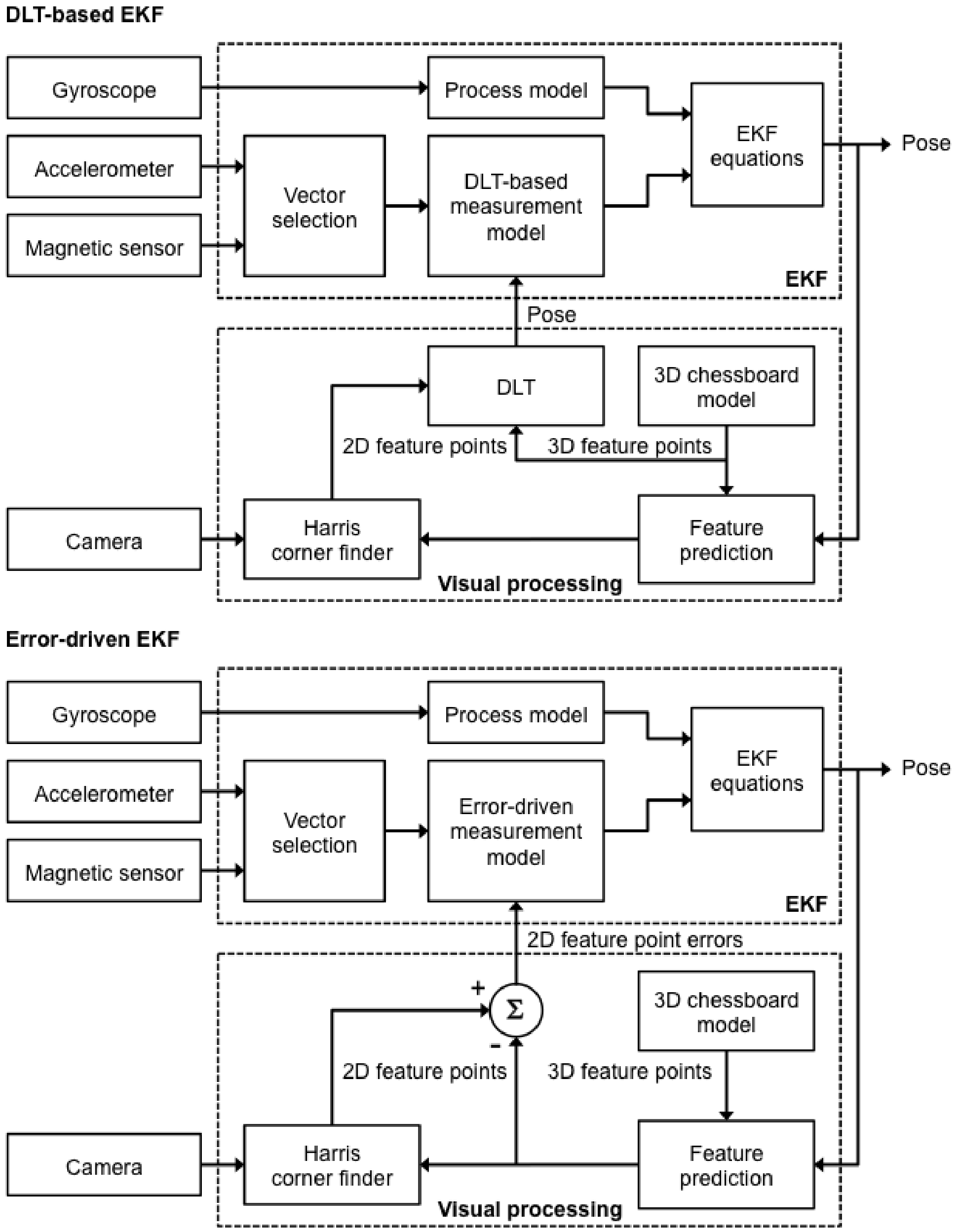
$$\boldsymbol{x}_k = \begin{bmatrix} \boldsymbol{F}_{\mathrm{R}\,k-1} & \mathbf{0}_{4\times 6} \\ \mathbf{0}_{6\times 4} & \boldsymbol{F}_{\mathrm{T}\,k-1} \end{bmatrix} \boldsymbol{x}_{k-1} \tag{16}$$

The covariance matrix of the process noise $\boldsymbol{w}_{k-1} = [\boldsymbol{w}_{\mathrm{R}\,k-1}^T \, \boldsymbol{w}_{\mathrm{T}\,k-1}^T]^T$ is:

$$\boldsymbol{Q}_{k-1} = \begin{bmatrix} \boldsymbol{Q}_{\mathrm{R}\,k-1} & \mathbf{0}_{4\times 6} \\ \mathbf{0}_{6\times 4} & \boldsymbol{Q}_{\mathrm{T}\,k-1} \end{bmatrix} \tag{17}$$

Two different sensor fusion strategies are considered to account for how to add the visual measurements $\boldsymbol{y}_k^{vis}$ to Equation (7), which leads to different dependencies between the output variables and the components of the system's state vector. Henceforth the two measurement models are called the *DLT-based model* and the *error-driven model*, hence the name DLT-based EKF and error-driven EKF for the corresponding sensor fusion methods, see Figure 3.

Figure 3. Block diagrams of the two Kalman-filter-based methods of sensor fusion.



A common element to both methods is the approach to visual feature tracking. While the purely vision-based method of pose estimation relies on the popular frame-to-frame KLT, visual feature tracking in either the DLT-based EKF or the error-driven EKF exploits the predicted *a priori* point features $\hat{\boldsymbol{p}}_{j,k}^{i}$ $j = 1, \ldots, N_f$ that are obtained from the projection of the 3D chessboard model in $\{\mathbf{i}\}$:

$$\hat{\boldsymbol{p}}_{j,k}^{i} = \mathbf{K}\,\mathbf{R}^{cb}\left(\hat{\mathbf{R}}_{k}^{bn}\left(\boldsymbol{Z}_{j}^{n} - \hat{\boldsymbol{b}}_{k}^{n}\right) - \boldsymbol{b}^{c}\right), j = 1, \ldots, N_f \tag{18}$$

where $\hat{\mathbf{R}}_{k}^{bn}$ and $\hat{\boldsymbol{b}}_{k}^{n}$ are derived from the *a priori* estimate of the state vector, and $\mathbf{K}$ is the camera calibration matrix [9]:

$$\mathbf{K} = \begin{bmatrix} f_x & \beta & -x_c \\ 0 & f_y & -y_c \\ 0 & 0 & 1 \end{bmatrix} \tag{19}$$

f_x and f_y are the two components of the focal length (theoretically, they should be equal), β takes accounts for any pixel misalignment within the optical sensor, while x_c and y_c are the coordinates of the principal point (image centre) relative to the origin of the frame $\{\mathbf{i}\}$. Equation (18) is based on the "pinhole model", according to which an ideal planar lens is assumed and optical distortion is neglected. Actually, the image point features are compensated for the distortion introduced by the lens system using the so-called Brown-Conrady model [31]. All camera intrinsic parameters, involved both in the camera calibration matrix and in the distortion model, were estimated during the camera calibration stage [32].

Features points $\hat{\boldsymbol{p}}^i_{j,k}$ are then used as initial conditions for the Harris corner finder. The Harris corner finder works by searching for the nearest black-and-white corner within a window that is centered around its predicted location [33]. The search window size, which is constrained between 5 and 20 pixels, is adaptively computed based on the predicted *a priori* error covariance.

For either method, the overall linearized measurement model can be written in the following form:

$$\mathbf{H}_k = \begin{bmatrix} \mathbf{H}_k^{mag} & \mathbf{0}_{3\times6} \\ \mathbf{H}_k^{acc} & \mathbf{0}_{3\times6} \\ \mathbf{H}_{\mathrm{R}\,k}^{vis} & \mathbf{H}_{\mathrm{T}\,k}^{vis} \end{bmatrix} \tag{20}$$

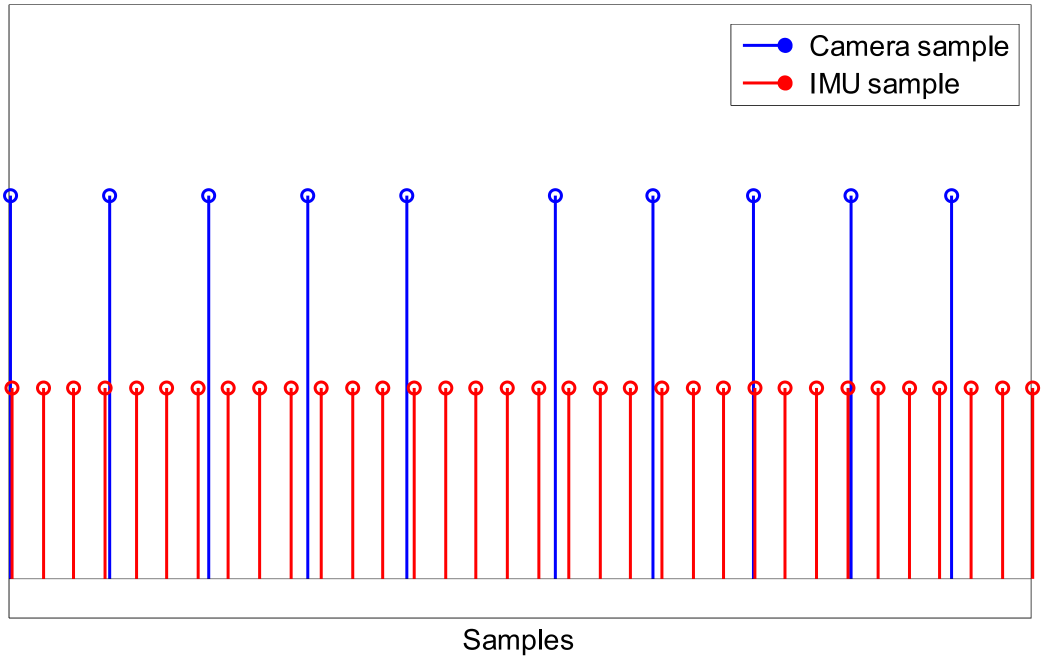
The measurement noise covariance matrix is written as follows:

$$\mathbf{R}_k = \begin{bmatrix} \mathbf{R}^{mag} & \mathbf{0}_3 & \mathbf{0}_{3\times N_{vis}} \\ \mathbf{0}_3 & \mathbf{R}^{acc} & \mathbf{0}_{3\times N_{vis}} \\ \mathbf{0}_{N_{vis}\times3} & \mathbf{0}_{N_{vis}\times3} & \mathbf{R}_k^{vis} \end{bmatrix} \tag{21}$$

The size of the matrices $\mathbf{H}_k$ and $\mathbf{R}_k$ depends on which EKF-based sensor fusion method we consider. Implicit in the formulation of Equation (20) is that inertial/magnetic sensing contributes only to the estimate of orientation, while visual sensing conveys information about all the six DOFs.

A multi-rate filtering strategy is needed to deal with the different sampling rates of IMU and camera measurements: the IMU measurement process runs at a rate of 100 Hz, while the camera measurement process is slower, running at a rate of approximately 30 fps (Figure 4). Both EKFs can be defined as multi-rate, which alludes to the transition between different measurement equations that must be performed within the filter depending on which measurements are available. Since the time instant when the inertial/magnetic and visual measurements are made is known to the system, the time lag between successive measurements Δ_k is also known, which allows propagating the state vector in the prediction stage and selecting which rows of the Jacobian matrix in Equation (20) would be actually set to zero in the update stage at any iteration step of the filter. In other words, in the time intervals between successive image frames from the camera only IMU measurements are to be processed, which implies that the measurement equations of both EKFs are identical to the measurement equations of the purely IMU-based method of orientation determination described in Section 2.1. Then, when a new image frame becomes available, the measurement equations are suitably changed in order to assimilate the visual information, leading to the measurement equations presented in Sections 2.3.1 and 2.3.2 for the two EKFs (see below).

Figure 4. Timestamps of camera (blue) and IMU (red) samples. The number of IMU samples between successive image frames is slightly variable due to the irregular sampling rate of the camera.



2.3.1. DLT-Based Measurement Model

The DLT method reviewed in Section 2.2 provides, for each incoming image frame, the estimate of the chessboard pose in $\{\mathbf{c}\}$, in terms of $\mathbf{R}_k^{cn}$ or $\bar{\boldsymbol{q}}_k^{cn}$ and $\boldsymbol{n}_k^c$. This is based on using the correspondences between the image point features and their corresponding corner points on the chessboard. The DLT output can be expressed directly in terms of the body pose in $\{\mathbf{n}\}$ using the following transformations:

$$\begin{aligned}\bar{\boldsymbol{q}}_k^{nb} &= (\bar{\boldsymbol{q}}_k^{cn})^{-1} \otimes \bar{\boldsymbol{q}}^{cb} \\ \boldsymbol{b}_k^n &= (\mathbf{R}_k^{cn})^T(\boldsymbol{b}^c - \boldsymbol{n}_k^c)\end{aligned} \tag{22}$$

We recall that $\bar{\boldsymbol{q}}^{cb}$*and* $\boldsymbol{b}^c$ are known from solving the IMU-camera relative pose calibration problem, as already described above.

The visual observation matrix can be simply written as:

$$\mathbf{H}^{vis} = \left[\mathbf{H}_R^{vis}\ \mathbf{H}_T^{vis}\right] = \begin{bmatrix} \boldsymbol{I}_4 & \mathbf{0}_{4\times6} \\ \mathbf{0}_{6\times4} & \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{bmatrix} \tag{23}$$

The measurement noise covariance matrix is:

$$\mathbf{R}_k^{vis} = \begin{bmatrix} \mathrm{cov}\left(\bar{\boldsymbol{q}}_k^{nb}\right) & \mathbf{0}_{4\times3} \\ \mathbf{0}_{3\times4} & \mathrm{cov}(\boldsymbol{b}_k^n) \end{bmatrix} \tag{24}$$

where:

$$\begin{aligned}\operatorname{cov}\left(\overline{\boldsymbol{q}}_k^{nb}\right) &= \frac{1}{4}\,\Xi\left(\overline{\boldsymbol{q}}_k^{nb}\right)\operatorname{cov}(\psi,\vartheta,\phi)\,\Xi^{\mathrm{T}}\left(\overline{\boldsymbol{q}}_k^{nb}\right)\\ \operatorname{cov}(\boldsymbol{b}_k^n) &= \left(\mathbf{R}_k^{cn}\right)^T\operatorname{cov}(\boldsymbol{n}_k^c)\,\mathbf{R}_k^{cn}\end{aligned} \tag{25}$$

In principle, the covariance matrix $\operatorname{cov}(\psi,\vartheta,\phi)$ of the Euler angles ψ,ϑ,ϕ and the covariance matrix $\operatorname{cov}(\boldsymbol{n}_k^c)$ of the translation vector $\boldsymbol{n}_k^c$ are provided by the DLT method using the methods described in [9]. However, a stable behavior of the DLT-based EKF is simply obtained by taking $\operatorname{cov}(\psi,\vartheta,\phi) = \mathbf{I}_3\cdot\sigma_\theta^2$ ($\sigma_\theta = 0.05°$) and $\operatorname{cov}(\boldsymbol{b}_k^n) = \mathbf{I}_3\cdot\sigma_b^2$ ($\sigma_b = 1$ mm). These values are in close agreement with the experimental uncertainty estimated during extensive experimental testing of the DLT method in our experimental setup (not reported in this paper).

2.3.2. Error-Driven Measurement Model

The feature projection errors at time t_k are the difference between the measured image point features with coordinates $\boldsymbol{z}_{j,k}^i, j = 1,\dots,N_f$ and the a priori predicted features points $\widehat{\boldsymbol{p}}_{j,k}^i$ (see Section 2.3).

The measurement equation can be written as:

$$\boldsymbol{y}_{j,k}^{vis} = \mathbf{K}\,\mathbf{R}^{cb}\left(\widehat{\mathbf{R}}_k^{bn}\left(\boldsymbol{Z}_j^n - \widehat{\boldsymbol{b}}_k^n\right) - \boldsymbol{b}^c\right) - \boldsymbol{z}_{j,k}^i \tag{26}$$

Since the dependence of the measurements $\boldsymbol{y}_{j,k}^{vis}$ from the quaternion $\overline{\boldsymbol{q}}_k^{nb}$ is nonlinear, the Jacobian matrix of the transformation from Equation (26) must be computed as part of the EKF linearization:

$$\frac{\partial\boldsymbol{y}_{j,k}^{vis}}{\partial\overline{\boldsymbol{q}}_k^{nb}} = \mathbf{K}\,\mathbf{R}^{cb}\frac{\partial\left(\left(\overline{\boldsymbol{q}}_k^{nb}\right)^{-1}\otimes\overline{\boldsymbol{Z}_j^n - \boldsymbol{b}_k^n}\otimes\overline{\boldsymbol{q}}_k^{nb}\right)}{\partial\overline{\boldsymbol{q}}_k^{nb}} = \mathbf{K}\,\mathbf{R}^{cb}\boldsymbol{\Psi}\left(\overline{\boldsymbol{q}}_k^{nb},\boldsymbol{Z}_j^n - \boldsymbol{b}_k^n\right) = \mathbf{H}_k^{vis} \tag{27}$$

The Jacobian matrix related to the translational part of the state vector can be written:

$$\boldsymbol{H}_{\mathrm{T}k}^{vis} = \left[\left(\frac{\partial\boldsymbol{y}_{j,k}^{vis}}{\partial\boldsymbol{b}_k^n}\right)^{(1)}\quad\mathbf{0}_{3\times1}\quad\left(\frac{\partial\boldsymbol{y}_{j,k}^{vis}}{\partial\boldsymbol{b}_k^n}\right)^{(2)}\quad\mathbf{0}_{3\times1}\quad\left(\frac{\partial\boldsymbol{y}_{j,k}^{vis}}{\partial\boldsymbol{b}_k^n}\right)^{(3)}\quad\mathbf{0}_{3\times1}\right] \tag{28}$$

where $\left(\frac{\partial\boldsymbol{y}_{j,k}^{vis}}{\partial\boldsymbol{b}_k^n}\right)^{(m)}$ denotes the *m*-column of $\frac{\partial\boldsymbol{y}_{j,k}^{vis}}{\partial\boldsymbol{b}_k^n}$:

$$\frac{\partial\boldsymbol{y}_{j,k}^{vis}}{\partial\boldsymbol{b}_k^n} = -\mathbf{K}\,\mathbf{R}^{cb}\mathbf{R}_k^{bn} \tag{29}$$

The visual measurement noise covariance matrix $\boldsymbol{R}^{vis}$ can be written as:

$$\mathbf{R}^{vis} = \mathbf{I}_{2N_f\times2N_f}\cdot{}^{vis}\sigma^2 \tag{30}$$

where the standard deviation ${}^{vis}\sigma$ measures the uncertainty of the Harris corner finder [33]. We chose the value ${}^{vis}\sigma = 0.75$ pixel, rather than the more optimistic value suggested in [33] $\left({}^{vis}\sigma = 0.1 \text{ pixel}\right)$, which gave rise to a more stable filter in our experiments.

2.4. Experimental Validation

Eight tracking experiments, each lasting 60 s, were conducted by moving the sensor unit assembly of Figure 1 freely by hand in all six DOFs, with the constraint to keep the chessboard area always within the camera FOV. The angular velocities were up to 40°/s and the linear accelerations were up to 0.6 m/s^2. An additional tracking experiment was performed by moving the sensor unit assembly 1° at a time.

The IMU sensors were calibrated using the in-field calibration techniques described in [34]. In particular, the gyroscope was compensated for the initial bias value by taking the average of its output during a rest period of 1 s, just before the IMU motion started (bias-capture procedure).

The following filtering methods were tested: the purely IMU-based method of orientation estimation (Section 2.1); the purely vision-based method of pose estimation (Section 2.2); and the two methods of sensor fusion named DLT-based EKF (Section 2.3.1) and error-driven EKF (Section 2.3.2). In all cases no gating technique was implemented in the EKFs to detect outliers due to mismatched features in consecutive image frames. The sensor data acquired during the tracking experiments were analyzed for the off-line validation study in five different filtering scenarios: (a) inertial/magnetic sensor measurements from the IMU were ignored by the filters; (b) inertial/magnetic sensor measurements from the IMU were assimilated in the filters; (c) the magnetic sensor measurements from the IMU were ignored by the filters; (d) gyro bias was not compensated by bias capture, in the situation when magnetic sensor measurements from the IMU were ignored by the filters; (e) a mechanism of intentional damage to the integrity of visual information was implemented and inertial/magnetic sensor measurements were assimilated by the filters. The rationale behind (c) was to stress the importance of magnetic sensor measurements for heading stabilization. The rationale behind (d) was to urge the capability of the proposed sensor fusion methods to accommodate slight imperfections that are typical of inertial sensors. Finally, the rationale behind (e) was to assess the tracking robustness of the sensor fusion methods against visual gaps. The mechanism for degrading the visual information made available to the DLT-based EKF and the error-driven EKF was implemented as follows: for each incoming image frame, a random sample of visual features with size randomly selected from 0 (*i.e.*, no deletions occurred) to the maximum number tolerated by each filter (*i.e.*, nine for the error-driven EKF and three for the DLT-based EKF) was discarded by setting the corresponding rows of the Jacobian matrix $\mathbf{H}^{vis}$ to zero (this trick allowed preventing the information associated with the selected features to influence the filtering process); at the next image frame, number and identity of the removed visual features were due to change independently based on the chosen random selection process. The filter parameter setting reported in Table 1 was chosen.

Table 1. Parameter tuning.

Process noise	
σ_g [°]	0.40
${}^{a}\sigma$ [m/s^2]	0.05
Measurement noise	
σ_h [mGauss]	2
σ_a [mg]	10
σ_θ [°]	0.05
σ_b [mm]	1
${}^{vis}\sigma$ [pixel]	0.75
Vector selection	
λ_h [mGauss]	20
λ_{dip} [°]	5
λ_g [mg]	20

The reference data were interpolated using cubic splines to the time instants when inertial/magnetic and visual measurements were made. Standard conversion formulae were then used to convert the reference and estimated quaternions in the corresponding Euler angles. The performance assessment was based on the root mean square errors (RMSEs) of the estimated roll, pitch and yaw angles. Moreover, the error quaternion $\Delta\bar{\boldsymbol{q}} = \bar{\boldsymbol{q}}_{\text{ref}}^{-1} \otimes \hat{\bar{\boldsymbol{q}}}$ represented the estimated rotation needed to bring the estimated body frame into $\{\mathbf{b}\}$: the scalar component of $\Delta\bar{\boldsymbol{q}}$, namely $\Delta\theta = 2\cos^{-1}(\Delta q_4)$ was used to compute the orientation RMSE. The RMSE of the estimated position was computed separately for each coordinate axis (e_X, e_Y and e_Z) and as a total position error $e_T = \sqrt{e_X^2 + e_Y^2 + e_Z^2}$. Finally, the RMSE values calculated in the eight tracking experiments were summarized using mean value and standard deviation (SD).

The filtering algorithms were implemented using Matlab; the experimental validation was carried out in off-line conditions. Since the 10-dimensional state vector was the same for either the DLT-based EKF or the error-driven EKF, the operations involved in the prediction stage were exactly the same, which took (approximately) 1 ms in the current implementation (standard laptop, 2.2 GHz clock frequency). Another common element was the vector matching process for the sensed acceleration and magnetic field vectors, which required 1 ms, while the computation of the inertial/magnetic Jacobian matrix took approximately 1 ms. The difference between the two EKFs was in the visual measurement equations: in the DLT-based EKF 10 measurement channels were deployed, in contrast with the 24 measurement channels needed by the error-driven EKF. The computation of the visual features required 14 ms in both filters, which included state propagation and prediction. In the DLT-based EKF, the DLT method was implemented at each iteration cycle, followed by the update of the time-varying measurement noise covariance matrix in Equation (24); conversely, in the error-driven EKF the computation of the visual Jacobian matrix—see Equations (27–29)—was needed at each iteration cycle. In conclusion, both filters would require 16 ms for each iteration cycle when an image frame was available for processing. The purely vision-based method was more computationally expensive (approximately, 28 ms), mainly because of the need for the pyramidal implementation of the KLT tracker. The purely IMU-based method took about 2 ms for iteration cycle.

3. Experimental Results

The RMSE values of the eight tracking experiments are summarized in mean value ± SD in Table 2, when all tested filtering methods are based on visual measurements only, and in Tables 3–5, where visual measurements are fused with inertial/magnetic measurements: in particular, Tables 4 and 5 report the summary statistics of the performance metrics when magnetic measurements are prevented from influencing the filtering process—the conditions under which data in Table 4 are produced differ from those valid for Table 5 depending whether the gyro bias capture is enabled (Table 4) or not (Table 5). The label TF, *i.e.*, Tracking Failure indicates the inability of the error-driven EKF to successfully complete the tracking task when the inertial/magnetic measurements are not integrated within the filter. The label N/A, *i.e.*, Not Available indicates the inability of the purely-IMU based method of orientation estimation to do positioning.

Table 2. Summary statistics of the performance metrics in the scenario (a).

	Purely-vision based	**DLT-based EKF**	**Error-driven EKF**
Yaw, °	0.45 ± 0.08	0.40 ± 0.04	TF
Pitch, °	0.64 ± 0.10	0.66 ± 0.13	TF
Roll, °	0.78 ± 0.14	0.73 ± 0.09	TF
Orientation, °	1.11 ± 0.16	1.09 ± 0.19	TF
X, mm	1.55 ± 0.42	1.54 ± 0.31	TF
Y, mm	2.67 ± 0.59	2.88 ± 0.43	TF
Z, mm	4.45 ± 0.71	2.14 ± 0.32	TF
Position, mm	5.45 ± 0.76	3.95 ± 0.52	TF

Table 3. Summary statistics of the performance metrics in the scenario (b).

	Purely-IMU based	**DLT-based EKF**	**Error-driven EKF**
Yaw, °	1.04 ± 0.27	0.41 ± 0.06	0.81 ± 0.16
Pitch, °	0.76 ± 0.19	0.63 ± 0.10	0.78 ± 0.31
Roll, °	0.96 ± 0.15	0.78 ± 0.10	0.92 ± 0.13
Orientation, °	1.61 ± 0.29	1.08 ± 0.13	1.46 ± 0.25
X, mm	N/A	1.37 ± 0.41	2.59 ± 0.56
Y, mm	N/A	2.82 ± 0.48	6.72 ± 1.20
Z, mm	N/A	1.96 ± 0.24	6.64 ± 2.41
Position, mm	N/A	3.40 ± 1.10	10.00 ± 1.75

Table 4. Summary statistics of the performance metrics in the scenario (c).

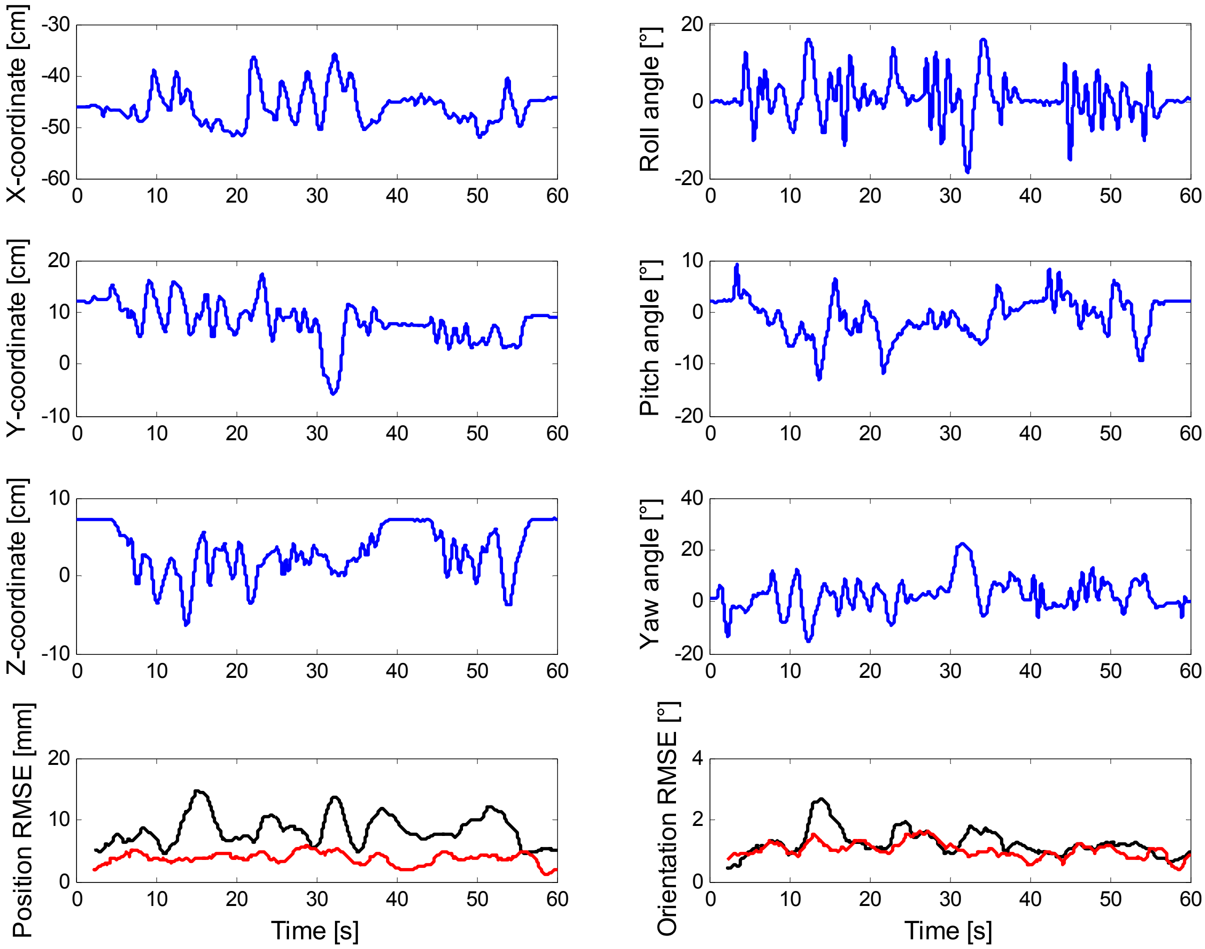
	Purely-IMU based	**DLT-based EKF**	**Error-driven EKF**
Yaw, °	2.16 ± 2.03	0.43 ± 0.06	2.34 ± 1.81
Pitch, °	0.80 ± 0.16	0.65 ± 0.09	0.84 ± 0.20
Roll, °	1.29 ± 1.40	0.81 ± 0.11	0.81 ± 0.15
Orientation, °	3.00 ± 1.95	1.10 ± 0.15	2.72 ± 1.63
X, mm	N/A	1.48 ± 0.44	4.47 ± 2.01
Y, mm	N/A	3.03 ± 0.32	20.67 ± 13.4
Z, mm	N/A	2.01 ± 0.29	5.89 ± 2.13
Position, mm	N/A	3.90 ± 0.51	22.22 ± 13.0

Table 5. Summary statistics of the performance metrics in the scenario (d).

	Purely-IMU based	**DLT-based EKF**	**Error-driven EKF**
Yaw, °	29.93 ± 0.90	0.41 ± 0.06	3.07 ± 0.62
Pitch, °	1.01 ± 0.27	0.69 ± 0.09	1.20 ± 0.51
Roll, °	1.16 ± 0.17	0.77 ± 0.10	1.00 ± 0.16
Orientation, °	29.99 ± 0.91	1.08 ± 0.13	3.41 ± 0.65
X, mm	N/A	1.36 ± 0.41	4.92 ± 1.58
Y, mm	N/A	2.82 ± 0.48	24.27 ± 6.33
Z, mm	N/A	1.97 ± 0.24	9.19 ± 4.80
Position, mm	N/A	3.71 ± 0.60	26.79 ± 6.27

The representative plots in Figure 5 are produced by running the DLT-based EKF and the error-driven EKF using sensor data from one of the eight tracking experiments in the scenario (b).

Figure 5. For one of the eight tracking experiments, plots of reference position and Euler angles of the body pose, together with plots of the position and orientation RMSE. For the sake of visualization, the RMSE values are computed over moving average windows of duration 5 s (DLT-based EKF, in red; error-driven EKF, in black).



The plot of Figure 6 concerns the results of tracking one rotational DOF at a time, when the error-driven EKF runs in the scenario (a). Finally, the results of eroding the amount of visual information made available to the filtering methods are presented in Figure 7.

Figure 6. Reference and filtered position and Euler angles that are produced using the error-driven EKF (pure vision). Reference data in blue; filtered data in red.

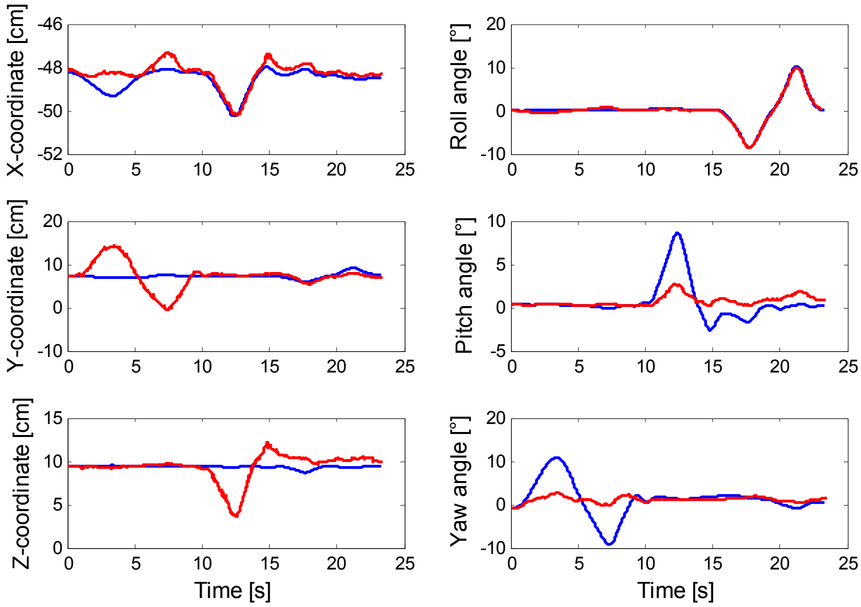
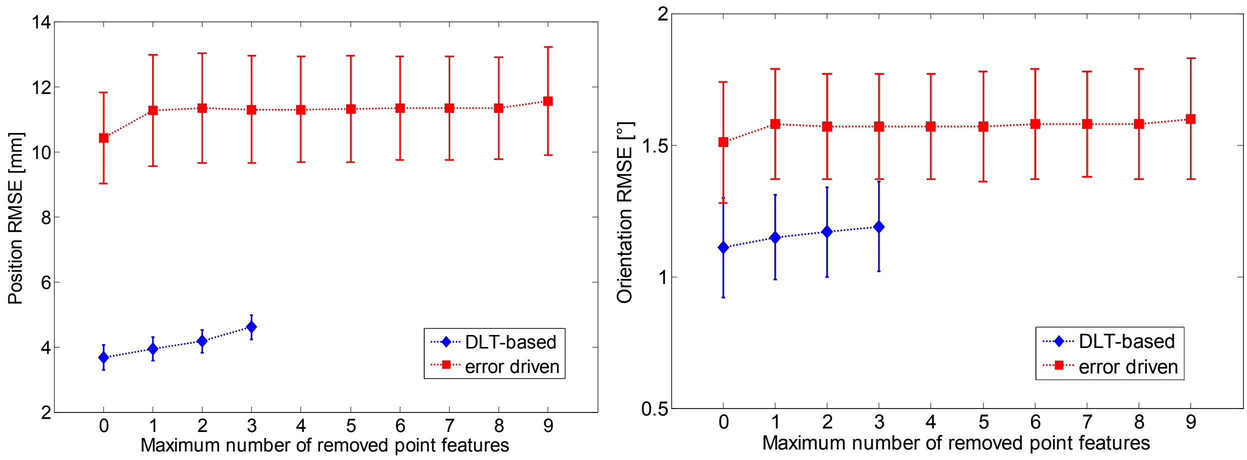


Figure 7. Error plots of the position and orientation RMSEs *versus* the number of image feature points that are removed at each iteration during processing by either the DLT-based EKF or the error-driven EKF (see text).



4. Discussion

In this paper positioning is not attempted using inertial/magnetic sensors alone, as it is done, e.g., in pedestrian navigation systems, when the IMU is attached to the foot. The exploitation of biomechanical constraints that concern the dynamics of human walking allows indeed mitigating the error growth incurred in the double-time integration process of gravity-compensated acceleration components: for instance, the cubic-time growth of positioning errors can be broken down to a linear-time growth by implementing zero-velocity updates (ZUPT) at the times when the foot is detected steady during walking [35]. This approach cannot be pursued in general, and in particular when the tracked motions are too slow and rest periods for ZUPT are infrequent, if any, which is the case in the tracking experiments discussed in this paper. In other words, positioning is possible in our experimental setup only because of the availability of monocular vision, provided that we can properly deal with the scale ambiguity in the translational ego-motion. The DLT-based EKF using vision alone and the purely vision-based method are characterized by the same accuracy of pose estimation in the experimental trials of this paper, as shown in Table 2; it is worth noting that, when inertial/magnetic measurements are incorporated in the filter, the predictive mechanism implemented in the DLT-based EKF allows it to perform the feature tracking task with the same efficiency as the KLT algorithm and much lower computational costs.

However, the informative contribution of the inertial/magnetic or just the inertial measurements to the DLT-based EKF is not relevant to boost the accuracy of pose estimation—for slow tracked motions, the DLT-based visual measurements are sufficient to obtain very accurate pose estimates—see Tables 2–5.

In contrast to the DLT-based EKF, the error-driven EKF benefits greatly from the integration of inertial/magnetic or from inertial measurements (to a lesser extent), without which it fails in the experimental trials of this paper. The error-driven EKF performs better, or even much better, than the purely IMU-based method in terms of attitude estimation accuracy, while yielding quite accurate estimates of position too. However, some problems of the error-driven EKF are raised, especially when magnetic measurements are not incorporated in the filtering process, which are not shown by the DLT-based EKF. Our explanation is that providing the sensor fusion method with direct measurements of the quaternion and translation vector of interest is much more informative than relying on visual projection errors as the error-driven EKF does.

The value of incorporating the magnetic sensor measurements in the sensor fusion process is assessed by analyzing the data reported in Tables 3–5. Since the visual measurements are highly informative on all six DOFs, the DLT-based EKF performs accurately even in the experimental scenarios (c) and (d) (Tables 4 and 5). Conversely, the error-driven EKF suffers substantially from lacking the magnetic sensor information, although the visual measurements allow somewhat mitigating the error growth in the orientation estimates. Nonetheless, the positioning accuracy is due to degrade significantly, especially in the experimental scenario (d), which is reflected in the quite high SDs attached to the RMSE average values in Tables 4 and 5.

The reason is that the error-driven EKF may suffer from gross mismatches between estimated and reference poses. In practice, wrong state vector estimates are produced, which do not preclude however the system from successfully tracking the image point features. This is a good instance of the problem of ambiguous/multiple solutions to the pose estimation problem. As discussed in [36,37], the motion of a planar target seen from perspective views can result ambiguous even if four or more coplanar points are used to generate the 2D/3D correspondences. A typical ambiguity problem is represented by the rotation/translation coupling [37] in which yaw or pitch angle variations are interpreted as translational displacements along the *Y*- or *Z*-axis, respectively, as shown in Figure 6—see Figure 1 for interpreting the meaning of the axes: changes of the yaw angle are wrongly interpreted as motion occurring along the *Y*-axis, in the same way as changes of the pitch angle are misleadingly interpreted as motion occurring along the *Z*-axis. Moreover the state parameter values that minimize the projection errors may be quite different from the physical orientation and translation from $\{\mathbf{b}\}$ to $\{\mathbf{n}\}$. This problem is due to the non-linear nature of the least-square method used by the error-driven EKF to generate the pose from the projection errors, which is prone to local minima. Visuo-inertial integration is a suitable means to deal with the problem of ambiguous/multiple solutions: the error-driven EKF is indeed capable of correctly disambiguating critical motions thanks to the IMU measurements, especially when measurements from the magnetic sensor are integrated and gyro bias is properly compensated by the bias capture procedure, as shown in Table 3.

The visual sabotage implemented in this paper is not as extreme as permanent losses of image point features would be, such as those occurring in case of occlusions, or when the ego-motion is so fast that part or all of the chessboard area escapes the camera FOV. We simply limit to randomly reduce number and location of coplanar feature points, sometimes even below the minimum number theoretically needed for pose estimation. The data reported in Figure 7 demonstrates the superiority, in terms of visual robustness terms, of the error-driven EKF over the DLT-based EKF. In fact, the former filter can tolerate reductions down to zero of the image point features without experiencing tracking losses of any kind while the latter absolutely needs a minimum number of six image point features. In addition, the RMSE values of the DLT-based increase progressively with the number of removed features, in contrast to the RMSE values of the error-driven EKF.

The main problem experienced in regard of loss of vision is as follows: since it is only the vision that does positioning, the position estimates tend to diverge fast when the system is blind, visually speaking. While the orientation estimates are continuously and accurately provided by the inertial/magnetic sensors, it is this diverging trend that explains why projection errors may rapidly grow to an extent that makes impossible for the system to maintain the track on the chosen fiducial markers. To make matters worse, we have decided not to implement any mechanism for monitoring the filter divergence based on the number of visual features registered, or any re-initialization procedure in case of divergence [38]: a Kalman-based filter would be capable, in principle, of recovering tracking losses of short duration using either the information on the motion trajectory captured by the dynamic model or the information from the inertial/magnetic sensors.

5. Conclusions

In this paper two approaches to fuse visual and inertial/magnetic measurements have been considered and correspondingly two EKFs have been developed to track the ego-motion in all six DOFs. They were analyzed with the aim to elucidate how the visual and inertial/magnetic measurements cooperate together and to which extent they do for ego-motion estimation. The two filters perform differently in terms of accuracy and robustness: in the DLT-based EKF the visual measurements seem to have a higher informational content as compared to the inertial/magnetic measurements, and the overall system shows remarkably good accuracy in estimating all six DOFs; conversely, in the error-driven EKF the inertial/magnetic measurements are fundamental for the correct operation of the filter, and the overall system can thus gain in robustness against loss of visual information, at the expense of accuracy in estimating all six DOFs. Moreover, the strategy of sensor fusion is interesting in other respects: on the one hand, the DLT-based EKF takes advantage of the inertial/magnetic measurements since visual features can be tracked without using tools like the KLT, which are computationally time-consuming; on the other hand, the error-driven EKF does positioning only because of its capability of exploiting the projection errors of the image point features.

That magnetic sensor measurements can be helpful to stabilize heading is highlighted in our results, although this statement cannot be overemphasized given the difficulties of motion tracking in magnetically perturbed environments [39]. Another limitation of the present work is that we have not considered the effects of fast motions on the filter behavior. Actually, we have implemented vector selection schemes for accelerometer and magnetic sensor measurements, as done, e.g., in [24]; however, due to the benign nature of the tracked motions and the magnetic environment surrounding the IMU, they were substantially inactive during all tracking experiments described in this paper. A possibility to deal with magnetically perturbed environments would be to augment the state vector with the magnetic disturbance as done, e.g., in [39]; a possibility to deal with aggressive movements would be to modify the state vector by including angular velocity and linear acceleration into it [2,18,40]. Both possibilities are technically feasible in our approach, and they are left for our ongoing work. We plan to improve this work in several other respects: in particular we intend to remove the limitations of working with fixed calibration patterns like the chessboard by exploiting natural features that are usually present in unprepared environments, paving the way to the implementation of an SFM system. Although this effort may greatly complicate the feature extraction/tracking steps, faster and more natural ego-motions would be considered in our experimental scenarios.

In conclusion, in this paper we proposed two different models of visual measurements to be used within Kalman-based filters that also incorporate inertial/magnetic measurements for estimating the ego-motion of a hand-held IMU/camera sensor unit. The two proposed EKFs were off-line analyzed in different experimental conditions: the DLT-based EKF was more accurate than the error-driven EKF, less robust against loss of visual features, and equivalent in terms of computational complexity. Orientation RMSEs of 1° (1.5°) and position RMSEs of 3.5 mm (10 mm) were achieved in our experiments by the DLT-based EKF (error-driven EKF). By contrast, the purely IMU-based EKF achieved orientation RMSEs of 1.6°.

Acknowledgements

This work has been supported in part by funds from the Italian Ministry of Education and Research.

References

1. Welch, G.; Foxlin, E. Motion tracking: No silver bullet, but a respectable arsenal. *IEEE Comput. Graph. Appl.* **2002**, *22*, 24–38.
2. Hol, J.; Schön, T.; Luinge, H.; Slycke, P.; Gustafsson, F. Robust real-time tracking by fusing measurements from inertial and vision sensors. *J. Real-Time Imag. Proc.* **2007**, *2*, 149–160.
3. Bleser, G.; Hendeby, G.; Miezal, M. Using Egocentric Vision to Achieve Robust Inertial Body Tracking under Magnetic Disturbances. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*, Basel, Switzerland, 26–29 October 2011; pp. 103–109.
4. Vieville, T.; Romann, F.; Hotz, B.; Mathieu, H.; Buffa, M.; Robert, L.; Facao, P.E.D.S.; Faugeras, O.D.; Audren, J.T. Autonomous Navigation of a Mobile Robot Using Inertial and Visual Cues. In *Proceedings ot the IEEE International Conference on Intelligent Robots and Systems*, 26–30 July 1993; pp. 360–367.

5. Kelly, J.; Sukhatme, G.S. Visual-inertial sensor fusion: Localization, mapping and sensor-to-sensor self-calibration. *Int. J. Robot. Res.* **2011**, *30*, 56–79.
6. Nützi, G.; Weiss, S.; Scaramuzza, D.; Siegwart, R. Fusion of IMU and vision for absolute scale estimation in monocular Slam. *J. Intell. Robot Syst.* **2011**, *61*, 287–299.
7. Achtelik, M.; Weiss, S.; Siegwart, R. Onboard IMU and Monocular Vision Based Control for MAVs in Unknown in- and Outdoor Environments. In *Proceedings of the IEEE International Conference of Robotics and Automation*, Shangai, China, 9–13 May 2011; pp. 3056–3063.
8. Tao, Y.; Hu, H.; Zhou, H. Integration of vision and inertial sensors for 3D arm motion tracking in home-based rehabilitation. *Int. J. Rob. Res.* **2007**, *26*, 607–624.
9. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*; Cambridge University Press: Cambridge, UK, 2003.
10. Sabatini, A.M. Estimating three-dimensional orientation of human body parts by inertial/magnetic sensing. *Sensors* **2011**, *11*, 1489–1525.
11. Foxlin, E. Pedestrian tracking with shoe-mounted inertial sensors. *IEEE Comput. Graph. Appl.* **2005**, *25*, 38–46.
12. Bebek, O.; Suster, M.A.; Rajgopal, S.; Fu, M.J.; Xuemei, H.; Cavusoglu, M.C.; Young, D.J.; Mehregany, M.; van den Bogert, A.J.; Mastrangelo, C.H. Personal navigation via high-resolution gait-corrected inertial measurement units. *IEEE Trans. Instrum. Meas.* **2010**, *59*, 3018–3027.
13. Corke, P.; Lobo, J.; Dias, J. An introduction to inertial and visual sensing. *Int. J. Rob. Res.* **2007**, *26*, 519–535.
14. Aron, M.; Simon, G.; Berger, M.-O. Use of inertial sensors to support video tracking: Research articles. *Comput. Animat. Virt. World.* **2007**, *18*, 57–68.
15. Klein, G.S.W.; Drummond, T.W. Tightly integrated sensor fusion for robust visual tracking. *Imag. Vis. Comput.* **2004**, *22*, 769–776.
16. Hwangbo, M.; Kim, J.-S.; Kanade, T. Gyro-aided feature tracking for a moving camera: Fusion, auto-calibration and gpu implementation. *Int. J. Robot. Res.* **2011**, *30*, 1755–1774.
17. Qian, G.; Chellappa, R.; Zheng, Q. Robust structure from motion estimation using inertial data. *J. Opt. Soc. Am. A* **2001**, *18*, 2982–2997.
18. Gemeiner, P.; Einramhof, P.; Vincze, M. Simultaneous motion and structure estimation by fusion of inertial and vision data. *Int. J. Robot. Res.* **2007**, *26*, 591–605.
19. Sabatini, A.M. Quaternion-based extended kalman filter for determining orientation by inertial and magnetic sensing. *IEEE Trans. Biomed. Eng.* **2006**, *53*, 1346–1356.
20. Yun, X.; Bachmann, E.R. Design, implementation, and experimental results of a quaternion-based kalman filter for human body motion tracking. *IEEE Trans. Robot.* **2006**, *22*, 1216–1227.
21. Roetenberg, D.; Slycke, P.J.; Veltink, P.H. Ambulatory position and orientation tracking fusing magnetic and inertial sensing. *IEEE Trans. Biomed. Eng.* **2007**, *54*, 883–890.
22. Shuster, M.D. Survey of attitude representations. *J. Astronaut. Sci.* **1993**, *41*, 439–517.
23. Lobo, J.; Dias, J. Relative pose calibration between visual and inertial sensors. *Int. J. Robot. Res.* **2007**, *26*, 561–575.
24. Harada, T.; Mori, T.; Sato, T. Development of a tiny orientation estimation device to operate under motion and magnetic disturbance. *Int. J. Robot. Res.* **2007**, *26*, 547–559.

25. Bouguet, J. Pyramidal Implementation of the Lucas Kanade Feature Tracker: Description of the Algorithm. Available online: http://robots.stanford.edu/cs223b04/algo_tracking.pdf (access on 22 November 2012).
26. Lucas, B.D.; Kanade, T. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Vancouver, BC, Canada, April 1981; pp. 674–679.
27. Shi, J.; Tomasi, C. Good Features to Track. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 21–23 June 1994; pp. 593–600.
28. Yang, Y.; Qixin, C.; Charles, L.; Zhen, Z. Pose Estimation Based on Four Coplanar Point Correspondences. In *Proceedings of the 6th International Conference on Fuzzy Systems and Knowledge Discovery*, Tianjin, China, 14–16 August 2009; pp. 410–414.
29. Abdel-Aziz, Y.I.; Karara, H.M. Direct Linear Transformation from Comparator Coordinates into Object Space Coordinates in Close-Range Photogrammetry. In *Proceedings of the Symposium on Close-Range Photogrammetric*, Urbana, IL, USA, January 1971; pp. 1–18.
30. Singer, R.A. Estimating optimal tracking filter performance for manned maneuvering targets. *IEEE Trans. Aerosp. Electron. Syst.* **1970**, *6*, 473–483.
31. Brown, D.C. Decentering distortion of lenses. *Photogramm. Eng.* **1966**, *32*, 444–462.
32. Bouguet, J.Y. Camera Calibration Toolbox for Matlab. Available from: http://www.vision.caltech.edu/bouguetj/calib_doc/ (accessed on 22 November 2012).
33. Harris, C.; Stephens, M. A combined Corner and Edge Detector. In *Proceedings of 4th Alvey Vision Conference*, Manchester, UK, 31 August–2 September 1988; pp. 147–151.
34. Ferraris, F.G.U.; Parvis, M. Procedure for effortless in-field calibration of three-axial rate gyro and accelerometers. *Sens. Mater.* **1995**, *7*, 311–330.
35. Skog, I.; Handel, P.; Nilsson, J.O.; Rantakokko, J. Zero-velocity detection—An algorithm evaluation. *IEEE Trans. Biomed. Eng.* **2010**, *57*, 2657–2666.
36. Schweighofer, G. Robust pose estimation from a planar target. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 2024–2030.
37. Dannilidis, K.; Nagel, H.H. The Coupling of Rotation and Translation in Motion Estimation of Planar Surfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, New York, NY, USA, 15–17 June 1993; pp. 188–193.
38. Bleser, G.; Stricker, D. Advanced Tracking Through Efficient Image Processing and Visual-Inertial Sensor Fusion. In *Proceedings of the IEEE International Conference on Virtual Reality*, Reno, NV, USA, March 2008; pp. 137–144.
39. Sabatini, A.M. Variable-state-dimension kalman-based filter for orientation determination using inertial and magnetic sensors. *Sensors* **2012**, *12*, 8491–8506.
40. Armesto, L.; Tornero, J.; Vincze, M. Fast ego-motion estimation with multi-rate fusion of inertial and vision. *Int. J. Robot. Res.* **2007**, *26*, 577–589.