Visuo-inertial fusion for homography-based filtering and estimation

Alexandre Eudes, Pascal Morin, Robert Mahony and Tarek Hamel

***Abstract*— The paper concerns visuo-inertial filtering and estimation based on homography and angular velocity measurements, i.e. data obtained from a mono-camera/IMU sensor. We extend recently developed nonlinear filters on the special linear group of homographies to the estimation of scene parameters and velocity of the sensor. A validation of the proposed solution and a comparative evaluation based on real data is presented.**

I. Introduction

Many robotic applications rely on the real-time estimation of the homography matrix that relates two images of the same planar scene. Among such applications, let us mention several recent results on homography-based visual servoing for ground [2], [8], [10], aerial [4], [17], or underwater robotics [5], [18]. SLAM in locally planar environments also makes use of homography matrices extensively [3], [16]. Several vision algorithms have been developed to compute the homography matrix from two image data sets. One can distinguish feature-based techniques, which rely on the extraction and matching of characteristic features in the image [1], [11] from direct methods, which process a dense pixel data set [16]. All these methods are computationally expensive. This can be a major issue for real-time onboard applications, e.g. in aerial robotics. Furthermore, many methods based on optimization algorithms (e.g. direct methods) only yield a small domain of convergence. This is another issue since the algorithm may fail in case of "large" displacements between consecutive images. Loss of local convergence, which is independent of the computational power, is often encountered in practice. Fusing the visual information with other sensory data can help to reduce these difficulties.

This paper concerns the fusion of visual and inertial data in the context of homography estimation. Many results have been reported recently on visuo-inertial fusion based on cartesian pose (i.e. position and orientation) measurements([14],[9],[15]). Indeed, data provided by an IMU (Inertial Measurement Unit) are directly related to the time-derivatives of the cartesian pose. Using inertial data in the context of homography estimation is much more challenging because the relation between the homography matrix and the cartesian pose is complex and depends on unknown quantities (reference plane normal and distance to the scene). A possibility would consist in first decomposing the homography matrix so as to recover cartesian pose (up to a scale factor) and then using existing fusion techniques for cartesian measurements. Multiple solutions to the decomposition problem [13], however, make such an approach difficult to implement when no a priori information on the visual environment is available [3].

In this paper, we build on a recently published article [12] in which nonlinear filters have been proposed to fuse homography data with inertial measurements. The contribution is twofold. Firstly, we extend the solution proposed in [12] in order to estimate additional variables, namely the normal to the scene and linear velocity (up to the scale factor) in body frame. This information can be exploited at different levels, e.g. for scene reconstruction, feedback control, or initialization of the vision algorithm. Secondly, we validate the results, including the filters proposed in [12], with real-data and compare the performance obtained with different solutions. This comparison shows the strong benefit resulting from the use of inertial data.

The present paper is related to SLAM results concerning motion and structure estimation based on mono-camera and inertial data (see, e.g. [6], [16]). With respect to those works, the objective of this paper is more limited since we essentially focus on the homography estimation without trying to recover the cartesian pose. The fusion methods proposed in this paper differ from those typically used in the SLAM context (nonlinear complementary filters versus the EKF). Complementary filters are of interest for their computational simplicity, large stability domains and explicit stability conditions, and consistence of the estimates with the geometry of the observation space (i.e., group of homographies, etc). We believe that the results we present have significance for the more general SLAM problem.

The paper is organized as follows. Section II provides technical background, including a brief review of the main results in [12]. Section III presents the extension of the results in [12] for normal and velocity estimation. Section IV concerns the validation and comparative evaluation of the proposed filters. The paper ends with some concluding remarks and perspective for future work.

II. Technical background

A. Notation

- $S(x)$ is the antisymmetric matrix associated with the cross product by x, i.e. $S(x)y = x \times y$ with $\times$ the cross product.
- I_n is the $n \times n$ identity matrix and $0_n \in \mathbb{R}^n$ is the null vector.

A. Eudes and P. Morin are with ISIR-UPMC, Paris, France, eudes@isir.upmc.fr, morin@isir.upmc.fr. These authors have been supported by the "Chaire d'excellence en Robotique RTE-UPMC"

Tarek Hamel is with I3S-CNRS, Nice-Sophia Antipolis, France, thamel@i3s.unice.fr

Robert Mahony is with Department of Engingeering, ANU, ACT, 0200, Australia, Robert.Mahony@anu.edu.au

- $\mathrm{tr}(M)$ is the trace of the matrix M and $\det(M)$ its determinant.
- The Special Linear Group $SL(3)$ and its Lie algebra $\mathfrak{sl}(3)$ are :

$$\begin{array}{rcl} SL(3) & = & \left\{ H \in \mathbb{R}^{3\times 3} \quad | \quad \det(H) = 1 \quad \right\} \\ \mathfrak{sl}(3) & = & \left\{ A \in \mathbb{R}^{3\times 3} \quad | \quad \mathrm{tr}(A) = 0 \quad \right\} \end{array}$$

- The adjoint operator $\mathrm{Ad} : SL(3) \times \mathfrak{sl}(3) \longrightarrow \mathfrak{sl}(3)$ is defined by

$$\mathrm{Ad}_G X = GXG^{-1}, \quad G \in SL(3), X \in \mathfrak{sl}(3)$$

- $\mathbb{P} : \mathbb{R}^{3\times 3} \longrightarrow \mathfrak{sl}(3)$ is the orthogonal projector defined by

$$\mathbb{P}(M) = M - \frac{\mathrm{tr}(M)}{3} I_3, \quad M \in \mathbb{R}^{3\times 3}$$

B. Homographies

We recall below well known facts about homography matrices (see, e.g., [11] for details). Consider two images $\mathbf{I}_A$ and $\mathbf{I}_B$ of the same planar scene taken by a monocular camera. Each image $\mathbf{I}_*$ ($* \in \{A, B\}$) is taken from a specific pose of the camera and we denote by $\mathcal{F}_*(* \in \{A, B\})$ an associated camera frame with origin corresponding to the optical center of the camera and third basis vector aligned with the optical axis. Furthermore, we denote by d_* and n_* respectively the distance from the origin of $\mathcal{F}_*$ to the planar scene and the normal to the scene expressed in $\mathcal{F}_*$.

Let R denote the rotation matrix from $\mathcal{F}_B$ to $\mathcal{F}_A$ and $p \in \mathbb{R}^3$ the coordinate vector of the origin of $\mathcal{F}_B$ expressed in $\mathcal{F}_A$. Then, the following classical relations hold:

$$d_B = d_A - n_B^T R^T p, \quad n_B = R^T n_A$$

The raw homography matrix

$$G = \gamma K (R + \frac{p n_B^T}{d_B}) K^{-1} \tag{1}$$

with K the calibration matrix of the camera and γ a scale factor, maps pixel coordinates from $\mathbf{I}_B$ to $\mathbf{I}_A$. Without loss of generality, γ can be chosen to scale the determinant of G to ensure that $\det(G) = 1$ and G is in $SL(3)$. Note that the scale factor γ can be computed as the second singular value of G (see [11],[13]). After calibration of the camera, the (Euclidean) homography matrix

$$H = K^{-1} G K = \gamma (R + \frac{p n_B^T}{d_B})$$

is obtained. Note that H and G have the same singular values and hence if G is scaled to lie in $SL(3)$ then $H \in SL(3)$. This matrix maps Euclidean coordinates of the scene's points from $\mathcal{F}_B$ to $\mathcal{F}_A$ and that γ is related to the distances from the scene to each camera frame: $\gamma^3 = \frac{d_B}{d_A}$.

C. Complementary filters on SL(3)

Several nonlinear observers have been proposed in [12] depending on availability of inertial information. We are interested here by two cases.

The first observer considered is based on the general form of the kinematics on $SL(3)$:

$$\dot{H} = H X \tag{2}$$

with $H \in SL(3)$ and $X \in \mathfrak{sl}(3)$. The observer is given by

$$\begin{cases} \dot{\hat{H}} = \hat{H}\, \mathrm{Ad}_{\tilde{H}} \left(\hat{X} - k_1 \mathbb{P} \left(\tilde{H}(I_3 - \tilde{H}) \right) \right) \\ \dot{\hat{X}} = -k_2 \mathbb{P} \left(\tilde{H}(I_3 - \tilde{H}) \right) \end{cases} \tag{3}$$

with $\hat{H} \in SL(3), X \in \mathfrak{sl}(3), \tilde{H} = \hat{H}^{-1} H$. It is shown in [12] that this observer ensures almost global asymptotic stability of $(I_3, 0)$ for the estimation error $(\tilde{H}, \tilde{X}) = (\hat{H}^{-1} H, X - \hat{X})$ (i.e., asymptotic convergence of the estimates to the original variables) provided that X is constant (see [12, Th. 3.2] for details). Although this condition is seldom satisfied in practice, this observer provides a simple solution to the problem of filtering homography measurements. Finally, note that this observer uses homography measurements only.

A second observer, that explicitly takes into account the rigid body kinematics of the camera motion, is proposed in [12]. The kinematics of the camera frame of reference is given by

$$\begin{cases} \dot{R} = R S(\omega) \\ \dot{p} = R V \end{cases} \tag{4}$$

with ω the angular velocity of $\mathcal{F}_B$ w.r.t. $\mathcal{F}_A$ expressed in $\mathcal{F}_A$ and V the linear velocity of $\mathcal{F}_B$ w.r.t. $\mathcal{F}_A$ expressed in $\mathcal{F}_B$. With this notation, one can show that the group velocity X in (2) is given by

$$\begin{aligned} X &= S(\omega) + \frac{V n_B^T}{d_B} - \frac{V n_B^T}{3 d_B} I_3 \\ &= S(\omega) + \frac{1}{\gamma^3} \mathbb{P}(M) \end{aligned}$$

with

$$M = \frac{V}{d_A} n_B^T \tag{5}$$

The following observer of H and M is proposed in [12]:

$$\begin{cases} \dot{\hat{H}} = \hat{H}\, \mathrm{Ad}_{\tilde{H}} \left(S(\omega) + \frac{1}{\gamma^3} \mathbb{P} \left(\hat{M} \right) - k_1 \mathbb{P} \left(\tilde{H}(I_3 - \tilde{H}) \right) \right) \\ \dot{\hat{M}} = \hat{M} S(\omega) - \frac{k_2}{\gamma^3} \mathbb{P} \left(\tilde{H}(I_3 - \tilde{H}) \right) \end{cases} \tag{6}$$

with $\hat{H} \in SL(3), \hat{M} \in \mathbb{R}^{3\times 3}$ and $\tilde{H} = \hat{H}^{-1} H$.

First, let us remark that this observer relies on measurements of H and ω. Thus, it can be implemented with a monocamera/IMU sensor since ω can be given by the IMU's rate gyro. Conditions under which the estimates $(\hat{H}, \hat{M})$ almost globally converge to (H, M) are given in [12, Cor. 5.5]. These conditions essentially reduce to the following: *i)* ω is persistently exciting, and *ii)* V is constant. The hypothesis of persistent excitation on the angular velocity is used to demonstrate the convergence of $\hat{M}$ to M. In the case of lack of persistent excitation, $\hat{M}$ converges only to $M + a(t) I_3$ with $a(t) \in \mathbb{R}$ but the convergence of $\hat{H}$ to H still holds. The

hypothesis of V constant is a strong assumption. Asymptotic stability of the observer for V constant, however, guarantees that the observer can provide accurate estimates when V is slowly time varying with respect to the filter dynamics. This will be illustrated later in the paper and experimentally verified.

III. Normal and velocity estimation

Observer (6) provides estimates of the matrix M in (5). The objective of this section is to show how to exploit these estimates in order to estimate the normal n_B and the velocity V up to a scale factor. Two difficulties must be overcome. First, as explained above, $\hat{M}$ is only guaranteed to converge to $M + a(t)I_3$ where a is unknown. It is evident from (5) then, that M contains no information on n_B when $V = 0$. This is a well known unobservability problem of the scene structure when the camera is motionless. Thus, one can only expect to obtain an accurate estimate of n_B in the presence of linear motion.

A. Estimation of the normal from $\hat{M}$

In order to simplify the notation, let $v = \frac{V}{d_A}$. Upon convergence of $\hat{M}$ to $M + aI_3$, it follows from (5) that

$$\begin{cases} \operatorname{tr}(\hat{M}) = & 3a + v^T n_B \\ \det(\hat{M}) = & a^3 + a^2 v^T n_B \end{cases} \tag{7}$$

Thus, a is solution of the third-order polynomial equation:

$$2a^3 - \operatorname{tr}(\hat{M})a^2 + \det(\hat{M}) = 0 \tag{8}$$

the roots of which can be computed explicitly. For each root a_k of this equation, one can compute $\bar{M}_k := \hat{M} - a_k I_3$. Observe that, upon convergence of $\hat{M}$ to $M + aI_3$ and when $a_k = a$,

$$\bar{M}_k = \hat{M} - a_k I_3 = \begin{pmatrix} v_1 n_B^T \\ v_2 n_B^T \\ v_3 n_B^T \end{pmatrix} \tag{9}$$

Since the last coordinate of the homography's normal is positive and $\|n_B\| = 1$, this implies that

$$v_i = \operatorname{sign}(\bar{M}_{k,(i,3)})\|\bar{M}_{k,i}\| \text{ and } n_B^T = \frac{\bar{M}_{k,i}}{v_i} \tag{10}$$

with $\bar{M}_{k,i}$ denoting the i-th row vector of $\bar{M}_k$. Equation (10) suggests to compute for each k an estimate v_k of v as follows:

$$v_{k,i} = \operatorname{sign}(\bar{M}_{k,(i,3)})\|\bar{M}_{k,i}\|$$

with $\bar{M}_{k,(i,j)}$ the (i,j) element of $\bar{M}_k$. Using (10) again, we propose to compute for each k a first estimate $n_{B,k}$ of n_B by solving a weighted linear least square problem:

$$\min_{n_{B,k}} \sum_{i \in L_k} \|\bar{M}_{i,k}\|^2 \|n_{B,k} - \frac{\bar{M}_{k,i}^T}{v_{k,i}}\|^2 \tag{11}$$

where $L_k = \{i \in 1,2,3 / \|\bar{M}_{k,i}\| > \delta_1\}$ with δ_1 positive. The idea is to limit the effect of measurement noise by weighting the pseudo-measurements $\frac{\bar{M}_{k,i}}{v_{k,i}}$ by $\|\bar{M}_{k,i}\|$ and discarding these measurements when $\|\bar{M}_{k,i}\| = |v_{k,i}|$ is smaller than δ_1, knowing that n_B is not observable when $v = 0$. The solution to this optimization problem is given by

$$n_{B,k} = \frac{\sum_{i \in L_k} \operatorname{sign}(\bar{M}_{k,(i,3)})\|\bar{M}_{k,i}\|\bar{M}_{k,i}^T}{\sum_{i \in L_k} \|\bar{M}_{k,i}\|^2} \tag{12}$$

Furthermore, $n_{B,k}$ is normalized to avoid numerical drift. Using the fact that $M(I - n_B n_B^T) = 0$, one can associate to each $n_{B,k}$ thus obtained the following score(ρ):

$$\rho_k = \|(\hat{M} - a_k I_3)(I - n_{B,k} n_{B,k}^T)\|$$

The value of k with the lowest score provides a first estimate n_B^* of n_B, an associated score ρ^*, and a candidate root value a^*. If $L_k = \emptyset$ for each k (i.e. $\|\bar{M}_{k,i}\| \leq \delta_1$ for each k and $i = 1,2,3$), meaning that there is not enough motion to deduce from $\hat{M}$ a reliable estimate of n_B, we set $n_B^* = e_3$ and $\rho^* = 1$. We will see further that this relatively arbitrary choice is unimportant.

B. Complementary filtering on the normal

The second step of the method is a complementary filter using the previously extracted normal and score (n_B^*, ρ^*) and the angular velocity measurement provided by the IMU. First, recall that $n_B = R^T n_A$ and $\dot{R} = RS(\omega)$, so that

$$\dot{n}_B = -S(\omega)n_B = S(n_B)\omega$$

Then the filter is built as a complementary filter on the unit sphere:

$$\dot{\hat{n}}_B = S(\hat{n}_B)[\omega - k(\rho^*)(\hat{n}_B \times n_B^*)]$$

with $k(\rho) = k_3(1+\exp(-100(\rho-\delta_2)))^{-1}$ and δ_2, k_3 positive scalars. Since $k(x)$ tends to zero when x is large, the role of the varying gain $k(\rho^*)$ is to use innovation in the filter only when the score associated with n_B^* is low enough. Otherwise, $\hat{n}_B$ is updated from angular velocity measurements only. Let us briefly establish the stability of this filter. Consider the candidate Lyapunov function

$$\mathcal{L} = 1 - n_B^T \hat{n}_B$$

Then, if $n_B^* = n_B$,

$$\begin{aligned} \dot{\mathcal{L}} &= -\dot{n}_B^T \hat{n}_B - n_B^T \dot{\hat{n}}_B \\ &= -(-S(\omega)n_B)^T \hat{n}_B - n_B^T S(\hat{n}_B)\omega \\ &\quad + k(\rho^*) n_B^T S(\hat{n}_B)(\hat{n}_B \times n_B) \\ &= -k(\rho^*)\|S(\hat{n}_B)n_B\|^2 \end{aligned} \tag{13}$$

Thus, this filter is stable and almost globally convergent, i.e. provided that $\hat{n}_B(0) \neq -n_B(0)$.

Finally, we obtain an estimate of v (i.e. velocity in body frame up to the scale factor d_A) as:

$$\hat{v} = (\hat{M} - a^* I_3)\hat{n}_B \tag{14}$$

IV. Validation and comparative evaluation

In this section we validate the estimation algorithms presented in the previous sections and evaluate their performance. These algorithms are used in conjunction with an ESM homography computer vision algorithm[2].For each image, the filter is used to make a prediction of the raw homography that is used as initialization of the ESM algorithm. The visual method provides two results: an homography estimation and the correlation score between the current image and the reference image. If the correlation score is good enough (> 0.85) the estimate is considered as "good" and is used as measurement in the filter. Three algorithms are compared:

1) ESM algorithm alone: the visual method is initialized by the homography estimated at the previous frame. This method is named ESMonly thereafter;
2) Filter (3): this filter, which uses no IMU data, is named filternoIMU hereafter;
3) Filter (6): this filter, which uses the rate gyro data, is named filterIMU.

Three issues are investigated: *i)* tracking quality and ability of the filter to follow the pattern in the presence of fast dynamics; *ii)* ability of the filter to interpolate the homography between two frames and provide estimation of the homography at higher rate; *iii)* quality of the normal and velocity estimation algorithm.

A. Experimental setup

We make use of a sensor consisting of a xSens MTiG IMU working at a frequency of 200 [Hz], and an AVT Stingray 125B camera that provides 40 images of 800×600 [pixel] resolution per second. The camera and the IMU are synchronised. The camera uses wide-angle lenses (focal 1.28 [mm]). The target is placed over a surface parallel to the ground and is printed out on a 376×282 [mm] sheet of paper to serve as a reference for the visual system. The reference image is 320×240 [pixel]. So the distance d_A can be determined as 0.527[m]. The processed video sequence presented in the accompanying video is 1321 frames long and presents high velocity motion (rotations up to 5[rad/s], translations, scaling change) and occlusions. In particular, a complete occlusion of the pattern occurs little after $t = 10$[s].

Four images of the sequence are presented on Figure 1. A "ground truth" of the correct homography for each frame of the sequence has been computed thanks to a global estimation of the homography by SIFT followed by the ESM algorithm. If the pattern is lost, we reset the algorithm with the ground-truth homography. The sequence is used at different sampling rates to obtain more challenging sequences and evaluate the performances of the proposed filters.

For both filters (3) and (6), the estimation gains have been chosen as $k_1 = 25$ and $k_2 = 250$. Following the notation of the description available at *http://esm.gforge.inria.fr/ESM.html*, the ESM algorithm is used with the following parameter values: $prec = 2$,$iter = 50$.

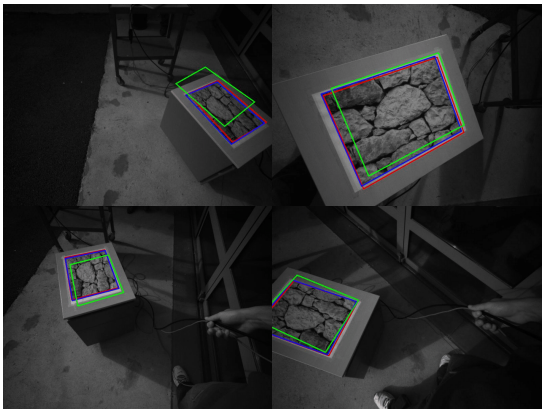

Fig. 1. Four images of the sequence at 20[Hz]: pattern position at previous frame (green), vision estimate (blue), and prediction of the filterIMU (red).

B. Tracking quality

In this section we measure the quantitative performance of the different estimators. This performance is reflected by the number of frames for which the homography is correctly estimated. We use the correlation score computed by the visual method to discriminate between well and badly estimated frames. A first tracking quality indicator is the percentage of well estimated frames. This indicator will be labelled as "%track". Another related criteria concerns the number of time-sequences for which estimation is successful. For that, we define a track as a continuous time-sequence during which the pattern is correctly tracked. We provide the number of tracks in the sequence (label "nb track") and also the mean and the maximum of track length. Table I presents the obtained results for the full sequence at different sampling rates (40[Hz], 20[Hz], 10[Hz]).

The ESMonly estimator works well at 40[Hz] since 95% of the sequence is correctly tracked but performance rapidly decreases as distance between images grow (72% at 20[Hz], and only 35% at 10[Hz]). It must be noted that the ESM estimator parameters are tuned for speed and not for performance, having in mind real-time applications.

The filternoIMU estimator outperforms the ESMOnly filter on the sequence at 40[Hz]. Tracks are on average twice longer and many losses of the pattern are avoided (11 tracks versus 19 for ESMonly). At 20[Hz] the performance is still better but the difference between these two solutions reduces. At 10[Hz] the filter degrades performance.

The filterIMU tracks almost all the sequence at both 40[Hz] and 20[Hz]. There is just one tracking failure, which occurs around time $t = 10$[s] due to the occlusion of the visual target. Improvement provided by the IMU is clearly demonstrated. At 10[Hz], the performance significantly deteriorates but this filter still outperforms the other ones.

Let us finally remark that these performances are obtained despite the fact that the assumption of constant velocity in body frame (upon which the filter stability was established) is violated, as can be seen on the video and on the velocity data presented further.

Frame rate	Method	%track	nb track	track length	
				mean	max
40Hz	ESM Only	94.31	19	65.36	463
	FilternoIMU	97.74	11	114.27	607
1321 img	FilterIMU	98.78	2	646.5	915
20Hz	ESM Only	72.38	59	8.0	89
	FilternoIMU	80.5	52	10.17	94
660 img	FilterIMU	97.42	2	321.5	456
10Hz	ESM Only	38.79	46	2.78	27
	FilternoIMU	32.36	58	1.72	4
330 img	FilterIMU	58.66	59	3.27	27

TABLE I

RATE OF GOOD TRACK FOR DIFFERENT FRAME-RATES AND METHODS: PERCENTAGE OF WELL ESTIMATED FRAMES, NUMBER OF TRACKS, MEAN AND MAXIMUM TRACK LENGTH ON THE SEQUENCE

C. Prediction quality

The results reported in this section have been obtained with the FilterIMU and the video sequence at 20[Hz]. To evaluate the prediction quality we consider the error between the ground truth and the predicted homographies. A first comparison could be made by considering the matrix norm of the difference (in $SL(3)$) between these two homographies. In order to obtain a more precise comparison, we decompose the raw homography transformation (2D) into elementary 2D transformations (translation, rotation, ...). Indeed, a raw homography can be uniquely decomposed as a product of a similarity transformation by an affine transformation, and finally by a projective transformation ([7]):

$$G = G_{33} \begin{pmatrix} sR(\theta) & t \\ 0_2^T & 1 \end{pmatrix} \begin{pmatrix} R(\phi)DR(\phi)^T & 0_2 \\ 0_2^T & 1 \end{pmatrix} \begin{pmatrix} I_2 & 0_2 \\ \ell^T & 1 \end{pmatrix}$$

with

$$R(\alpha) = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}, \qquad D = \begin{pmatrix} s_1 & 0 \\ 0 & \frac{1}{s_1} \end{pmatrix}$$
$$t = \begin{pmatrix} t_x & t_y \end{pmatrix}^T, \ \ell = \begin{pmatrix} \ell_1 & \ell_2 \end{pmatrix}^T, \qquad \phi \in [0, \frac{\pi}{2}]$$

The 8 parameters of this decomposition represent: a rotation parameter θ, two translation parameters t_x, t_y, one scale factor $s > 0$, two anisotropic scaling parameters with the shear angle ϕ and related scale s_1, and finally two parameters ℓ_1, ℓ_2 related to infinite line behaviour (vanishing point, horizon).

Let

$$T^{\text{oref}} = \begin{pmatrix} I_2 & \frac{pref}{2} \\ 0_2^T & 1 \end{pmatrix}$$

with $pref$ the size in [px] of the reference pattern image I_A (320×240 in our experimental setup). We will decompose $G^{-1}T^{\text{oref}}$ since G^{-1} provides an error in the current image and the shift by the translation T^{oref} allows us to split-up translation and rotation, i.e. the translation t is the vector from origin (top left corner) to the center of the pattern in the image and it is not corrupted by other parameters.

We are interested by two comparisons with respect to the ground-truth: the homography predicted by the FilterIMU ($pred$) and the homography obtained after visual processing of the last image (vis). For each frame used by the filter, we compare coefficients of the decomposition of $G_*^{-1}T^{\text{oref}}$ ($* \in \{pred, vis\}$) with coefficients of the decomposition of the ground truth homography $G_{gt}^{-1}T^{\text{oref}}$ by using the following error:

$$cmp(*) = (t_x^* - t_x^{gt}, t_y^* - t_y^{gt}, \theta^* - \theta^{gt}, \phi^* - \phi^{gt}, \frac{s^*}{s^{gt}}, \frac{s_1^*}{s_1^{gt}}, \frac{\ell_1^*}{\ell_1^{gt}}, \frac{\ell_2^*}{\ell_2^{gt}}) \qquad * \in \{pred, vis\}$$

Table II presents parameter statistics of the comparison with the ground-truth of the prediction ($cmp(pred)$) and the previous image ($cmp(vis)$). Figure 4 presents the comparison on θ, s, t_y for prediction and previous image homographies.

From the translation part of Figure 4, we see that the filter gives good prediction of the translation. Some errors still remain (3.1[px] in average, max 22[px] in the portion of the sequence with important scale changes) but by comparison with vision homography (11.5[px] in average and 66[px]max) the errors remain small. This figure shows that the filter works well for fast linear motions, although the assumption of constant velocity used in the stability analysis is not satisfied.

Inspection of the rotationnal part (variable θ) shows that the IMU is very reliable as the rotation error is about 0.2[deg] with very low standard deviation whereas the vision homography rotation is about (3[deg] in average). For other coefficients, the error for the predicted homography is smaller on average than the error for the previous image homography and in most cases this remains true all along the sequence.

All these results show that the prediction is accurate and is always better than the measurement from previous image. In practice, the prediction could be used as output of the filter thus allowing to have homography estimates at IMU rate. All the more so as the prediction process is fast, so it does not much increase latency.

	Prediction from i_0 to i_0+2		Previous image i_0	
	compared with ground truth on i_0+2			
error on	mean	std	mean	std
t_x [px]	3.09	3.14	11.4	10
t_y [px]	2.56	2.81	13.2	11.3
θ [deg]	0.222	0.155	3	2.64
ϕ[deg]	0.00322	0.0029	0.0114	0.0105
ℓ_1	2.71e-05	9.53e-06	3.45e-05	3.23e-05
ℓ_2	1.04e-05	5.91e-06	3.12e-05	2.96e-05
s	0.999	0.00681	0.999	0.0439
s_1	1.01	0.00443	0.999	0.0379

TABLE II

PREDICTION QUALITY STATISTICS: MEAN AND STANDARD VARIATION FOR THE ERROR BETWEEN GROUD-TRUTH AT i_0+2 AND PREDICTED HOMOGRAPHY AT i_0+2/VISION HOMOGRAPHY AT i_0.

D. Normal and velocity estimation

This section concerns the evaluation of the normal and velocity estimator proposed in Section III.

The ground-truth for the normal is obtained by decomposing (see [11] p.136 for the decomposition algorithm) the inverse of the ground-truth homography in 3D rotation, translation and normal:

$$H^{-1} = \frac{1}{\gamma}(R^T - \frac{1}{d_A}R^T p n_A^T) \tag{15}$$

From the decomposition we get $(R^T, \frac{1}{d_A}R^T p, n_A)$ The decomposition is not unique and gives us two possible homographies and thus two normals. Since the target (sheet of paper) was placed on the ground, we know that the normal is quasi-vertical and we keep as ground-truth the solution closest to the vertical. The ground-truth normal in body frame can be computed as $n_B = R^T n_A$. An approximation $\hat{v}_{gt}$ of the ground-truth velocity is obtained applying a derivative filter to the ground-truth position $\frac{p}{d_A}$ (i.e., position up to scale factor obtained by decomposition of the ground-truth homography):

$$\begin{cases} \dot{\hat{p}}_{gt} = \hat{v}_{gt} - k_4(\hat{p}_{gt} - \dfrac{p}{d_A}) \\ \dot{\hat{v}}_{gt} = -k_5(\hat{p}_{gt} - \dfrac{p}{d_A}) \end{cases} \tag{16}$$

with $k_4 = 27$ and $k_5 = 225$.

Algorithms proposed in Section III are applied with the following parameters: $k_3 = 2$, $\delta_1 = 0.33, \delta_2 = 0.3$. The 20[Hz] sequence is used and the filter for normal estimation is initialized at random for the two tracks of the sequence (recall that the pattern is lost around $t = 10$[s] due to an occlusion of the visual target). Figure 2 shows the angle error in degree between the ground-truth n_B and the value $\hat{n}_B$ estimated by the filter. One can observe the fast convergence of $\hat{n}_B$ to the ground-truth despite the fact that the initial error is large at the begining of each track, thus validating the claim for large stability domains of the proposed filters. After the transcient convergence phase the error remains small ($\approx$ 5[deg]). It must be noted that when the linear velocity is very small the filter relies only on the gyrometer measurements (normal is not observable in this case). Drift could occur in this situation and capacity of the filter to converge from large initial errors is all the more important.

Figure 3 shows a comparison of the velocity at a scale factor $\hat{v}$ estimated by the filter (i.e. Eq. (14)) and the ground-truth $\hat{v}_{gt}$. For legibility of the figures, we only show results on the time-interval $[15; 35]$s. Horizontal and vertical velocity are well estimated, with a small time-lag of about 80[ms]. The velocity in z (in the optical axis direction) is less reliable with the presence of offsets.

Conclusion

We have presented experimental validations of new non-linear filters for visuo-inertial fusion and extensions of these filters have been proposed in order to recover the visual target normal and velocity in body-frame (up to a scale factor). Validations show the efficiency of the fusion algorithms and the capacity of the filters to converge from large initial errors. Extensions of this work are multiple. The proposed solution allows one to reduce the frequency of vision acquisition/processing. This is a key aspect in real-time embarked applications. Exploiting accelerometer measurements may reduce the frequency of vision processing still more. The quality of normal estimation also suggests to make a step further in the direction of estimation in cartesian space. Finally, we plan to apply this work to the feedback control of UAVs with limited computing power.

References

[1] A. Agarwal, C.V. Jawahar, and P. Narayanan. A survey of planar homography estimation techniques. Technical report, Centre for Visual Information Technology, 2005.

[2] S. Benhimane and E. Malis. Homography-based 2d visual tracking and servoing. *International Journal of Computer Vision*, 2007.

[3] F. Caballero, L. Merino, J. Ferruz, and A. Ollero. Vision-based odometry and slam for medium and high altitude flying uavs. *Unmanned Aircraft Systems*, 27:137–161, 2009.

[4] T.F. Gonçalves, J.R. Azinheira, and P. Rives. Homography-based visual servoing of an aircraft for automatic approach and landing. In *IEEE Conf. on Robotics and Automation*, pages 9–14, 2010.

[5] R. Garcia, J. Batle, X. Cufi, and J. Amat. Positioning an underwater vehicle through image mosaicking. In *IEEE Conf. on Robotics and Automation*, pages 2779–2784, 2001.

[6] P. Gemeiner, P. Einramhof, and M. Vincze. Simultaneous motion and structure estimation by fusion of inertial and vision data. *The International Journal of Robotics Research*, 26:591–605, 2007.

[7] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*, volume 2. Cambridge Univ Press, 2000.

[8] J.Chen, W.E. Dixon, D.M. Dawson, and M. McIntyre. Homography-based visual servo tracking control of a wheeled mobile robot. *IEEE Trans. on Robotics*, 22:407–416, 2006.

[9] Jonathan Kelly and Gaurav S Sukhatme. Visual-inertial sensor fusion: Localization, mapping and sensor-to-sensor self-calibration. *The International Journal of Robotics Research*, 30(1):56–79, 2011.

[10] G. López-Nicolás, N.R. Gans, S. Bhattacharya, C. Sagüés, J.J. Guerrero, and S. Hutchinson. Homography-based control scheme for mobile robots with nonholonomic and field-of-view constraints. *IEEE Trans. on Systems, Man, and Cybernetics: Part B*, 10:1115–1127, 2010.

[11] Y. Ma, S. Soatto, J. Kosecka, and S.S. Sastry. *An Invitation to 3-D Vision: From Images to Geometric Models*. SpringerVerlag, 2003.

[12] R. Mahony, T. Hamel, P. Morin, and E. Malis. Nonlinear complementary filters on the special linear group. *International Journal of Control*, 85:1557–1573, 2012.

[13] E. Malis and M. Vargas. Deeper understanding of the homography decomposition for vision-based control. Technical Report 6303, INRIA, 2007. Available at http://hal.inria.fr/inria-00174036/fr/.

[14] Faraz M Mirzaei and Stergios I Roumeliotis. A kalman filter-based algorithm for imu-camera calibration: Observability analysis and performance evaluation. *Robotics, IEEE Transactions on*, 24(5):1143–1156, 2008.

[15] G. Scandaroli, P. Morin, and G. Silveira. A nonlinear observer approach for concurrent estimation of pose, imu bias and camera-to-imu rotation. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 3335–3341, 2011.

[16] F. Servant, P. Houlier, and E. Marchand. Improving monocular plane-based slam with inertial measures. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 3810–3815, 2010.

[17] D. Suter, T. Hamel, and R. Mahony. Visual servo control using homography estimation for the stabilization of an x4-flyer. In *IEEE Conference on Decision and Control*, 2002.

[18] S. van der Zwaan and J. Santos-Victor. Real-time vision-based station keeping for underwater robots. In *OCEANS, 2001. MTS/IEEE Conference and Exhibition*, pages 1058–1065, 2001.

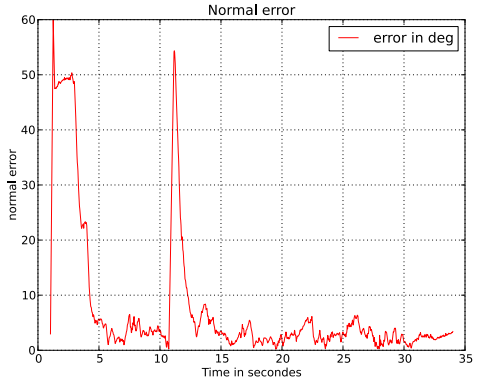

Fig. 2. Estimation error (in degree) between the normal and its estimate.

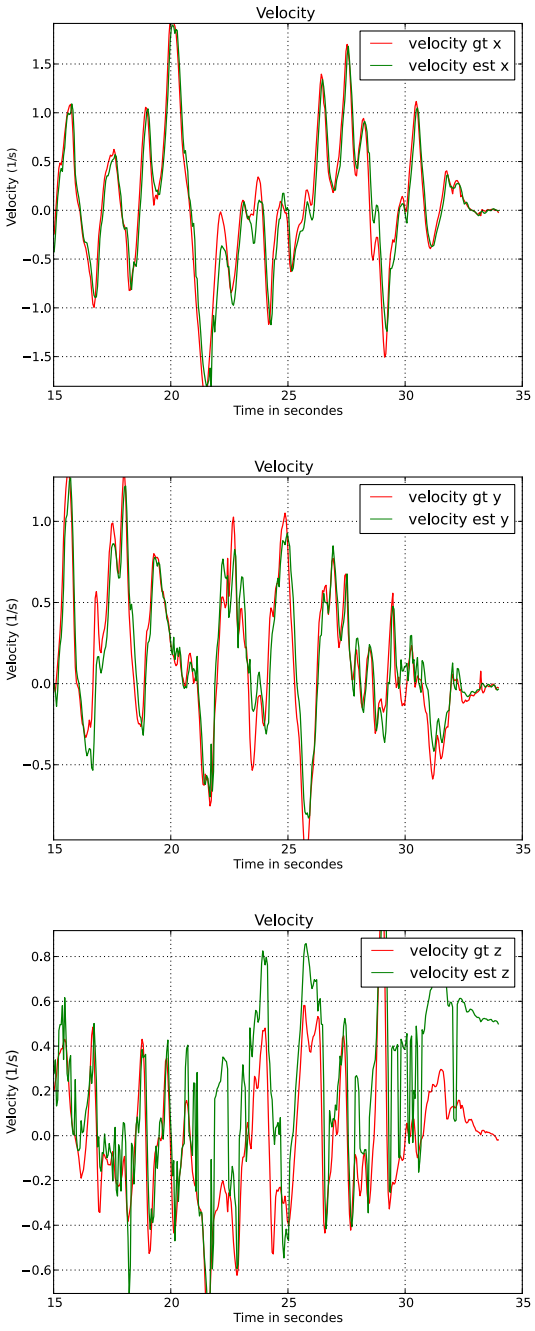

Fig. 3. Estimation of velocity at a scale factor (≈1.9 real velocity) in body frame: estimated velocity (green) and ground-truth (red).

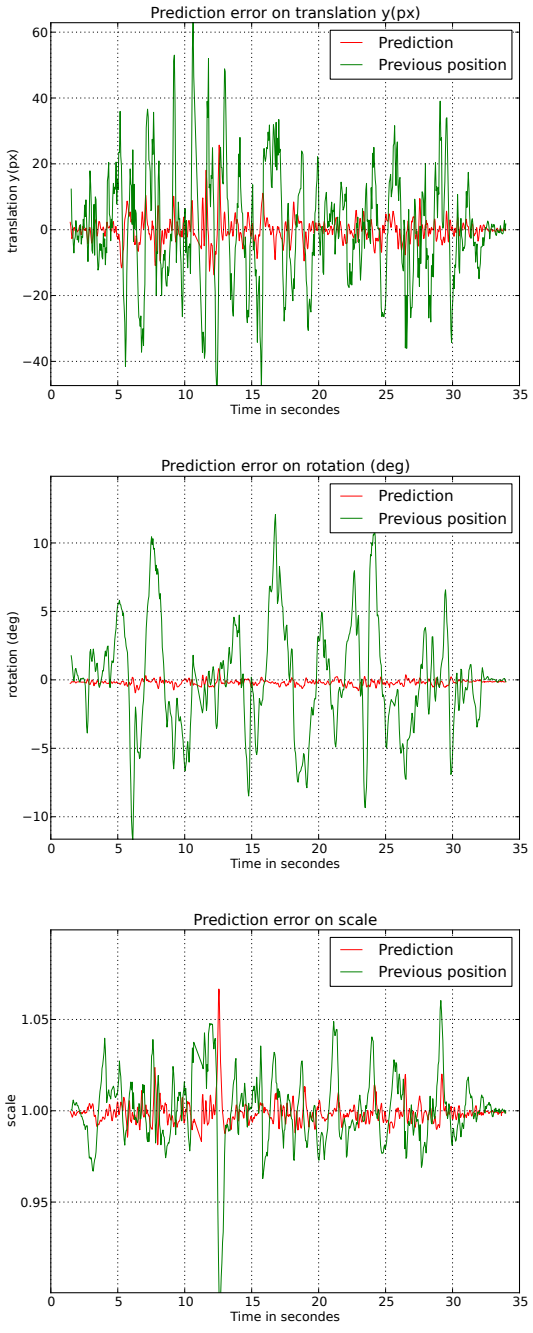

Fig. 4. Prediction quality: mean and standard variation for the error between groud-truth and the predicted H (in red) and the error for the homography in previous frames in green for, from top to bottom: translation on y, rotation, scale error.