2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN), 30 Sept. - 3 Oct. 2019, Pisa, Italy

Stationary Detector for Monocular Visual-Inertial SLAM

Richard Guillemard, François Hélénon, Bruno Petit, Vincent Gay-Bellile and Mathieu Carrier
CEA, LVML, Laboratoire de Vision pour la Modélisation et la Localisation
Point Courrier 94, Gif-sur-Yvette, F-91191 France
Email: {richard.guillemard, bruno.petit2, vincent.gay-bellile, mathieu.carrier}@cea.fr, francois.helenon@wanadoo.fr

***Abstract*— Monocular Visual-Inertial SLAM (VISLAM) algorithms are very popular solutions for accurate indoor localization. However, they may suffer from speed divergence when the system is at rest as illustrated on Figure 1. In this paper we propose to tackle this issue. For that we investigate the detection of time epochs when a visual-inertial sensor rig is stationary. Two kind of stops are deduced from raw sensor data. SoftStop when the system is at rest with a slight movement noise (*e.g* a human at rest holding the system) and HardStop when the system is perfectly at rest (*e.g* a robot at rest holding the system). We propose an inertial detector and a visual detector to decide if the system is on move, on SoftStop or HardStop and describe how to take advantage of this additional information in a VISLAM. A significant accuracy gain and better robustness against divergence is demonstrated on our datasets.**

I. Introduction

Monocular Visual-Inertial SLAM (VISLAM) algorithms are the most widespread and efficient solutions to solve indoor localization problem. Both cameras and Inertial Measurement Units (IMU) are quite cheap and are easily found together in smartphones for example. These sensors are very complementary. On one hand IMU brings robustness to visual-only localization giving a motion prediction at each frame and tackling the issue of camera denial of service (*e.g* motion blur, light saturation, pitch black). On the other hand camera reduce the divergence of cheap IMU-only localization systems and allows to observe gyroscope and accelerometer bias. There are a lot of implementation of VISLAM, which can be split in two main families, the filter-based [5] [11] and the graph optimization-based [6] [4] algorithms (called bundle adjustment).

However, this sensors combination, despite being efficient, has its own flaws. For monocular case, when the sensor rig is at rest, the visual constraints are poor since new 3D points cannot be accurately reconstructed. They do not prevent from velocity divergence induced by integration of noise preponderant inertial data as illustrated on Figure 1.

Velocity drift is well known in inertial-only localization literature. Two major procedures used to avoid velocity divergence are Pseudo Velocity Update and Zero Velocity Update (ZUPT) [10]. The former can be directly used in VISLAM since the only pre-requisite is to know dynamic limit of the system. The latter needs to detect when the system is at rest. Since a human can detect he is at rest through his eyes and internal ear, a monocular visual-inertial sensor rig should be able to do the same.

978-1-7281-1788-1/19/$31.00

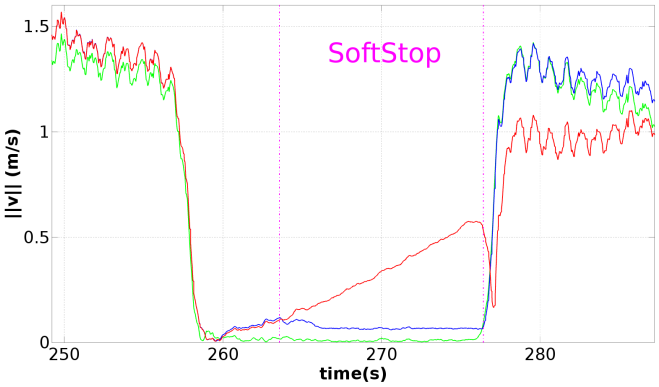

Fig. 1. Speed estimation of monocular VISLAM algorithms with the visual-inertial system described in (Section V). The velocity norm of the ground truth is represented in green, the one estimated by the MSCKF algorithm [5] in red and the velocity norm estimated by the proposed solution *i.e* MSCKF + Stationary Detector in blue. When velocity starts to diverge, SoftStop Measurement prevents velocity from growing, allowing for a better velocity estimation when the system moves again.

Thus we decided to find a visual-inertial criterion to determine if the system is at rest and to distinguish between SoftStop, *i.e* the system is at rest with residual motion noise, and HardStop, *i.e* the system is perfectly at rest. This additional information is inserted in a Multi State Constraint Kalman Filter (MSCKF) [5] framework in order to reduce velocity drift, thus increasing accuracy and robustness to divergence. We choose MSCKF since the Kalman Filter framework is well suited to add any kind of measurements, and it is one of the fastest VISLAM algorithm of the state of the art. The Stationary Detector can be also used in a graph based VISLAM algorithm [6] [4], in the keyframe selection strategy for example. The contributions of this paper are:

1) The distinction between SoftStop and HardStop with an inertial criterion [9]
2) A visual criterion to determine a SoftStop and an HardStop
3) An HardStop measurement
4) The addition of stationary information in the MSCKF framework

The outline of this paper is as follow. In Section II, the basis of the MSCKF are briefly reminded. Section III describes the design of the Stationary Detector. Then, Section IV focuses on properly using stationary measurement in MSCKF framework. Finally, Section V presents experimental results on our visual-inertial dataset.

II. MSCKF Framework

This section describes briefly the MSCKF Framework [5] and demonstrate its weakness during stationary period. We use the following abbreviation for 3D transformations:

$$P_{AfromB} = P_{AB} = \begin{bmatrix} R_{AB} & t_{AB} \\ 0_{1\times 3} & 1 \end{bmatrix} \in SE(3),$$

where $R_{AB} \in SO(3)$ is the rotation from frame B to frame A, $t_{AB} \in \mathbb{R}^3$ is the 3D coordinate of frame B in A and q_{AB} is the quaternion which represents the same rotation as R_{AB}. Let G be the gravitational frame, where the gravity is alongside the z axis, B be the body frame of our sensor rig, I be the inertial frame and C be the camera frame. For simplicity, we consider that B and I are the same frame and only use I in the following.

A. Filter State

There is two parts in the MSCKF state, the first part is the Current body state x_C:

$$\mathrm{x}_C = \begin{bmatrix} q_{GI}^\top & t_{GI}^\top & v_{GI}^\top & b_g^\top & b_a^\top \end{bmatrix}^\top,$$

where $v_{GI} \in \mathbb{R}^3$ is the velocity of the sensor rig expressed in G, vectors $b_g \in \mathbb{R}^3$ and $b_a \in \mathbb{R}^3$ are the biases of the measures of angular velocity and linear acceleration from the IMU.

The second part of the MSCKF state is the Window x_W of the N last body states associated to the N last images recorded by the camera:

$$\mathrm{x}^i = \begin{bmatrix} q_{GI}^{i\top} & t_{GI}^{i\top} \end{bmatrix}^\top$$
$$\mathrm{x}_W = \begin{bmatrix} \mathrm{x}^{1\top} & \cdots & \mathrm{x}^{N\top} \end{bmatrix}^\top.$$

So the full MSCKF state is:

$$\mathrm{x} = \begin{bmatrix} \mathrm{x}_C^\top & \mathrm{x}_W^\top \end{bmatrix}^\top.$$

B. Propagation Model

The IMU includes a 3 axis gyroscope and a 3 axis accelerometer. The gyrometer allows to predict q_{GI} through angular velocity integration and accelerometer is used to predict v_{GI} through acceleration integration. The MSCKF uses the following sensor model for IMU propagation:

$$\begin{aligned} \tilde{\omega}_I &= \omega_I + b_g + \eta_g \\ \tilde{\mathrm{a}}_I &= \mathrm{a}_I - R_{GI}^\top \mathbf{g}_G + b_a + \eta_a, \end{aligned} \tag{1}$$

where $\tilde{\omega}_I \in \mathbb{R}^3$ and $\tilde{\mathrm{a}}_I \in \mathbb{R}^3$ are the gyroscope and accelerometer raw measurements and $\omega_I \in \mathbb{R}^3$ and $\mathrm{a}_I \in \mathbb{R}^3$ are the angular velocity and acceleration experienced by the IMU. $\mathbf{g}_G$ is the local gravity vector with magnitude g. η_g and η_a are the gyroscope and accelerometer noise. From this equation, the predicted state $\mathrm{x}_{k+1|k}$ and covariance $P_{k+1|k}$ at time step $k+1$ can be estimated, see [11] for details.

C. Measurement Model

The MSCKF measurement consists to perform the classical Extended Kalman Filter (EKF) equations:

$$\begin{aligned} K &= P_{k+1|k} H^\top (H P_{k+1|k} H^\top + \Sigma_z)^{-1} \\ \delta\mathrm{x} &= K(z - h(\mathrm{x})) \\ P_{k+1|k+1} &= (I - KH) P_{k+1|k}, \end{aligned} \tag{2}$$

where h is the measurement prediction function, H is its Jacobian, z is the sensor output and Σ_z the covariance of z. The measurement model of the MSCKF is the comparison between observed 2D coordinate z and estimated 2D reprojection $h(\mathrm{x})$ of a 3D point. Each 3D point Q^j, $j \in [|1, ..., M|]$, expressed in G has N 2D observations $z^{i,j}$ on the body state window x_W and is estimated via triangulation. Once Q^j is known, it is possible to compute the reprojection function $h^j(\mathrm{x}_W)$:

$$\begin{aligned} h^j(\mathrm{x}^i) &= \Pi(P_{CI} P_{IG}^i(Q^j)) = \Pi(X^{i,j}, Y^{i,j}, Z^{i,j}) \\ \Pi(X^{i,j}, Y^{i,j}, Z^{i,j}) &= \begin{bmatrix} o_x \\ o_y \end{bmatrix} + \begin{bmatrix} f_x & 0 \\ 0 & f_y \end{bmatrix} \begin{bmatrix} \frac{X^{i,j}}{Z^{i,j}} \\ \frac{Y^{i,j}}{Z^{i,j}} \end{bmatrix} \\ h^j(\mathrm{x}_W) &= \begin{bmatrix} h^j(\mathrm{x}^1)^\top & \cdots & h^j(\mathrm{x}^N)^\top \end{bmatrix}^\top, \end{aligned} \tag{3}$$

where $X^{i,j}$, $Y^{i,j}$ and $Z^{i,j}$ are 3D coordinates of Q^j in frame C at timestamp i, Π is the 2D reprojection function, o_x, o_y, f_x and f_y are the optical center and the focal of the camera obtained with intrinsics calibration.

Linearizing the measurement model at the current estimate, the residual r^j of the measurement can be approximated as:

$$r^j = z^j - h^j(\mathrm{x}_W) = H_W^j \mathrm{x}_W + H_Q^j Q^j + \eta^j, \tag{4}$$

where H_Q^j the Jacobian of h^j with regard to Q^j (H_Q^j is discarded via left nullspace computation). Its formulation can be found in [11]. The Jacobian H_W^j of h^j with regard to x_W is given by:

$$\begin{aligned} H_W^j &= \begin{bmatrix} H_W^{1,j\top} & \cdots & H_W^{N,j\top} \end{bmatrix}^\top \\ H_W^{i,j} &= H_\Pi^{i,j} R_{CI} R_{IG}^i \begin{bmatrix} \lfloor Q^j - t_{GI}^i \rfloor_\times & -I_3 \end{bmatrix} \\ H_\Pi^{i,j} &= \frac{1}{Z^{i,j}} \begin{bmatrix} f_x & 0 & -\frac{X^{i,j} f_x}{Z^{i,j}} \\ 0 & f_y & -\frac{Y^{i,j} f_y}{Z^{i,j}} \end{bmatrix}, \end{aligned} \tag{5}$$

where $H_\Pi^{i,j}$ is the Jacobian of Π and $\lfloor \cdot \rfloor_\times$ the matrix cross product operator. All the equations 4 are stacked for each 3D point observed to obtain $r = z - h(\mathrm{x})$ and all H_W^j are stacked to obtain H, allowing to perform the EKF equations (2), with $\Sigma_z = \sigma_{cam}^2 I$.

D. Velocity Divergence at rest

In monocular case, when the sensor rig is at rest, the lack of temporal baseline leads to a very poor estimation of Q^j depth,

e.g all $Z^{i,j}$ are close to the infinite. The Jacobian equations 5 of the MSCKF for the position component $T^{i,j}$ becomes:

$$T^{i,j} = -H_{\Pi}^{i,j} R_{CI} R_{IG}^{i}$$
$$T^{i,j\infty} = \lim_{Z^{i,j}\to\infty} T^{i,j} = 0.$$

For the rotation component $R^{i,j}$, using the following equality $R\lfloor u \rfloor_{\times} = \lfloor Ru \rfloor_{\times} R$ valid for every rotation $R \in SO(3)$ and $u \in \mathbb{R}^3$:

$$R^{i,j} = H_{\Pi}^{i,j} R_{CI} R_{IG}^{i} \left\lfloor Q^{j} \right\rfloor_{\times} = H_{\Pi}^{i,j} \left\lfloor R_{CI} R_{IG}^{i} Q^{j} \right\rfloor_{\times} R_{CI} R_{IG}^{i}$$
$$= H_{\Pi}^{i,j} \begin{bmatrix} X^{i,j} \\ Y^{i,j} \\ Z^{i,j} \end{bmatrix}_{\times} R_{CI} R_{IG}^{i}.$$

Computing explicitly the above equation conducts to the following results:

$$R^{i,j\infty} = \lim_{Z^{i,j}\to\infty} R^{i,j} = \begin{bmatrix} 0 & -f_x & 0 \\ f_y & 0 & 0 \end{bmatrix} R_{CI} R_{IG}^{i}.$$

Finally:

$$H_{W}^{i,j\infty} = \lim_{Z^{i,j}\to\infty} H_{W}^{i,j} = \begin{bmatrix} R^{i,j\infty} & 0_{3\times 3} \end{bmatrix}. \tag{6}$$

The expression of $H_{W}^{i,j\infty}$ demonstrates that vision constraints only rotation when 3D points are triangulated at the infinite. That's why the MSCKF algorithm drift in position and speed at rest.

III. Stationary Detector

To prevent MSCKF algorithm from drifting at rest, the Stationary Detector will label every images with three values: Move, SoftStop *i.e* the system is at rest with residual motion noise and HardStop, *i.e* the system is perfectly at rest.

A. Prerequisite

The Stationary Detector aims to decide whether, during a time interval of N observations between the instants n and $n+N-1$, one sensor is in Move, SoftStop or HardStop.

In [9], the stationary detection problem is formalized as a binary hypothesis testing problem, where the detector can choose between the two hypotheses $\mathcal{H}^0$ and $\mathcal{H}^1$:

$$\mathcal{H}^0 : \text{Sensor is moving}$$
$$\mathcal{H}^1 : \text{Sensor is stationary.}$$

We use this formulation, the $\mathcal{H}^1$ hypothesis will include both SoftStop and HardStop states. Their distinction will be made in a second step. Both the false-alarm probability, $P_{FA} = Pr\{\mathcal{H}^1|\mathcal{H}^0\}$ (i.e., the probability of choosing $\mathcal{H}^1$ when $\mathcal{H}^0$ is true) and the probability of detection $P_D = Pr\{\mathcal{H}^1|\mathcal{H}^1\}$ (i.e., the probability of choosing $\mathcal{H}^1$ when $\mathcal{H}^1$ is true) determine the performance of detection. The Neyman-Pearson theorem allows to maximize P_D for a given P_{FA}. Let $\mathcal{Z}_n$ be the set of sensor measurements at all instants $k \in \Omega_n = [|n, ..., n+N-1|]$. Let $p(\mathcal{Z}_n; \mathcal{H}^0)$ and $p(\mathcal{Z}_n; \mathcal{H}^1)$ be the Probability Density Functions (PDF) of the observations for the two hypotheses:

Theorem 1: To maximize P_D for $P_{FA} = \alpha$, choose $\mathcal{H}^1$ if

$$L(\mathcal{Z}_n) > \gamma \text{ with } L(\mathcal{Z}_n) = \frac{p(\mathcal{Z}_n; \mathcal{H}^1)}{p(\mathcal{Z}_n; \mathcal{H}^0)}, \tag{7}$$

where the threshold γ is determined from:

$$P_{FA} = \int_{\{\mathcal{Z}_n : L(\mathcal{Z}_n) > \gamma\}} p(\mathcal{Z}_n; \mathcal{H}^0) d\mathcal{Z}_n = \alpha.$$

The test (7) is called the Likelihood Ratio Test (LRT). The function $L(\mathcal{Z}_n)$ in (7) is known as the likelihood ratio since it indicates the likelihood of the $\mathcal{H}^1$ hypothesis versus the $\mathcal{H}^0$ hypothesis. To perform the LRT, the PDFs of the observations $\mathcal{Z}_n$ for $\mathcal{H}^0$ and $\mathcal{H}^1$ must be known, however, it's impossible to predict all observations in $\mathcal{Z}_n$ with only $\mathcal{H}^0$ or $\mathcal{H}^1$ information, so they have to be approximated by hypotheses test method.

That's why θ^i, the set of unknown parameters for each hypothesis $\mathcal{H}^i$ with $i \in \{0, 1\}$ must be introduced. The LRT is performed by substituting θ^i with their Maximum Likelihood Estimates (MLE) $\hat{\theta}^i$ assuming $\mathcal{H}^i$ is true. The test is called Generalized Likelihood Ratio Test (GLRT) and consists in selecting $\mathcal{H}^1$ if:

$$L(\mathcal{Z}_n) > \gamma \text{ with } L(\mathcal{Z}_n) = \frac{p(\mathcal{Z}_n; \hat{\theta}^1, \mathcal{H}^1)}{p(\mathcal{Z}_n; \hat{\theta}^0, \mathcal{H}^0)}. \tag{8}$$

The proposed Stationary Detector is based on two main criteria, an inertial criterion and a visual criterion. Let's define $\mathcal{Z}_n$, θ^i, $\hat{\theta}^i$ and $p(\mathcal{Z}_n; \hat{\theta}^i, \mathcal{H}^i)$ for each sensor to compute the GLRT and select between $\mathcal{H}^0$ and $\mathcal{H}^1$. Once $\mathcal{H}^1$ is detected, the sensor is either in SoftStop or in HardStop. A second test must be performed after the GLRT to distinguish between those two states.

B. IMU Stationary Criterion

The same modelization as [9] will be used for the inertial GLRT. The IMU sensor model is simpler as the one of equation 1 since only raw data are used in the GLRT. The presence of b_g and b_a, which depend of the localization algorithm, may corrupt the GLRT during localization divergence. So let $y^k \in \mathbb{R}^6$ and $s^k \in \mathbb{R}^6$ be the vector

$$y^k = \begin{bmatrix} y_a^k \\ y_\omega^k \end{bmatrix} = \begin{bmatrix} \tilde{\mathrm{a}}_I^k \\ \tilde{\omega}_I^k \end{bmatrix} \text{ and } s^k = \begin{bmatrix} s_a^k \\ s_\omega^k \end{bmatrix} = \begin{bmatrix} \mathrm{a}_I^k - \mathbf{g}_I \\ \omega_I^k \end{bmatrix} \tag{9}$$

denoting respectively the output of the IMU and the IMU-experienced specific force and angular rate at time instant $k \in \Omega_n$. Thus, the set of IMU measurements $\mathcal{Z}_n$ is equal to $\{y^k\}_{k\in\Omega_n}$ The following simple model is applied to y^k:

$$y^k = s^k + \eta^k,$$

where

$$\eta^k = \begin{bmatrix} \eta_{\mathrm{a}} \\ \eta_\omega \end{bmatrix}.$$

Here, $\mathbf{g}_I$ is the local gravity vector in IMU frame, and η^k corresponds to the accelerometer and gyroscope noise. The noise is assumed to be independent identically distributed

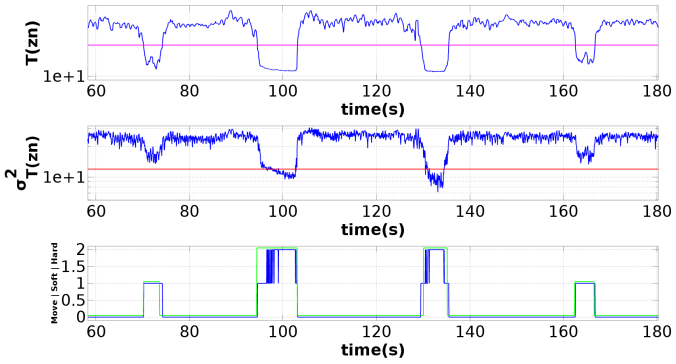

Fig. 2. Inertial Stationary Criterion on a dataset obtained with a DUO MLX (Section V). A SoftStop, two HardStops and a SoftStop are performed consecutively and labeled by hand. Top: $T(\mathcal{Z}_n)$ allows to distinguish between Move and any stop but it is impossible to distinguish between SoftStop and HardStop. Middle: $\sigma_{T(\mathcal{Z}_n)}$, it is possible to distinguish between SoftStop and HardStop. Bottom: IMU stationary decision (blue) and the ground truth stationary decision (green).

zero-mean, Gaussian with covariance matrix:

$$\Sigma_{imu} = \mathbb{E}\{\eta^k \eta^{k\top}\} = \begin{bmatrix} \sigma_a^2 I_3 & 0_3 \\ 0_3 & \sigma_g^2 I_3 \end{bmatrix},$$

with σ_a and σ_g denoting the accelerometer and gyroscope noise densities. Those values can be computed with Allan variance [1].

The inertial GLRT [9] is given by :

$$L(\mathcal{Z}_n) = \exp\left(- \sum_{k \in \Omega_n} \frac{1}{2\sigma_a^2} ||y_a^k - g.\frac{\overline{y}_a^n}{||\overline{y}_a^n||}||^2 + \frac{1}{2\sigma_g^2}||y_\omega^k||^2 \right), \quad (10)$$

where

$$\overline{y}_a^n = \frac{1}{N} \sum_{k \in \Omega_n} y_a^k.$$

There is no need to compute exponential function since the logarithm is monotonically increasing, the GLRT can be simplified into:

$$T(\mathcal{Z}_n) < \gamma' \text{ with } T(\mathcal{Z}_n) = -\frac{2}{N} \ln L(\mathcal{Z}_n).$$

Thus:

$$T(\mathcal{Z}_n) = \frac{1}{N} \sum_{k \in \Omega_n} \left(\frac{1}{\sigma_a^2} ||y_a^k - g.\frac{\overline{y}_a^n}{||\overline{y}_a^n||}||^2 + \frac{1}{\sigma_g^2}||y_\omega^k||^2 \right). \quad (11)$$

This can be interpreted as follows, the GLRT chooses the hypothesis that the IMU is stationary if the weighted sum of a gyroscopic score and an accelerometric score falls below the threshold $\gamma' = -\frac{2}{N} ln(\gamma)$. The gyrometric score is low when the gyrometer output is close to zero, *e.g* when there is no angular velocity. The accelerometric score is low when each accelerometer data are identical vectors of magnitude g. We analyzed the behavior of $T(\mathcal{Z}_n)$ in order to find an empiric value for γ' and to maximize the stationary detection without false alarm.

However, this test cannot distinguish between SoftStop or HardStop. Figure 2 shows that $T(\mathcal{Z}_n)$ is relatively low during both SoftStop and HardStop, so it cannot be a relevant criterion. Nevertheless, $T(\mathcal{Z}_n)$ vibrates way more during SoftStop than HardStop, so computing the variance $\sigma^2_{T(\mathcal{Z}_n)}$ of $T(\mathcal{Z}_n)$, is a good way to find a distinctive criterion:

$$\sigma^2_{T(\mathcal{Z}_n)} = \frac{1}{N} \sum_{k=n-N}^{k=n} \left(||T(\mathcal{Z}_k) - \overline{T(\mathcal{Z}_n)}||^2 \right). \quad (12)$$

Being given both $T(\mathcal{Z}_n)$ and $\sigma^2_{T(\mathcal{Z}_n)}$, it is possible to accurately find if the system is moving, in SoftStop or in HardStop, as illustrated in Figure 2.

C. Visual Stationary Criterion

The visual stationary criterion is based on 2D points analysis. Each tracked 2D point contributes to the state decision (Move, SoftStop or HardStop). For each tracked 2D point, the following model is used, let $y^k \in \mathbb{R}^2$ be the vector of the 2D coordinates of one 2D point on image $k \in \Omega_n$ and $s^k \in \mathbb{R}^2$ be the 2D coordinates corresponding to the reprojection of the real 3D points on the camera image k:

$$y^k = \begin{bmatrix} y_{\text{x}}^k \\ y_{\text{y}}^k \end{bmatrix} \text{ and } s^k = \begin{bmatrix} s_{\text{x}}^k \\ s_{\text{y}}^k \end{bmatrix}.$$

The set of camera measurements $\mathcal{Z}_n$ is equal to $\{y^k\}_{k \in \Omega_n}$. Similarly to the IMU, the following signal model is applied:

$$y^k = s^k + \eta^k, \quad (13)$$

where

$$\eta^k = \begin{bmatrix} \eta_{\text{x}} \\ \eta_{\text{y}} \end{bmatrix}.$$

Here, η^k aggregates the image noise and the pixel detection noise. η^k is always depicted as an independent identically distributed zero-mean Gaussian with the same intensity on x and y axis:

$$\Sigma_{cam} = \mathbb{E}\{\eta^k \eta^{k\top}\} = \sigma_{cam}^2 I_2. \quad (14)$$

If the camera is stationary, then each tracked 2D points should be exactly the same with the exception of those associated to moving objects. Mathematically, the visual signal fulfills the following condition:

$$\mathcal{H}^0 : \exists k, k' \in \Omega_n \text{ so that } s^k \neq s^{k'}$$
$$\mathcal{H}^1 : \forall k, k' \in \Omega_n \ , \ s^k = s^{k'} = s^n.$$

So it follows that under the hypothesis $\mathcal{H}^0$, the signal is totally unknown and cannot be predicted, whereas under $\mathcal{H}^1$, only one 2D coordinate is unknown, the knowledge of this coordinate allows to predict the entire signal:

$$\mathcal{H}^0 : \theta = \{s^k\}_{k \in \Omega_n}$$
$$\mathcal{H}^1 : \theta = s^n.$$

Since all measurements y^k are considered independent and according to the camera sensor described in 13 and 14, the

expression of $p(\mathcal{Z}_n;\theta,\mathcal{H}^i)$ is given by:

$$p(\mathcal{Z}_n;\theta,\mathcal{H}^i) = \prod_{k\in\Omega_n} p(y^k;\theta^i,\mathcal{H}^i) \tag{15}$$

$$p(y^k;\theta^i,\mathcal{H}^i) = \frac{1}{2\pi\sigma_{cam}^2}\exp\left(-\frac{1}{2\sigma_{cam}^2}||y^k - s^k||^2\right).$$

The MLE $\hat{\theta}^i$ is calculated by maximizing (15) with respect to the unknown parameters θ^i. Under $\mathcal{H}^0$, the result is straightforward since all the signal is completely unknown and $\hat{\theta}^0 = \{s^k\}_{k\in\Omega_n}$, so:

$$p(\mathcal{Z}_n;\hat{\theta}^0,\mathcal{H}^0) = \frac{1}{2\pi\sigma_{cam}^2}. \tag{16}$$

Under $\mathcal{H}^1$, the MLE of $\hat{\theta}^1 = s^n$ is calculated by maximizing $p(\mathcal{Z}_n;s^n,\mathcal{H}^1)$ with respect to $s^n \in \mathbb{R}^2$:

$$p(\mathcal{Z}_n;s^n,\mathcal{H}^1) = \frac{1}{2\pi\sigma_{cam}^2}\exp\left(-\frac{1}{2\sigma_{cam}^2}||y^k - s^n||^2\right). \tag{17}$$

The maximization of $p(\mathcal{Z}_n;s^n,\mathcal{H}^1)$ leads to $\hat{s}^n$:

$$\begin{aligned}\hat{s}^n &= \arg\max_{s\in\mathbb{R}^2}(p(\mathcal{Z}_n;s,\mathcal{H}^1))\\ &= \arg\min_{s\in\mathbb{R}^2}(\sum_{k\in\Omega_n}||y^k - s||^2)\\ &= \arg\min_{s\in\mathbb{R}^2}(\sum_{k\in\Omega_n}(||y^k||^2 - 2s^\top y^k + ||s||^2))\\ &= \arg\min_{s\in\mathbb{R}^2}(||s||^2 - 2s^\top\overline{y}^n) = \overline{y}^n,\end{aligned} \tag{18}$$

where

$$\overline{y}^n = \frac{1}{N}\sum_{k\in\Omega_n} y^k.$$

Substituting $\hat{s}^n$ into (17) gives:

$$p(\mathcal{Z}_n;\hat{\theta}^1 = \hat{s}^n,\mathcal{H}^1) = \frac{1}{2\pi\sigma_{cam}^2}\exp\left(-\frac{1}{2\sigma_{cam}^2}||y^k - \overline{y}^n||^2\right). \tag{19}$$

Finally, the GLRT is computed with equations (8), (16) and (19):

$$L(\mathcal{Z}_n) = \exp\left(-\frac{1}{2\sigma_{cam}^2}\sum_{k\in\Omega_n}||y^k - \overline{y}^n||^2\right) \tag{20}$$

$$T(\mathcal{Z}_n) = \frac{1}{N}\sum_{k\in\Omega_n}\left(\frac{1}{\sigma_{cam}^2}(||y_{\mathrm{x}}^k - \overline{y}_{\mathrm{x}}^n||^2 + ||y_{\mathrm{y}}^k - \overline{y}_{\mathrm{y}}^n||^2)\right). \tag{21}$$

It is difficult to interpret this test unless the covariance matrix of the data $\mathcal{Z}_n = \{y^k\}_{k\in\Omega_n}$ is introduced:

$$\Sigma_{\mathcal{Z}_n} = \mathbb{E}[(y - \mathbb{E}[y])(y - \mathbb{E}[y])^\top] = \begin{bmatrix}\sigma_{\mathrm{x}}^2 & \sigma_{\mathrm{xy}}\\ \sigma_{\mathrm{xy}} & \sigma_{\mathrm{y}}^2\end{bmatrix}.$$

The surface area $S = det(\Sigma_{\mathcal{Z}_n}) = \sigma_{\mathrm{x}}^2\sigma_{\mathrm{y}}^2 - \sigma_{\mathrm{xy}}^2$ corresponds to the vibration of the 2D point around its mean. If the camera does not move, $S = 0$, and S grows bigger when the movement is stronger. S and $T(\mathcal{Z}_n)$ are correlated:

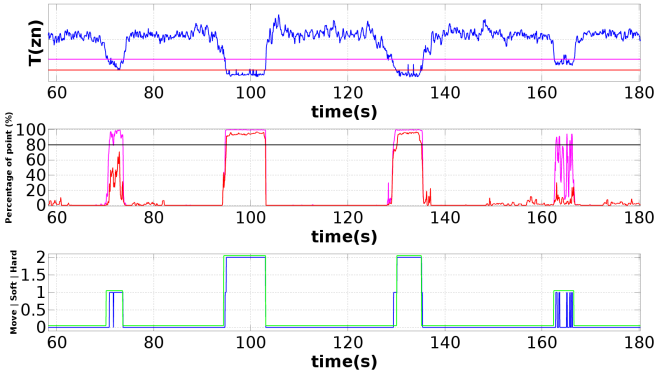

Fig. 3. Camera Stationary Criterion on a dataset obtained with DUO MLX (Section V). A SoftStop, two HardStops and a SoftStop are performed consecutively and labeled by hand. Top: $T(\mathcal{Z}_n)$ for one tracked 2D feature (Harris features matched with SURF descriptors), it is possible to distinguish between SoftStop and HardStop. Middle: percentage of SoftStop points (magenta) and HardStop points (red), this criterion is much more stable than for a single point. Bottom: camera stationary decision (blue) and the ground truth stationary decision (green).

$$(T(\mathcal{Z}_n))^2 = \left(\frac{\sigma_x^2 + \sigma_y^2}{\sigma_{cam}^2}\right)^2$$

$$\begin{aligned}(\sigma_x^2 + \sigma_y^2)^2 &> 4\sigma_x^2\sigma_y^2\\ &> 4(\sigma_x^2\sigma_y^2 - \sigma_{xy}^2)\\ &> 4S.\end{aligned}$$

Finally:

$$T(\mathcal{Z}_n) > 2\frac{\sqrt{S}}{\sigma_{cam}^2}. \tag{22}$$

According to inequality 22, if $T(\mathcal{Z}_n)$ is minimized, S decreases. Since S is notably correlated to the movement of the camera, it points out a good physical interpretation for $T(\mathcal{Z}_n)$.

We once again analyzed the behavior of $T(\mathcal{Z}_n)$ in order to find an empiric value for γ', exhibiting the maximum stationary detection without false alarm. Depicted on figure 3, a two threshold approach allows to distinguish SoftStop and HardStop:

$$\begin{cases}\text{if } T(\mathcal{Z}_n) > \gamma_{soft}\text{ , the camera moves}\\ \text{if } \gamma_{soft} > T(\mathcal{Z}_n) > \gamma_{hard}\text{ , the camera is in SoftStop}\\ \text{if } \gamma_{hard} > T(\mathcal{Z}_n)\text{ , the camera is in HardStop.}\end{cases}$$

After computing the decision of every tracked 2D points, a global decision is taken. If more than 80 percents of every tracked point conclude SoftStop or HardStop, then this decision is taken. This threshold should be carefully chosen to avoid false alarm. The visual criterion for stationary detection is taken into account if the number of 2D tracked points is sufficient (more than 50 points).

D. Visual-Inertial Stationary Test

Despite their efficiency, both inertial and visual stationary criteria have limitations. If the IMU operates on a constant velocity platform, the observation would be the same as being at rest and the inertial stationary test could be tricked that way.

If the camera moves towards a scene at the infinite, the observation is the same as being at rest and the visual stationary test could also be tricked. Moreover, the visual test fails if the observed scene is moving.

The Stationary Detector consists in combining both visual and inertial decisions, thus avoiding those drawbacks:

$$Soft_{system} = (Soft_{imu} \text{ AND } Soft_{cam}) \text{ OR } Hard_{imu}$$
$$Hard_{system} = Hard_{cam} \text{ AND } Hard_{imu},$$

with the following property (by construction of inertial and visual criteria):

$$Hard_{imu} \implies Soft_{imu}$$
$$Hard_{cam} \implies Soft_{cam}.$$

It is mandatory that the Stationary Detector is as restrictive as possible, a wrong SoftStop or HardStop measurement could drastically degrade localization. That's why the system needs both sensor to detect SoftStop or HardStop instead of just one of them. That way, if one sensor is tricked with one of the two above drawbacks, the system will stay in Move and not insert wrong stationary measurement. Both sensor can still be tricked at the same time, but it is much more harder than just one of them. It is not an issue if SoftStop or HardStop are not detected during the entire time the system is at rest, since sporadic measurement is sufficient to correct diverging localization.

The inertial HardStop criterion is very sensitive to any kind of move and difficult to obtain, so if it occurs when the vision says move, we still consider that it is a SoftStop. This way, if the visual criterion is corrupted by a moving scene, the system will not diverge thanks to SoftStop measurement.

The visual ineterial stationary test has a negligeable impact on the MSCKF executation time. In fact the most computational process, *i.e* the detection and the matching of 2D points, is already perfomed by the MSCKF and is thus reused by our Stationary Detector.

IV. Stationary Measurement

This section describes the measures to include in the MSCKF framework when the Stationary Detector finds a SoftStop or an HardStop. For both stationary state, h, z and Σ_z are described. Once those values are known, the classical EKF measurement equation (2) can be performed.

A. SoftStop Measurement

In this case, the measurement consists in applying a Zero Velocity Update (ZUPT) [10]:

$$z = 0_{3\times1} \text{ , } h(\mathrm{x}) = v_{GI} \text{ and } \Sigma_z = \sigma_{soft} I_3. \tag{23}$$

However, the ZUPT measurement is fundamentally wrong in SoftStop state since $v_{GI} \neq 0$ due to residual movement. It is not a problem for an inertial only navigation system since the estimated velocity is often way more wrong than fixing $v_{GI} = 0$. However, monocular visual-inertial system may estimate a very low velocity at rest by themselves, depending on IMU quality and the duration of stationary periods. That's why a χ^2 95% activation test is required to check if a ZUPT is needed:

$$v_{GI}^\top \Sigma_z^{-1} v_{GI} < \chi^2(3, 95). \tag{24}$$

If this test is true, it means the system is already observing the good velocity so it is useless to add wrong information with the ZUPT. Else, the velocity of the system is diverging and requires to be fixed.

B. HardStop Measurement

When an image is labeled HardStop, it is possible to add more information in the localization than with a single ZUPT. Since the system is supposed to be totally at rest, the inertial equations 1 simplify as follows:

$$\tilde{\omega}_I = b_g + \eta_g$$
$$\tilde{\mathrm{a}}_I = -R_{GI}^\top \mathbf{g}_G + b_a + \eta_a.$$

Moreover, if the HardStop lasts during a sufficient time, it is possible to compute the mean of those equations, eliminating the sensor noise:

$$\overline{\tilde{\omega}_I} \approx b_g$$
$$\overline{\tilde{\mathrm{a}}_I} \approx -R_{GI}^\top \mathbf{g}_G + b_a. \tag{25}$$

From this observation, the HardStop output z and the measurement prediction function h are deduced:

$$z = \begin{bmatrix} 0_{3\times1} \\ \overline{\tilde{\omega}_I} \\ \overline{\tilde{\mathrm{a}}_I} \end{bmatrix} \text{ and } h(\mathrm{x}) = \begin{bmatrix} v_{GI} \\ b_g \\ -R_{GI}^\top \mathbf{g}_G + b_a \end{bmatrix}. \tag{26}$$

As for Σ_z, the velocity is perfectly null so an arbitrary low value such as $\sigma_v = 0.0003 m/s$ will fit well. For the gyroscope bias and accelerometer bias, gyroscope and accelerometer random walks σ_{wg} and σ_{wa} are a good estimation of the covariance. Those values can be computed with Allan variance [1]. Finally:

$$\Sigma_z = \begin{bmatrix} \sigma_v^2 I_3 & 0_3 & 0_3 \\ 0_3 & \sigma_{wg}^2 I_3 & 0_3 \\ 0_3 & 0_3 & \sigma_{wa}^2 I_3 \end{bmatrix}.$$

To conclude, the HardStop measurement not only resets velocity to 0, it also allows to reestimate b_g, and to correct $R_{GI}^\top$ and b_a.

V. Experiments

It would have been interesting to validate the proposed Stationary Detector on well known visual-inertial datasets such as EuRoC Mav [3] or TUM [8]. However, they do not exhibit any stationary period during sequences. The system is at rest only at the beginning and at the end of each acquisition. That's

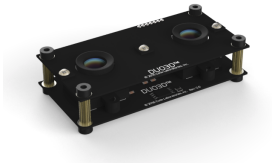

Fig. 4. DUO MLX, a stereo camera/IMU with fish-eye lens and a 3 cm baseline.

Fig. 5. Our 4 cameras/IMU helmet. Copyright ©CEA, NEXTER, ANR, DGA

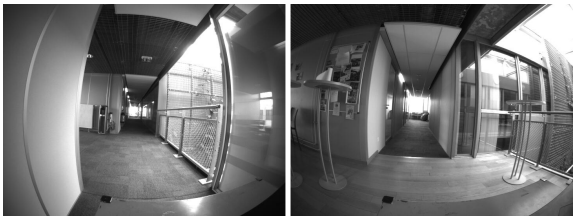

Fig. 6. Illustration of the robot sequence.

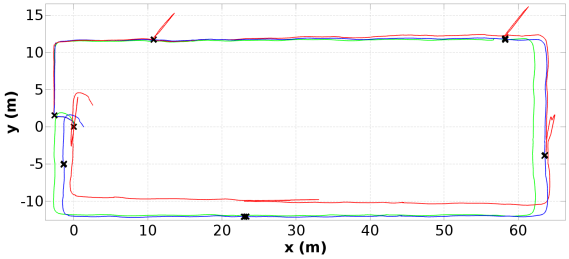

Fig. 7. The robot sequence. In green , 4-cam/imu MSCKF localization that constitutes our ground truth. In blue, 1-cam/imu MSCKF localization with stationary measurement. In red standard 1-cam/imu MSCKF [11] localization. The black cross indicates the places where HardStops occur. Adding HardStop stationary measurements in the MSCKF framework results in a more accurate localization due to the absence of divergence at rest.

why we designed our own dataset with two visual-inertial sensor systems. We present here two acquisitions of relevant use case scenario, a robot trajectory and a pedestrian trajectory.

A. Sensor sets

Our main system is a head-mounted 4-cameras system, with a front and a back 30 cm baseline stereo. Each camera is a FLIR Blackfly S with a global shutter sensor (Figure 5). The IMU of the helmet is the SBG Ellipse-N. The localization obtained by a 4-cam/imu MSCKF is accurate enough to be considered as a ground truth for every trajectories and to observe small velocities at rest (due to stereo baseline). To evaluate our proposed solution, only the IMU and one frontal camera of the system is used. We compare the standard MSCKF [11] with and without the proposed Stationary Detector.

We also use the DUO MLX, which is a stereo camera with fish-eye lens and a 3 cm baseline (Figure 4). Its IMU is an InvenSense MPU-6050. This device (the IMU and the left camera) is only used to generate test datasets for inertial and visual stationary criteria analysis (Figures 2 and 3), and their thresholds γ^{imu}_{soft}, γ^{imu}_{hard}, γ^{cam}_{soft}, γ^{cam}_{hard} estimation. Those values are then used by the helmet system to demonstrate the genericity of the Stationary Detector.

Both sensor sets were calibrated with the open source camera/IMU calibration framework Kalibr [7].

B. Robot trajectory

For this use case scenario, we put the helmet on a robot. It performs a 183 m long trajectory inside a corridor environment (Figure 6) with several stops of 30 seconds as illustrated on Figure 7. At each stop, the monocular MSCKF estimates an increasing velocity between 0.4 m/s and 0.7 m/s as illustrated on Figure 8. These speed divergence at rest induce errors in position up to 5 m. When the system moves again, the MSCKF partially succeed to reestimate an accurate speed and thus to correct the position drift. However, adding the stationary test to the MSCKF allows to estimate correctly the velocity at rest resulting in a much better velocity estimation when the system moves again. This results in a more accurate localization. In fact, errors in position at the end of the sequence are 1.1m and 2.37m for the MSCKF with and without stationary detector respectively. Hardstop are detected whenever the robot is at rest on this sequence. This is only possible for robots whose engines do not cause strong vibrations, otherwise SoftStop may be detected rather than HardStop.

C. Pedestrian trajectory

For this use case scenario, the helmet is worn by a human who walks during a 326 m long trajectory inside a hall environment (Figure 9). In this configuration, it is impossible to detect HardStop since there is constantly human noise movement on the head, only SoftStop are detectable. Figure 10 shows the trajectory performed and where SoftStops occur. SoftStop measurement allows the MSCKF to limit velocity divergence as illustrated on Figure 1 (for one stop of the sequence corresponding to the lower right part of the Figure 10). However, contrary to a MSCKF with HardStop measurement, the velocity estimated at rest with SoftStop is not exactly zero since the confidence in its measure $\sigma_{soft} = 0.045m/s$ is less strong than the HardStop one $\sigma_{hard} = 0.0003m/s$, resulting in

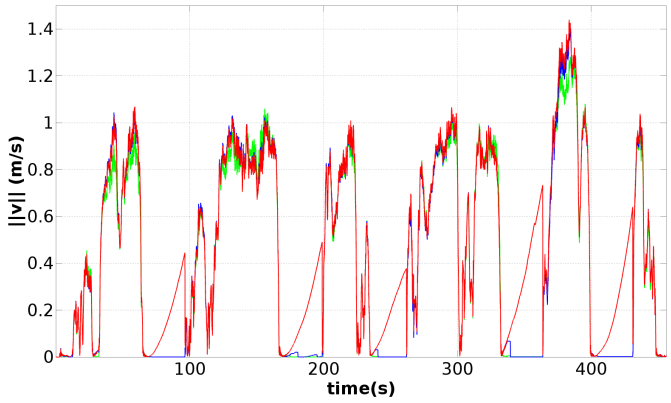

Fig. 8. The robot sequence. In green , velocity norm estimated by the 4-cam/imu MSCKF that constitutes our ground truth. In blue, velocity norm estimated by the 1-cam/imu MSCKF with stationary measurement. In red velocity norm estimated by the standard 1-cam/imu MSCKF [11] velocity. HardStops measurement allows the monocular MSCKF to avoid velocity divergence and to accurately estimate the speed during the whole sequence.

Fig. 9. Illustration of the pedestrian sequence.

a weaker speed constraint. Thus the position at rest cannot be completely removed during a SoftStop even if it is drastically reduced compared to no stationary detection. When the system moves again, the velocity is more accurately estimated with the proposed solution yielding a more accurate localization. Indeed, the standard MSCKF has a 13.1 m position error and the MSCKF with Stationary Detector a 4.8 m position error at end of the sequence.

VI. Conclusion

In this paper, we present a Stationary Detector which can be integrated in any monocular VISLAM algorithms. It takes advantage of the rawest inertial and visual data, so it does not depend on localization status and keep working when divergence occurs. We present how to insert stationary measurement in the MSCKF framework. A gain of precision and robustness to velocity divergence is demonstrated on our own dataset. Further work will improve the visual stationary criterion with semantic image segmentation, to focus only on 2D points of the static parts of the scene. An other perspective is to adaptatively estimate σ_{soft} by using a deep leraning based approach inspired of [2].

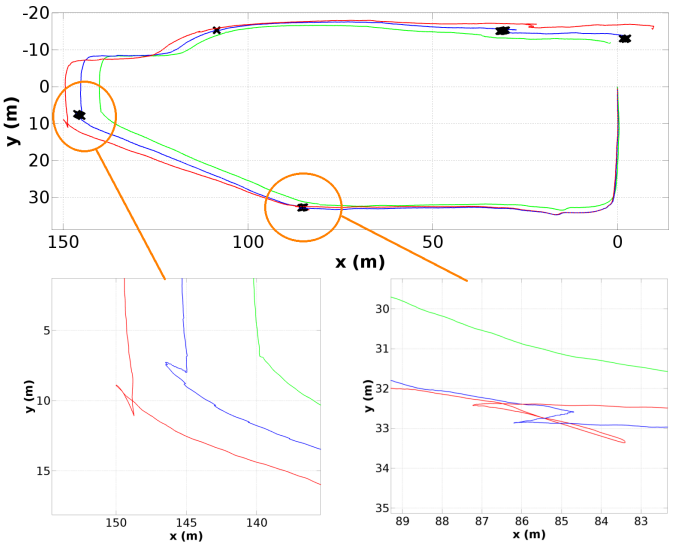

Fig. 10. The pedestrian sequence. In green , 4-cam/imu MSCKF localization that constitutes our ground truth. In blue, 1-cam/imu MSCKF localization with stationary measurement. In red standard 1-cam/imu MSCKF [11] localization. The black cross indicates the places where SoftStops occur. Adding SoftStop stationary measurements in the MSCKF framework results in a more accurate localization due to restrained divergence at rest. Figure 1 plots velocity norm of the trajectory zoom on the right.

VII. Acknowledgments

This work was supported by the Agence Nationale de la Recherche (ANR) and the Direction Générale de l'Armement (DGA).

References

[1] D. W. Allan, "Statistics of atomic frequency standards," vol. 54, no. 2, pp. 221–230, 1966.
[2] M. Brossard, A. Barrau, and S. Bonnabel, "AI-IMU dead-reckoning," *CoRR*, 2019.
[3] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. Achtelik, and R. Siegwart, "The EuRoC micro aerial vehicle datasets," *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.
[4] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "Imu preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation," *RSS*, 2015.
[5] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint Kalman filter for vision-aided inertial navigation," *ICRA*, 2007.
[6] R. Mur-Artal and J. D. Tardós, "Visual-Inertial Monocular SLAM with Map Reuse," *CoRR*, 2016.
[7] J. Rehder, J. Nikolic, T. Schneider, T. Hinzmann, and R. Siegwart, "Extending kalibr: Calibrating the extrinsics of multiple imus and of individual axes," *ICRA*, 2016.
[8] D. Schubert, T. Goll, N. Demmel, V. Usenko, J. Stueckler, and D. Cremers, "The TUM VI Benchmark for Evaluating Visual-Inertial Odometry," *IROS*, 2018.
[9] I. Skog, P. Handel, J. Nilsson, and J. Rantakokko, "Zero-Velocity Detection—An Algorithm Evaluation," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 11, pp. 2657–2666, 2010.
[10] A. Solin, S. C. Reina, E. Rahtu, and J. Kannala, "Inertial Odometry on Handheld Smartphones," *CoRR*, 2017.
[11] K. Sun, K. Mohta, B. Pfrommer, M. Watterson, S. Liu, Y. Mulgaonkar, C. J. Taylor, and V. Kumar, "Robust stereo visual inertial odometry for fast autonomous flight," *CoRR*, 2017.