2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

IDOL: A Framework for IMU-DVS Odometry using Lines

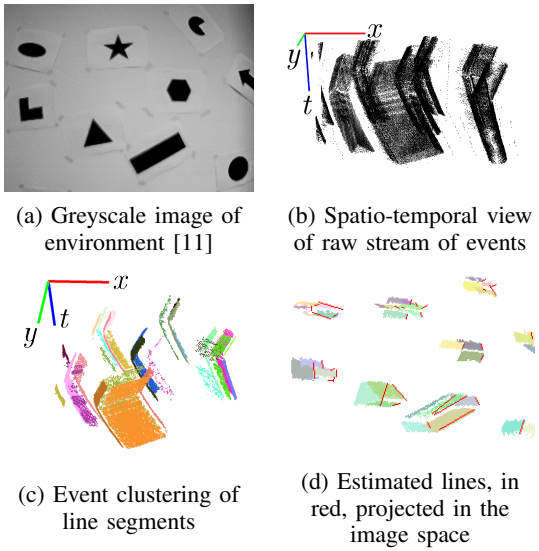
Cedric Le Gentil[1,2,*], Florian Tschopp[1,*], Ignacio Alzugaray[3],
Teresa Vidal-Calleja[2], Roland Siegwart[1], and Juan Nieto[1]

***Abstract*— In this paper, we introduce IDOL, an optimization-based framework for IMU-DVS Odometry using Lines. Event cameras, also called Dynamic Vision Sensors (DVSs), generate highly asynchronous streams of events triggered upon illumination changes for each individual pixel. This novel paradigm presents advantages in low illumination conditions and high-speed motions. Nonetheless, this unconventional sensing modality brings new challenges to perform scene reconstruction or motion estimation. The proposed method offers to leverage a continuous-time representation of the inertial readings to associate each event with timely accurate inertial data. The method's front-end extracts event clusters that belong to line segments in the environment whereas the back-end estimates the system's trajectory alongside the lines' 3D position by minimizing point-to-line distances between individual events and the lines' projection in the image space. A novel attraction/repulsion mechanism is presented to accurately estimate the lines' extremities, avoiding their explicit detection in the event data. The proposed method is benchmarked against a state-of-the-art frame-based visual-inertial odometry framework using public datasets. The results show that IDOL performs at the same order of magnitude on most datasets and even shows better orientation estimates. These findings can have a great impact on new algorithms for DVS.**

I. Introduction

In mobile robotics, having an understanding of the environment and the system's position in it is essential [1]. While being cost-effective, visual Simultaneous Localization And Mapping (SLAM) and Visual Odometry (VO) have been shown to achieve high accuracy and robustness in many applications [2]–[4]. By adding an Inertial Measurement Unit (IMU), the accuracy and robustness of VO can be further improved [5]–[8]. However, there are still scenarios which are challenging for visual-inertial systems such as under very fast motions or in scenes with High Dynamic Range (HDR) of illumination.

Dynamic Vision Sensors (DVSs), also called event-based cameras, are new sensor types that have a huge potential in addressing aforementioned limitations due to their extremely high temporal resolution and their HDR operative range [9], [10]. In contrast to traditional frame-based cameras, which periodically output intensity values for all pixels, these sensors output event tuples $e = \{t, x, y, p\}$, where t is the timestamp of the event, x, y are the image coordinates, and p is the event polarity. Such an event is triggered only in case a pixel's intensity change is larger than a threshold, with p being the direction of that change. Consequently, in a static scene, events are only generated when the camera is in motion. The events form an asynchronous stream of data that provides reliable information even in the presence of fast motion. As of today, most developments in the event-based literature focused on techniques to aggregate the event data into key-frames in order to apply or adapt conventional frame-based VO algorithms. To fully leverage the potential of this novel type of modality, new algorithms and ego-motion estimation paradigms need to be developed.



(a) Greyscale image of environment [11]

(b) Spatio-temporal view of raw stream of events

(c) Event clustering of line segments

(d) Estimated lines, in red, projected in the image space

Fig. 1: The proposed method, IDOL, estimates the system's ego-motion based on the segmentation and position estimation of line segments in the environment.

Most of the state-of-the-art work in both traditional and event-based VO rely on the tracking of point-like features over time. Those feature-tracks are used in a filter or optimization-based back-end to estimate the camera motion and the 3D location of the observed points. Human-made structures, however, are built with regular geometric shapes such as lines, making point landmarks not necessarily the best representation for visual tracking in all the scenarios.

In this paper, we investigate the potential of directly using asynchronous events without any frame-like accumulation, for ego-motion estimation. A key element is the use of a

* Equal contribution
[1]Authors are members of the Autonomous Systems Lab, ETH Zurich, Switzerland; {firstname.lastname}@mavt.ethz.ch
[2]Authors are members of the Centre for Autonomous Systems, School of Mechanical and Mechatronic Engineering, University of Technology Sydney, Sydney, New South Wales, Australia cedric.legentil@student.uts.edu.au, teresa.vidalcalleja@uts.edu.au
[3]The author is member of the Vision for Robotics Lab, ETH Zurich, Switzerland; ialzugaray@mavt.ethz.ch
This work was partly supported by Siemens Mobility, Germany and the ETH Mobility Initiative under the project *PROMPT*.

continuous-time representation of the inertial data to constrain the system pose at any time without relying on any motion model. By tightly coupling IMU and event data together via the generation of inertial information at each event's timestamp, a visual-inertial odometry (VIO) formulation that addresses and leverages the data asynchronism rigorously is achieved. Furthermore, instead of using traditional point features, we represent the environment using line segments. We introduce a new pipeline, IMU-DVS Odometry with Lines (IDOL), that detects line features in the event data and performs VIO by individually considering asynchronous events in a batch-optimization framework.

The remainder of the paper is organized as follows: In Section II, a summary of related work in event-based motion estimation, continuous state and measurement representation is presented. Section III provides an overview of the proposed framework IDOL while Section IV and Section V gives more details about the front-end and back-end, respectively. In Section VI, IDOL is evaluated on public datasets and compared to state-of-the-art in traditional frame-based VIO. Finally, Section VII provides a conclusion with an outlook on future work.

II. Related Work

A. Event-based motion estimation

The major differences between the traditional frame-based and event-based vision make the latter an especially appealing sensing modality for the task of VO in challenging scenarios where the performance of traditional imagers is compromised such as in HDR scenes or under fast camera motions. Despite the relatively recent interest of the community in event cameras, we can already identify several works on VO employing event cameras over the last years, starting from the first 2D SLAM approach by Weikersdorfer *et al*. [12]. Years later, Kim *et al*. [13] achieved the estimation of the 6-degree of freedom (DoF) of the camera pose on generic 3D scenes using probabilistic filtering. Rebecq *et al*. [14] proposed a parallel mapping and tracking approach that iteratively co-localize the pose of the camera against a local map of edges represented by a voxel grid.

Aforementioned approaches avoid the explicit definition of features at expenses of being relatively demanding in terms of computational resources. Modern event-based VIO pipelines such as [15]–[17], however, rely on detection and tracking of corners employing intermediate image-like representations from the accumulation of events. These approaches make use of IMUs, profiting from their high-rate of inertial measurements and making them an appealing type of sensor to be combined with event cameras.

While the bulk of event-based VO approaches still rely on the traditional concept of key-frames, it is only natural that continuous-time approaches would emerge. One of the seminal works is introduced by Mueggler *et al*. [18], in which the state of the camera is estimated associating events with line segments evaluated both in simulation and real-world experiments using fiducial markers detected using intensity images. The same authors later expand their approach to consider inertial measurements in [19].

Over the years, significant efforts in the community have been dedicated to the definition of reliable visual features for event data. It has led to the development of mainly corner detection and tracking approaches (*e.g.* [20]–[23]), while the progress of line-based features has been comparatively less notable. Among other works, we could highlight the approaches proposed by Brändli *et al*. [24], detecting and tracking line clusters in an event-by-event fashion, and Everding *et al*. [25], whose propose a plane fitting approach on the event stream based on principal component analysis to track and cluster lines.

The proposed approach takes inspiration from previous continuous-time formulations and their integration with IMU measurements as in [18] and [19]. However, we also draw concepts from event-driven line tracking and clustering approaches such as the methods described in [24] and [25], avoiding the need for supplementary sensing modalities as frame-based images and operating directly on the asynchronous event stream.

B. Continuous state and measurement representation

The asynchronism of event cameras represents a major technical challenge for state estimation. While originally focused on discrete-time estimation, the use of rolling-shutter-like sensors (lidar, rolling-shutter camera, etc.), as well as multi-sensor platforms, lead the robotics community to develop continuous-time state estimation theory and methods. A large number of frameworks assume motion-models, often constant velocity, to interpolate the state variables in between discrete estimation timestamps [26], [27]. In [28], Furgale *et al*. introduce a fully continuous framework that considers the state as being the linear combination of temporal basis functions. Anderson and Barfoot [29] present a probabilistic approach to efficiently infer the state variables using Gaussian Process (GP) regression over a discrete maximum a posteriori estimation.

A different paradigm is presented in [30], [31] where GPs are used as continuous representations of the inertial data allowing the characterization of the system's pose in a continuous manner while relying on a discrete state. This is the approach employed in the proposed method with the use of Gaussian preintegrated measurements (GPMs) originally presented in [32].

III. Method overview

The proposed method aims at estimating the ego-motion of an event-based visual-inertial system. To this end, line segments are detected in the event data provided by the camera, and both the system's trajectory and the position of the 3D lines are simultaneously estimated. Addressing the asynchronicity of the event stream, a continuous representation of the inertial data is used to associate each event with inertial measurements. The state is then estimated by means of a discrete-state batch on-manifold optimization that accounts for the events individually.

A. Problem formulation

Let us consider a rigidly mounted event-camera and a 6-DoF IMU. The camera and IMU reference frames at time t_i are respectively noted $\mathfrak{F}_C^{t_i}$ and $\mathfrak{F}_I^{t_i}$. The relative transformation from $\mathfrak{F}_I^{t_i}$ to $\mathfrak{F}_C^{t_i}$ is characterized by the rotation matrix $\mathbf{R}_I^C$ and the translation vector $\mathbf{p}_I^C$. Homogeneous transformation will be used for the rest of the paper, therefore rotation matrices and translations/positions will be associated with 4×4 transformation matrices with the same combination of subscripts and superscripts,

$$\mathbf{T}_a^b = \begin{bmatrix} \mathbf{R}_a^b & \mathbf{p}_a^b \\ \mathbf{0}^\top & 1 \end{bmatrix} \text{ and } \mathbf{T}_a^{b\,-1} = \begin{bmatrix} \mathbf{R}_a^{b\,\top} & -\mathbf{R}_a^{b\,\top}\mathbf{p}_a^b \\ \mathbf{0}^\top & 1 \end{bmatrix}. \quad (1)$$

The 6-DoF IMU acquires proper acceleration $\tilde{\mathbf{f}}_{\mathfrak{t}_q}$ and angular velocity $\tilde{\boldsymbol{\omega}}_{\mathfrak{t}_q}$ measurements at time $\mathfrak{t}_q$ ($q = 1, \cdots, Q$). These readings are combined together into GPMs [32] assuming known IMU biases. Section III-B presents a brief introduction on GPMs and the mechanism of post-integration bias correction. The event-camera data is collected as an asynchronous stream of events $\mathbf{e}^i = \begin{bmatrix} e_x^i & e_y^i \end{bmatrix}$ at time t_i, with e_x^i and e_y^i being the pixel coordinates in the image space. Note that in this work, we do not consider the polarity of the events. This stream is arbitrarily organised in M windows of N events. A 3D point is projected into the image space with the function $\pi(\bullet)$ according to a pinhole camera-model.

The system's IMU orientation $\mathbf{R}_W^{\tau_m}$, position $\mathbf{p}_W^{\tau_m}$, and velocity $\mathbf{v}_W^{\tau_m}$ are estimated at the timestamp τ_m of the first event of each window ($m = 1, \cdots, M$) with respect to the fixed world frame $\mathfrak{F}_W$. The proposed method also estimates corrections, $\hat{\mathbf{b}}_f^m$ and $\hat{\mathbf{b}}_\omega^m$, to the bias priors used during the preintegration.

Along the sensor's trajectory, events are clustered into segments that correspond to 3D lines in the environment (front-end). The accumulation of these event-line associations form the set $\mathcal{A}$. The positions of these lines are estimated simultaneously to the IMU pose and velocities mentioned above. Each line is parameterized by two 3D points as

$$\mathbf{L}_W^l = \begin{bmatrix} \mathbf{l}_{a\,W}^{l\,\top} & \mathbf{l}_{b\,W}^{l\,\top} \end{bmatrix}, \quad (2)$$

with $l = 1, \cdots, L$. While this representation over-parameterizes 3D lines, it allows the proposed method to fit the lines' extremities to the actual line segments through a mechanism of attraction/repulsion detailed in Section IV-B.

The proposed method estimates the state $\mathcal{S} = (\mathbf{R}_W^{\tau_1}, \cdots, \mathbf{R}_W^{\tau_M}, \ \mathbf{p}_W^{\tau_2}, \cdots, \mathbf{p}_W^{\tau_M}, \ \mathbf{v}_W^{\tau_2}, \cdots, \mathbf{v}_W^{\tau_M}, \ \hat{\mathbf{b}}_f^{\tau_1}, \cdots, \hat{\mathbf{b}}_f^{\tau_M}, \hat{\mathbf{b}}_\omega^{\tau_1}, \cdots, \hat{\mathbf{b}}_\omega^{\tau_M}, \mathbf{L}_W^1, \cdots, \mathbf{L}_W^L)$ using maximum likelihood estimation that corresponds to the minimization of the cost function C:

$$\mathcal{S}^* = \underset{S}{\operatorname{argmin}}\, C(\mathcal{S}),$$

$$C(\mathcal{S}) = \sum_{\alpha^i \in \mathcal{A}} \left(\|r_l^{\alpha^i}\|^2_{\Sigma_{\mathbf{r}_l^{\alpha^i}}} + \|r_s^{\alpha^i}\|^2_{\Sigma_{\mathbf{r}_s^{\alpha^i}}} \right) + \sum_{l=1}^{L} \|r_a^l\|^2_{\Sigma_{\mathbf{r}_a^l}} + \sum_{m=1}^{M-1} \left(\|\mathbf{r}_f^m\|^2_{\Sigma_{\mathbf{r}_f^m}} + \|\mathbf{r}_\omega^m\|^2_{\Sigma_{\mathbf{r}_\omega^m}} + \|\mathbf{r}_I^m\|^2_{\Sigma_{\mathbf{r}_I^m}} \right), \quad (3)$$

with r_l being event-to-line distances for each event-line association in $\mathcal{A}$, r_s and r_a being repulsion and attraction forces between each of the lines' extremities, respectively, $\mathbf{r}_f$ and $\mathbf{r}_\omega$ constraints on the IMU biases random-walk, and $\mathbf{r}_I$ being direct pose and velocity constraints between two consecutive timestamps of the estimated trajectory based on the IMU readings. These factors (back-end) are detailed in Section IV. Note that $\Sigma_\bullet$ is the covariance matrix of the variable $\bullet$.

B. Gaussian Preintegrated Measurement

The GPMs [32] rely on the use of GP regression and linear operators to preintegrate (analytically for the position and velocity parts) the inertial measurements between any given timestamps. Therefore, given the state $\mathcal{S}$, the pose and velocity of the system can be queried at t_i as

$$\begin{aligned} \mathbf{p}_W^{t_i} &= \mathbf{p}_W^{\tau_m} + (t_i - \tau_m)\mathbf{v}_W^{\tau_m} + (t_i - \tau_m)^2 \frac{\mathbf{g}}{2} + \mathbf{R}_W^{\tau_m} \Delta\mathbf{p}_{\tau_m}^{t_i} \\ \mathbf{v}_W^{t_i} &= \mathbf{v}_W^{\tau_m} + (t_i - \tau_m)\mathbf{g} + \mathbf{R}_W^{\tau_m} \Delta\mathbf{v}_{\tau_m}^{t_i} \\ \mathbf{R}_W^{t_i} &= \mathbf{R}_W^{\tau_m} \Delta\mathbf{R}_{\tau_m}^{t_i} \end{aligned} \quad (4)$$

where $\mathbf{g}$ is the known gravity vector in $\mathfrak{F}_W$, and $\Delta\mathbf{p}_{\tau_m}^{t_i}$, $\Delta\mathbf{v}_{\tau_m}^{t_i}$, and $\Delta\mathbf{R}_{\tau_m}^{t_i}$ the position, velocity and rotation GPMs, respectively. To prevent overly verbose notation, the superscript ^{t_i} and $^{\tau_m}$ refer to the IMU frame at time t_i and τ_m.

These pseudo-measurements are functions of the acceleration biases $\mathbf{b}_f$, and gyroscope biases $\mathbf{b}_\omega$. Unfortunately, these values are not accurately known at the time of preintegration. Consequently, as in [33], the GPM approach uses the first-order Taylor expansion of each of the preintegrated measurements $\Delta\mathbf{p}_{\tau_m}^{t_i}$, $\Delta\mathbf{v}_{\tau_m}^{t_i}$, and $\Delta\mathbf{R}_{\tau_m}^{t_i}$, and assumes the biases are individually constant in each of the M windows. The expansion is based on the approximation that $\mathbf{b}_f^m \approx \bar{\mathbf{b}}_f^m + \hat{\mathbf{b}}_f^m$ and $\mathbf{b}_\omega^m \approx \bar{\mathbf{b}}_\omega^m + \hat{\mathbf{b}}_\omega^m$, where $\bar{\mathbf{b}}_f^m$ and $\bar{\mathbf{b}}_\omega^m$ are the prior knowledge of the biases used to compute the GPMs, and $\hat{\mathbf{b}}_f^m$ and $\hat{\mathbf{b}}_\omega^m$ are their first-order correction.

IV. Back-end

This section describes the different elements of the cost function $C(\mathcal{S})$ presented in Equation (3).

A. Event-to-line factors

The event-to-line factors correspond to the point-to-line distances between events in the image space and the image projection of the associated 3D lines. Let us consider an event-line association $\alpha^i = \{\mathbf{e}^i, t_i, \mathbf{L}_W^l\}$. The projections $\mathbf{d}_{a_l}^{t_i}$ and $\mathbf{d}_{b_l}^{t_i}$ of the line extremities $\mathbf{l}_{a\,W}^{l}$ and $\mathbf{l}_{b\,W}^{l}$ into the camera image at time t_i are obtained using Equation (4) and the extrinsic calibration $\mathbf{T}_I^C$

$$\mathbf{d}_{a_l}^{t_i} = \pi(\mathbf{T}_I^{C\,\top}\mathbf{T}_W^{t_i\,\top}\mathbf{l}_{a\,W}^{l}) \text{ and } \mathbf{d}_{b_l}^{t_i} = \pi(\mathbf{T}_I^{C\,\top}\mathbf{T}_W^{t_i\,\top}\mathbf{l}_{b\,W}^{l}). \quad (5)$$

The point-to-line distance residual $r_l^{\alpha^i}$ is then equal to

$$r_l^{\alpha^i} = \frac{\|(\mathbf{e}^i - \mathbf{d}_{a_l}^{t_i}) \times (\mathbf{d}_{b_l}^{t_i} - \mathbf{d}_{a_l}^{t_i})\|}{\|\mathbf{d}_{b_l}^{t_i} - \mathbf{d}_{a_l}^{t_i}\|} \quad (6)$$

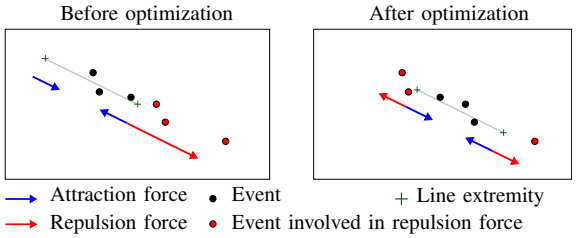


Fig. 2: Illustration of the attraction/repulsion mechanism used to fit the line extremities to the observed segment, as well as to constrain the over-parameterized two-point line representation. The estimated 3D lines extremities are projected into the image space. The extremities are subject to a constant attraction force toward one another. The events that are "outside" the line projection induce a repulsion force that push the extremities apart. After optimization, the estimated line fits the actual segments.

B. Line attraction and repulsion factors

The two-3D-point representation in Equation (2) is an over-parameterization of an infinite 3D line, and the event-to-line factors do not fully define the position of the line points. In other words, without additional constraints, there is ambiguity on the estimated state as an infinite number of point pairs can characterize the same line. Having extra unconstrained DoFs may lead to failure of the optimization process. To address this issue, we introduce an attraction/repulsion strategy that fits the line extremities to the cluster of events, constraining, therefore, the extra DoFs of the two-point line parameterization. Intuitively, the two components of this mechanism can be thought as of, on the one side, an inherent attraction force between the line extremities, and on the other, a set of forces generated by the events pushing the extremities apart as illustrated in Figure 2.

Formally, the attraction component is implemented as a residual equal to the square-root of the pixel-distance between the line extremities after projection into the image space at τ_m, the time of the window during which the line has been first observed:

$$r_a^l = \sqrt{\|\mathbf{d}_{b_l}^{\tau_m} - \mathbf{d}_{a_l}^{\tau_m}\|}. \tag{7}$$

The repulsion component is generated by the points around the extremities of the line. Given an event-line association α^i the position d^i of $\mathbf{e}^i$ along the line $\mathbf{d}_{a_l}^{t_i}/\mathbf{d}_{b_l}^{t_i}$ is computed as

$$d^i = \frac{(\mathbf{e}^i - \mathbf{d}_{a_l}^{t_i})^\top (\mathbf{d}_{b_l}^{t_i} - \mathbf{d}_{a_l}^{t_i})}{\|\mathbf{d}_{b_l}^{t_i} - \mathbf{d}_{a_l}^{t_i}\|}. \tag{8}$$

The events that are "outside" the line lead to residuals equal to the distance along the line to the closest extremity:

$$r_s^{\alpha^i} = \begin{cases} d^i & \text{if } d^i < 0 \\ \|\mathbf{d}_{b_l}^{t_i} - \mathbf{d}_{a_l}^{t_i}\| - d^i & \text{if } d^i > \|\mathbf{d}_{b_l}^{t_i} - \mathbf{d}_{a_l}^{t_i}\| \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Note that this approach does not require the front-end to extract line extremities among the event data. The events that are projected outside the estimated line segments automatically generate repulsion forces. Therefore, the positions of the line extremities are best estimated when solely a small number of events generate repulsion forces. As repulsion forces are invariant to the lines' length but correlated to the number of events involved, the attraction forces also need to be somewhat length-invariant to prevent the need for any additional balancing mechanism between the forces. Intuitively, in least-square optimization problems, the Jacobians of the cost function represent forces that constrain the state estimate. Consequently, the attraction forces are made length-invariant by the use of the square root in r_a^m, making the magnitude of the attraction force "constant" (preventing long line estimates to be "squashed" if Euclidean norms were directly used).

C. IMU and bias factors

The inertial measurements are used to constrain the system's trajectory from one window to the next. The IMU factors' residual $\mathbf{r}_I^m = \begin{bmatrix}\mathbf{r}_{I_p}^m & \mathbf{r}_{I_v}^m & \mathbf{r}_{I_r}^m\end{bmatrix}$ are obtained by manipulating Equation (4),

$$\begin{aligned}
\mathbf{r}_{I_p}^m &= \mathbf{R}_W^{\tau_m\top}(\mathbf{p}_W^{\tau_{m+1}} - \mathbf{p}_W^{\tau_m} - \Delta\tau_m \mathbf{v}_W^{\tau_m} - \Delta\tau_m^2 \frac{\mathbf{g}}{2}) - \Delta\mathbf{p}_{\tau_m}^{\tau_{m+1}} \\
\mathbf{r}_{I_v}^m &= \mathbf{R}_W^{\tau_m\top}(\mathbf{v}_W^{\tau_{m+1}} - \mathbf{v}_W^{\tau_m} - \Delta\tau_m \mathbf{g}) - \Delta\mathbf{v}_{\tau_m}^{\tau_{m+1}} \\
\mathbf{r}_{I_r}^m &= \mathrm{Log}(\Delta\mathbf{R}_{\tau_m}^{\tau_{m+1}\top}\mathbf{R}_W^{\tau_m\top}\mathbf{R}_W^{\tau_{m+1}})
\end{aligned} \tag{10}$$

with $\Delta\tau_m = \tau_{m+1} - \tau_m$, and $\mathrm{Log}(\bullet)$ the mapping from $SO(3)$ (rotation matrix) to $\mathfrak{so}(3)$ (axis-angle).

The proposed method models the temporal evolution of the biases with a Brownian motion. Consequently, the bias factors' residuals are defined as

$$\begin{aligned}
\mathbf{r}_f^m &= \bar{\mathbf{b}}_f^{m+1} + \hat{\mathbf{b}}_f^{m+1} - \bar{\mathbf{b}}_f^m - \hat{\mathbf{b}}_f^m \\
\mathbf{r}_\omega^m &= \bar{\mathbf{b}}_\omega^{m+1} + \hat{\mathbf{b}}_\omega^{m+1} - \bar{\mathbf{b}}_\omega^m - \hat{\mathbf{b}}_\omega^m.
\end{aligned} \tag{11}$$

V. Front-end

Similarly to [25], the proposed method considers the stream of events as 3D information, where the first two components are the events' coordinates in the image space, and the third coordinate is the events' timestamps arbitrarily normalized: $\mathbf{x}^i = \begin{bmatrix}e_x^i & e_y^i & t_i/c\end{bmatrix}^\top$. The value of c is chosen according to the average level of texture in the scene. The front-end consists of clustering events that are triggered by the same physical line in the environment according to the premise that 3D lines translate to locally planar patches in the 3D spatio-temporal representation of the event stream. At the moment, the presented approach does not aim for real-time operation, and thus it uses windows of N events to perform the event clustering. The event data in each window can be seen as a point cloud with the 3D-points' coordinates defined as $\mathbf{x}^i$. Normal vectors are estimated for each of the points based on the eigenvectors of the covariance matrix built with the neighbouring points. Points are considered neighbours if their Euclidean distance in the 3D spatio-temporal space is under a certain given threshold. Note that the proposed method does not explicitly fit planes to the 3D event data throughout the windows, but computes the normals based on a local neighbourhood of events. The x and y components of the normals are normalized to unit vectors $\mathbf{n}^i$.

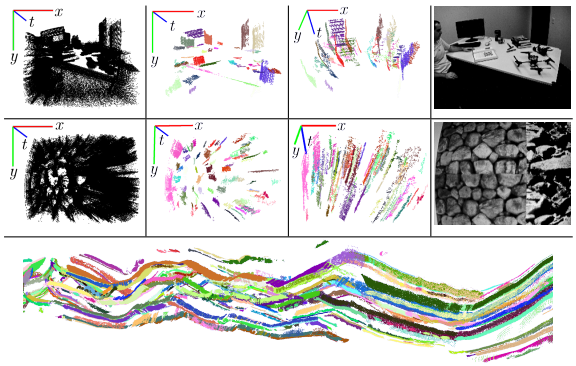


Fig. 3: Examples of line clusters extracted in different dataset. The first two rows (raw event data, line clusters viewpoint A, and B, corresponding greyscale image) correspond to windows of 200k events. Bottom is cluster extraction on a very large window of 2M events in `shapes_6dof`.

The actual clustering is applied in a region growing fashion inspired by the connected component segmentation implemented in [34]. Colinearity of the x-y normal vectors ($\mathbf{n}^{i\top}\mathbf{n}^{j} > n_{\text{thr}}$) and the point-to-line distance in the image space ($|\mathbf{n}^{i\top}(\mathbf{e}^{j}-\mathbf{e}^{i})| < e_{\text{thr}}$) are used as the similarity criteria to assert that two neighbouring points belong to the same cluster. Figure 1 and 3 show examples of event clusters in the three types of datasets used in Section VI. By considering only the x-y normals with a region growing algorithm, the proposed method allows for line clustering in a large variety of scenarios that are not restricted by the nature of the system's motion (e.g. constant velocity).

The segment association between two consecutive windows is conducted by appending the last events of a window to the beginning of the following one and using the same similarity criteria described for the in-window clustering. Each connected segment is attached to a line $\mathbf{L}_W^l$ and the event-line associations $\alpha^i = \{\mathbf{e}^i, t_i, \mathbf{L}_W^l\}$ are pushed to the set $\mathcal{A}$. Note that, in the current implementation, the event-to-line association is only performed at the front-end level. There is no additional strategy to associate new segments to existing line estimates. This would greatly improve the accuracy and robustness of the proposed method in scenarios where a same line results in multiple clusters that cannot be matched via the aforementioned procedures, or when lines reappear in the field of view after having left it.

VI. Experiments

Our implementation uses *Ceres*[1] for the non-linear least-square optimization of Equation (3). At the current stage of development, IDOL is computationally expensive as the full batch optimization is conducted every time a new event-window is processed. The rapidly growing number of residuals leads to prohibiting estimation time of the system's Hessian. We arbitrarily chose to compute the full batch optimization until $t = 24\,\text{s}$. Intuitively, increasing the length of the full-batch time-window robustifies the system with respect to front-end failures. Then we switch to a sliding window optimization over the last 30 event-windows with the last marginalized pose estimate fixed.

Due to the lack of publicly available event-based VIO algorithms, ROVIO [5] was chosen as a comparison. ROVIO is a light-weight VIO algorithm that operates on traditional intensity images and is built upon an extended Kalman filter (EKF) back-end. The front-end of ROVIO extracts patch-based features that are matched based on a photometric error resulting in a semi-dense approach. The system has been shown to be very robust, even under very aggressive motions [5].
Since IDOL, in contrast to ROVIO, performs batch optimization, for a fair comparison we also included results obtained with maplab [35] by building a visual-inertial (VI) pose-graph using ROVIO and jointly optimizing the trajectory and sparse BRISK landmarks [36] using the maplab toolbox.

A. Datasets and Evaluations

In order to test IDOL in challenging conditions and evaluate the performance compared to state-of-the-art in robust traditional VIO, tests on multiple real-world datasets including aggressive motions of the Event Dataset and Simulator [11] were performed. The Event Dataset contains sensor data from a DAVIS240 [10] including the event data, traditional grey-scale images and IMU measurements. In particular, the indoors scenarios `shapes`, `poster` and `dynamic` were chosen as they include 6-DoF ground-truth from an indoor positioning system. Both translation-only and full 6-DoF versions of these datasets have been used. Rotation-only datasets were omitted since translation is necessary for good observability of the scene depth [15]. In its current state, our front-end does not offer enough robustness to address the `boxes` scenarios of [11] because of their very high level of texture in the scene. In addition to reporting qualitative results of the state progression and root mean squared errors (RMSEs) of aligned trajectories, the evaluation is also performed using a trajectory-segment based approach [37] with segment lengths corresponding to $\{10, 20, 30, 40, 50\}\,\%$ of the trajectory length.

B. Results

Figures 4-6 depict the state progression of the ground-truth and estimations of IDOL, ROVIO and ROVIO+maplab on `shapes_6dof`, `poster_translation` and `dynamic_translation`. It becomes visible that, even though there is some drift in the translation estimation, IDOL is able to estimate the camera pose and velocity, and especially the camera rotation, with high accuracy.
After the initial stages of each of the experiments, we observe that the proposed VIO pipeline's translation estimate tends to diverge, which can be explained as a combination of different factors. One can see that the state variables are generally well estimated during the full-batch optimization that takes places at the initial instants of each experiment and only starts drifting seconds after switching to the sliding-window mode. Our implementation could benefit from leveraging the uncertainty of the estimated

[1] http://ceres-solver.org/

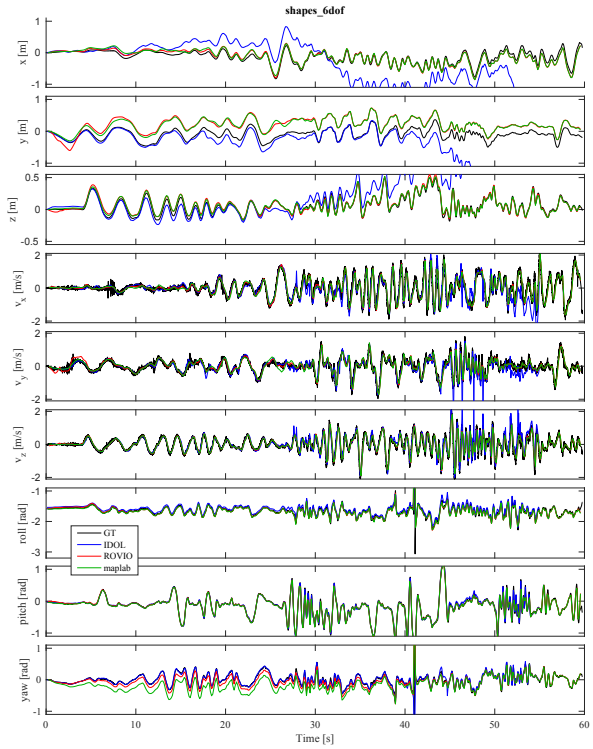


Fig. 4: Pose and velocity estimates' progression of the different algorithms on the `shapes_6dof` dataset.

pose across consecutive windows of optimization, properly marginalizing previous states. We must also consider that, in its current form, the front-end described in Section V produces rather short line-tracks. Figure 7 depicts the average track length across the datasets. One can observe the correlation between the drop of the track length around $t = 27\text{s}$ and $t = 24\text{s}$ in `shapes_6dof` and `poster_translation`, respectively, and the sudden increase of the translation errors. Short track lengths and the absence of a strategy to re-detect previously observed lines do not account for a good depth estimation due to the low parallax. Consequently, the translation estimates are notably affected whereas the rotation estimates, not as dependent on the depth estimate of the scene, still perform in a competitive range for small increments. Note that despite a growing drift of the translation estimates likely due to our front-end's weaknesses, IDOL performs accurate velocity estimation all along the trajectory validating its effectiveness for VIO. Figure 7 also shows that our front-end does not perform equally across the different datasets as per the different levels of noise in event data (due to high texture scenes) and the absence of noise filtering strategy. Table I reports RMSE of translation and rotation estimation of all algorithms on each of the datasets. The metric is highly depending on the translation estimate as the alignment method used for VIO minimizes the translation error to find a position and yaw offset for the whole trajectory. Accordingly, if even only one component of the translation estimate diverges, the orientation RMSE is highly influenced by the poor trajectory alignment. Consequently, the results depicted in Table I are obtained by aligning only the 50

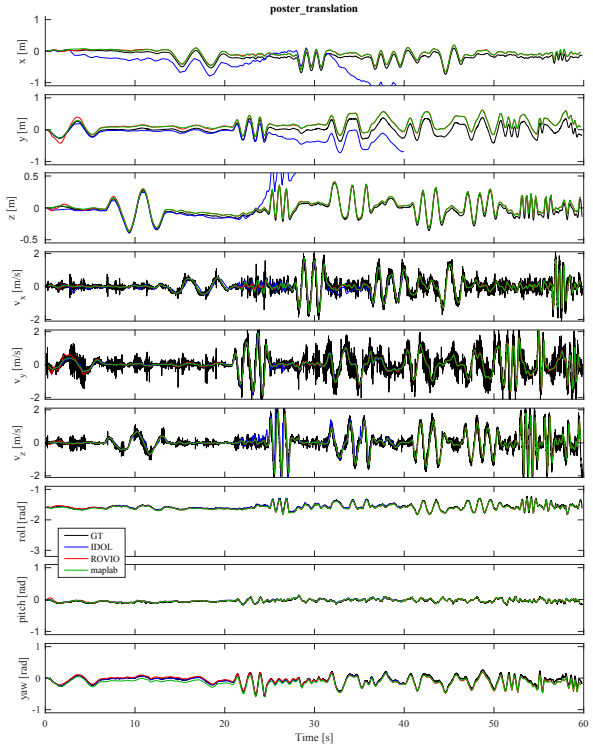


Fig. 5: Pose and velocity estimates' progression of the different algorithms on the `poster_translation` dataset.

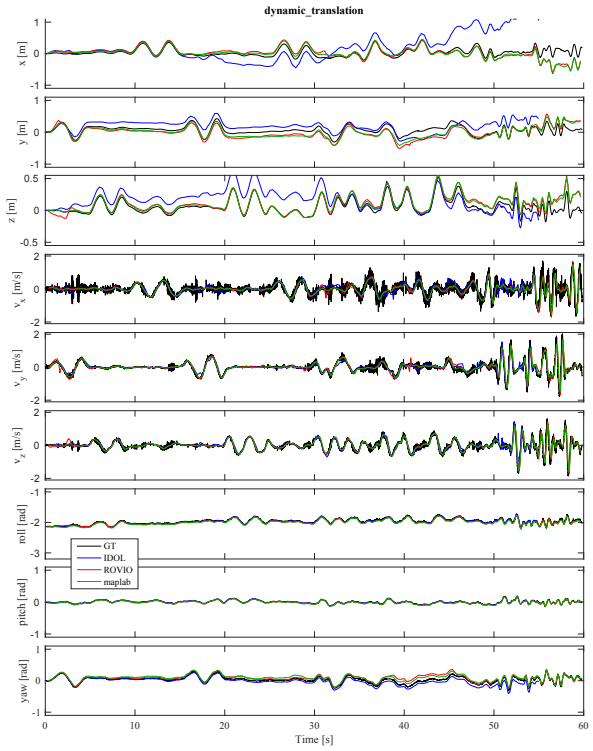


Fig. 6: Pose and velocity estimates' progression of the different algorithms on the `dynamic_translation` dataset.

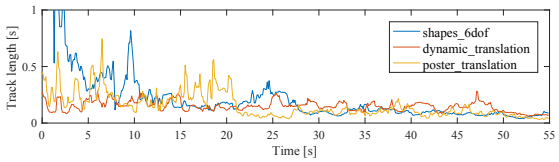
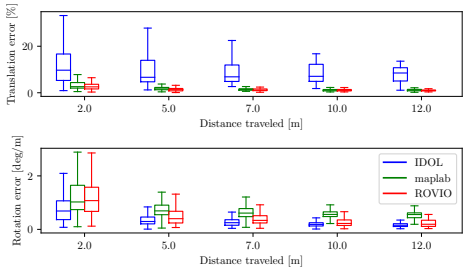
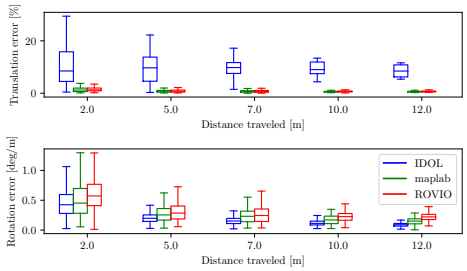


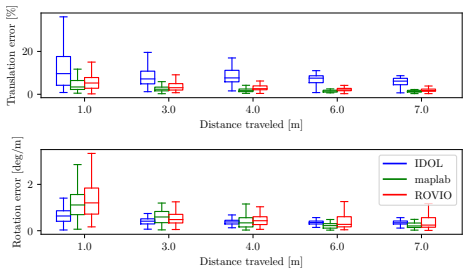
Fig. 7: Average line-track lengths (moving average) across different datasets.



(a) `shapes_6dof`



(b) `poster_translation`



(c) `dynamic_translation`

Fig. 8: Translation and orientation error of the first 40 s using the different algorithms upon different segment lengths.

first poses of the estimation and only taking the first 40 s into account. Typically, IDOL performs worse than ROVIO and ROVIO+maplab in terms of translation estimation but still on the same order of magnitude, especially on `shapes_6dof` and `dynamic_translation`. Moreover, competitive results in orientation estimation can be achieved for most datasets. Only `shapes_translation` and `poster_translation` show a worse rotation estimation compared to ROVIO mainly due to diverged translation estimation (see Figures 4-6).

More insights into the alignment issue can be found in the segment based evaluation shown in Figure 8. It becomes clear that, in its current state, IDOL can outperform both ROVIO and ROVIO+maplab in incremental rotation estimation but mainly lacks in translation estimation.

The results displayed above demonstrate the ability of IDOL to perform VIO in various real-world scenarios. Overall, the proposed method performs in the same order of magnitude than the compared frame-based methods according to the presented results. However, note that, at submission time, we are using an unrefined implementation of the event-based front-end, in which no noise filtering or re-detection scheme is applied. For instance, the knowledge of the line extremities' position could help the data association in scenarios in which the camera re-observes previously mapped areas. Additionally, no protection against outliers has been placed in the front-end nor in the back-end. In this regard, substantial improvements in performance are to be expected from the adoption of more robust strategies from the event-based and frame-based VIO literature with, for example, the adoption of a robust loss function.

VII. Conclusions

This paper introduced IDOL, a novel pipeline for event-based VIO using lines as features. Unlike most of the event-based methods in the literature, the proposed method does not aggregate events into frames that are later used in a traditional frame-based VIO pipeline. Here, the events are considered individually as part of a batch optimization that estimates the position of 3D lines alongside the system's pose and velocity. This framework leverages the GPMs to characterize the system's trajectory based on a continuous representation of the inertial data.

We demonstrated the feasibility of event-based VIO with lines in different real-world scenarios. While the presented system does not have real-time capabilities yet, IDOL results show the potential to become an efficient, robust, and accurate event-based VIO method. Across quantitative benchmarking against state-of-the-art frame-based VIO algorithms, IDOL demonstrated accuracy in the same order of magnitude for translation, and competitive results for orientation.

Future work includes the exploration of different optimization strategies and the implementation of a probabilistic marginalization of the past state variables to substitute for the current growing full-batch optimization. The computational cost of IDOL can be addressed at different levels. For example, one could optimize the implementation by using graphic processing unit (GPU) computation as most of the operations are highly parallelizable (normal estimations for each of the events in their spatio-temporal representation, computation of the residuals and Jacobians, inference of the per-event GPMs, etc.). At a higher level, reducing the size of the optimization problem would greatly benefit the method's efficiency. We will investigate different strategies to reduce the number of residuals while not sacrificing the amount of information used in the state estimation. Finally, combining our line-based features with corner-based features is likely to improve both robustness and accuracy.

References

[1] R. Siegwart, I. R. Nourbakhsh, and D. Scaramuzza, *Introduction to Autonomous Mobile Robots*, 2nd ed., Ronald C. Arkin, Ed. Cambridge: MIT Press, 2011.

	shapes				poster				dynamic			
	6dof		translation		6dof		translation		6dof		translation	
	e_t	e_r	e_t	e_r	e_t	e_r	e_t	e_r	e_t	e_r	e_t	e_r
IDOL	0.52	8.35	0.51	18.62	0.62	6.15	0.70	10.82	0.54	6.08	0.25	4.92
ROVIO	0.34	10.24	0.05	1.56	0.15	8.21	0.13	2.09	0.16	10.27	0.12	6.19
ROVIO + maplab	0.25	12.91	0.05	2.25	0.12	3.53	0.09	1.92	0.08	6.09	0.09	3.15

TABLE I: Overall RMSE of translation estimation e_t [m] and orientation estimation e_r [deg] for IDOL, ROVIO and ROVIO+maplab on different datasets ($0 - 40$ s, aligning first 50 poses).

[2] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," *IEEE TRANSACTIONS ON ROBOTICS*, vol. 31, no. 5, 2015.

[3] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, 12 2016.

[4] J. Engel, V. Koltun, and D. Cremers, "Direct Sparse Odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[5] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct EKF-based approach," in *IEEE International Conference on Intelligent Robots and Systems*, Seattle, WA, USA, 2015, pp. 298–304.

[6] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, 2015.

[7] T. Qin, P. Li, and S. Shen, "VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[8] F. Tschopp, T. Schneider, A. W. Palmer, N. Nourani-Vatani, C. Cadena Lerma, R. Siegwart, and J. Nieto, "Experimental Comparison of Visual-Aided Odometry Methods for Rail Vehicles," *IEEE Robotics and Automation Letters*, 2019.

[9] P. Lichtsteiner, C. Posch, T. Delbruck, and S. Member, "A 128x128 120 dB 15 μs Latency Asynchronous Temporal Contrast Vision Sensor," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[10] C. Brandli, R. Berner, Minhao Yang, Shih-Chii Liu, and T. Delbruck, "A 240 x 180 130 dB 3 us Latency Global Shutter Spatiotemporal Vision Sensor," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, 10 2014.

[11] E. Mueggler, H. Rebecq, G. Gallego, T. Delbruck, and D. Scaramuzza, "The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM," *International Journal of Robotics Research*, vol. 36, no. 2, pp. 142–149, 2017.

[12] D. Weikersdorfer, R. Hoffmann, and J. Conradt, "Simultaneous localization and mapping for event-based vision systems," in *Computer Vision Systems*, 2013, pp. 133–142.

[13] H. Kim, S. Leutenegger, and A. J. Davison, "Real-time 3D reconstruction and 6-DoF tracking with an event camera," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 349–364.

[14] H. Rebecq, T. Horstschaefer, G. Gallego, and D. Scaramuzza, "EVO: A geometric approach to event-based 6-DOF parallel tracking and mapping in real time," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 593–600, 2017.

[15] A. Zhu, N. Atanasov, and K. Daniilidis, "Event-based visual inertial odometry," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[16] H. Rebecq, T. Horstschaefer, and D. Scaramuzza, "Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[17] A. R. Vidal, H. Rebecq, T. Horstschaefer, and D. Scaramuzza, "Ultimate slam? combining events, images, and imu for robust visual slam in hdr and high-speed scenarios," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 994–1001, April 2018.

[18] E. Mueggler, G. Gallego, and D. Scaramuzza, "Continuous-time trajectory estimation for event-based vision sensors," in *Robotics: Science and Systems XI*, no. EPFL-CONF-214686, 2015.

[19] E. Mueggler, G. Gallego, H. Rebecq, and D. Scaramuzza, "Continuous-time visual-inertial odometry for event cameras," *IEEE Transactions on Robotics*, vol. 34, no. 6, pp. 1425–1440, Dec 2018.

[20] V. Vasco, A. Glover, and C. Bartolozzi, "Fast event-based harris corner detection exploiting the advantages of event-driven cameras," in *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, Oct 2016, pp. 4144–4149.

[21] J. Manderscheid, A. Sironi, N. Bourdis, D. Migliore, and V. Lepetit, "Speed invariant time surface for learning to detect corner points with event-based cameras," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[22] I. Alzugaray and M. Chli, "Asynchronous corner detection and tracking for event cameras in real time," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3177–3184, Oct 2018.

[23] ——, "ACE: An efficient asynchronous corner tracker for event cameras," in *International Conference on 3D Vision (3DV)*, Sep 2018.

[24] C. Brandli, J. Strubel, S. Keller, D. Scaramuzza, and T. Delbruck, "ELiSeD – an event-based line segment detector," in *International Conference on Event-Based Control, Communication and Signal Processing (EBCCSP)*, June 2016, pp. 1–7.

[25] L. Everding and J. Conradt, "Low-latency line tracking using event-based dynamic vision sensors," *Frontiers in Neurorobotics*, vol. 12, p. 4, 2018.

[26] S. Hong, H. Ko, and J. Kim, "VICP: Velocity updating iterative closest point algorithm," *Proceedings - IEEE International Conference on Robotics and Automation*, no. Section 3, pp. 1893–1898, 2010.

[27] M. Bosse and R. Zlot, "Continuous 3D Scan-Matching with a Spinning 2D Laser," *IEEE International Conference on Robotics and Automation*, 2009.

[28] P. Furgale, T. D. Barfoot, and G. Sibley, "Continuous-Time Batch Estimation using Temporal Basis Functions," *IEEE International Conference on Robotics and Automation*, pp. 2088–2095, 2012.

[29] S. Anderson and T. D. Barfoot, "Full STEAM ahead: Exactly sparse Gaussian process regression for batch continuous-time trajectory estimation on SE(3)," *IEEE International Conference on Intelligent Robots and Systems*, vol. 2015-Decem, no. 3, pp. 157–164, 2015.

[30] C. Le Gentil, T. Vidal-Calleja, and S. Huang, "3D Lidar-IMU Calibration based on Upsampled Preintegrated Measurements for Motion Distortion Correction," *IEEE International Conference on Robotics and Automation*, 2018.

[31] ——, "IN2LAMA : INertial Lidar Localisation And MApping," *IEEE International Conference on Robotics and Automation*, 2019.

[32] ——, "Gaussian Process Preintegration for Inertial-Aided State Estimation," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2108–2114, 2020.

[33] T. Lupton and S. Sukkarieh, "Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions," *IEEE Transactions on Robotics*, vol. 28, no. 1, pp. 61–76, 2012.

[34] K. Zampogiannis, C. Fermüller, and Y. Aloimonos, "Cilantro: A lean, versatile, and efficient library for point cloud data processing," *MM 2018 - Proceedings of the 2018 ACM Multimedia Conference*, pp. 1364–1367, 2018.

[35] T. Schneider, M. Dymczyk, M. Fehr, K. Egger, S. Lynen, I. Gilitschenski, and R. Siegwart, "maplab: An Open Framework for Research in Visual-inertial Mapping and Localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1418–1425, 11 2018.

[36] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proceedings of the International Conference on Computer Vision (ICCV)*, Nov 2011, pp. 2548–2555.

[37] Z. Zhang and D. Scaramuzza, "A Tutorial on Quantitative Trajectory Evaluation for Visual(-Inertial) Odometry," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS*. Madrid, Spain: IEEE, 2018, pp. 7244–7251.