Working Paper

Real-time Visual-inertial Localization using Summary Maps

Author(s):
Schneider, Thomas; Dymczyk, Marcin; Hinzmann, Timo; Gilitschenski, Igor; Lynen, Simon; Siegwart, Roland

Publication Date:
2015

Permanent Link:
https://doi.org/10.3929/ethz-b-000332326

Rights / License:
In Copyright - Non-Commercial Use Permitted
Real-time Visual-inertial Localization using Summary Maps

Thomas Schneider*, Marcin Dymczyk*, Timo Hinzmann*, Igor Gilitschenski*, Simon Lynen*†, and Roland Siegwart*

*Autonomous Systems Lab, ETH Zurich †Google Inc.

Abstract—Localization in a global reference frame constitutes a fundamental milestone towards high-level applications in robotics such as autonomous navigation and obstacle avoidance. Visual-inertial SLAM became a compelling approach for this task, despite its inherent drift and local pose estimation. A solution to these shortcomings, however, can be achieved by matching against maps built in previous sessions. Commonly, a careful data selection is performed to keep the map size traceable and thus enable localization in real-time. Although, such a map summarization usually guarantees global localization coverage, the accuracy suffers due to fewer matches.

In this work, we aim at mitigating this effect by directly integrating the scarce 2d-3d matches with visual feature tracks and inertial measurements in the framework of a sliding-window based optimization. We compare our approach to motion tracking data and demonstrate that such a joint estimation yields smoother and more accurate global pose estimates than related methods that loosely integrate 6-DoF localization poses with VIO. Finally, we evaluate the impact of varying map summarization parameters on the trade-off between map-size and localization accuracy and demonstrate that our approach allows for a more aggressive summarization while retaining the robustness and accuracy achieved with larger maps.

I. INTRODUCTION AND RELATED WORK

Being able to precisely determine the location of agents w.r.t. a known frame of reference, for instance using GPS is key to nearly all robotic applications. However, when such global positioning methods are unavailable alternative sensor modalities, such as cameras and time-of-flight sensors in conjunction with state estimation algorithms, become essential. In contrast to GPS signals, these sensor modalities estimate the pose in a local frame of reference that is tied to a map built using Simultaneous Localization and Mapping (SLAM). This lack of a common frame of reference has pivotal impact on collaborative robotic applications such as industrial inspection or disaster relief [1]. In recent years, several approaches have been proposed to co-localize multiple agents in a shared map: Strategies for collaborative mapping can involve a central server for micro-aerial vehicles (MAVs) [2], or teams of ground and aerial robots [3]. On the other hand, collaborative mapping can be performed in a distributed fashion as proposed in [4, 5]. In particular, a team of heterogeneous robots can profit from collaborative mapping and localization as different viewpoints and sensor modalities can be combined into a joint map. The achieved synergetic effects can be diverse for instance extending the range of perception [1].

Fusing camera and inertial sensor measurements has proven to yield highly accurate motion estimates [6, 7]. Therefore, such visual-inertial odometry (VIO) systems have become a common choice for robotic navigation. In parallel, the research community has developed high performance algorithms for loop-closure [8] and vision-based localization against a known 3d-model [10, 11, 12]. Furthermore, variants for real-time operation with [13] and without [14] the need for a server connection have been demonstrated. Combining a visual SLAM system with efficient large-scale localization [13, 14] provides a compelling approach to localize multiple agents in a common frame of reference. These methods localize in maps consisting of 3d-points and descriptors representing the appearance of these landmarks from observing keyframes. These localization maps are often built from multiple VIO trajectories that are merged into a global map and refined using visual-inertial weighted least-squares [15]. Several approaches have been described for merging these sub-maps in a subsequent post-processing step [4, 6], which may also incorporate additional sensors [11] besides vision. Localization of multiple-robots in a common frame-of-reference relies on a compact representation of the localization map. This is important to keep the memory requirement to a minimum and thus allow for low transmission times when sharing maps with other agents. Recently, several methods were presented to select and only retain the most informative localization landmarks of a map [17, 18, 19]. Such methods usually guarantee full localization coverage at a minimal count of localization landmarks. The number of
2d-3d matches available for localization, however, will drop and lead to a decrease in smoothness and accuracy of the global pose estimates.

In contrast to state of the art approaches [13, 20], we aim at a tight integration of inertial measurements with visual feature-tracks and 2d-3d constraints to the global localization map. For this reason, we use a fixed-lag-smoother as our state estimation framework instead of the filter formulation proposed in [14]. The proposed joint estimation of local VIO and global pose improves the smoothness and accuracy of the global pose when localizing against such highly compact maps. To sparsify landmarks and thus reduce the map size, we employ the method of map summarization described in our previous work [19]. This method only retains the most informative landmarks from a map that was initially built from several VIO trajectories and refined with a visual-inertial least squares optimization. During runtime the system concurrently performs localization to the known reference model and visual-odometry in yet unmapped areas. This allows both having an accurate estimate w.r.t. other agents and also mapping previously unvisited areas.

The contributions of this paper are:

- a localization estimator that tightly integrates inertial measurements and visual feature tracks with 2d-3d matches for a global pose w.r.t. a summary map and a local VIO pose estimate,
- an evaluation of the localization error against motion-capture data and a related loosely-coupled approach,
- a demonstration of the improvements in smoothness and accuracy of the global poses estimated in highly compact maps,
- and a validation of the map concept, presented in our previous work [19], against absolute ground-truth data in a real-world scenario with a team of heterogeneous robots shown in Fig. 1.

The remainder of the paper is structured as follows. Section II describes the mapping process that is used to create the compact localization summary maps from several VIO trajectories. Section III introduces the visual-inertial localization that tightly integrates visual feature tracks and inertial data with 2d-3d map matches into a globally pose estimate. Finally, Section IV evaluates the localization error against ground-truth from a motion-capture system and compares it against a related loosely-coupled approach.

II. MAPPING BACKEND

A global localization map is created by, first, running several mapping sessions using visual-inertial odometry to create several local sub-maps. This is carried out using the proposed VIL system from Section III without a reference map. Next, the resulting sub-maps are merged into one global map using appearance-based matching followed by a refinement using a visual-inertial least-squares minimization. And finally, the global map is summarized to only retain the most-informative landmarks for localization. The VIL estimator can then be used to localize against the resulting summary map in real-time. An overview of this process is shown in Fig. 3.

A. Map Representation

The map consists of several sub-maps, so-called missions that each represent a separate agent trajectory. Every mission has its own local frame of reference \( M_i \). These frames are anchored w.r.t. the global frame of reference by a transformation \( T_{GM_i} \). By introducing such additional frames of reference, we can align individual trajectories without modifying keyframe and landmark positions. The transformations \( T_{GM_i} \) are estimated as part of a least-squares minimization involving loop-closure constraints to other missions and/or GPS signals.

Each mission stores a graph of keyframes and a set of corresponding visual landmarks. The vertices of these graphs correspond to keyframes containing visual measurements, the local pose \( T_{M_i,B} \), and the inertial states (velocity and IMU biases). These keyframes are connected by edges holding IMU measurements. Landmarks are observed from and thus associated with multiple keyframes which are part of one mission or in case of loop-closures from multiple missions.

B. Merging of Local Maps

In order to create a global localization map we merge several local VIO maps that have been obtained by running the visual-inertial localization estimator without a reference map (Section III). This is achieved by loop-closing all individual trajectories and performing an alignment based on these constraints. We utilize an implementation of the visual descriptor based loop-closure system described in [9] that associates 2d-image descriptors and 3d-points in the maps. First, the high dimensional binary BRISK descriptors [21] are projected to a lower dimensional real-valued space (here: 10 dimensions) and then inserted into an index. Depending on the map size this is either formed by a KD-tree or
a multi-dimensional product vocabulary \[23\] augmented by KD-trees as proposed in [9]. The raw 2d-3d matches are first filtered using a covisibility graph \[24\] – an approximate set-cover problem is solved in which only 3d landmarks that form a cluster in the covisibility graph are returned. The matches from the dominant cluster are passed to a Perspective-n-Point (PnP) solver in a RANSAC loop to ensure geometric consistency. Once loop-closure correspondences are established an alignment transformation \( T_{GM} \) is computed for each mission \( M_i \) is estimated. Therefore, a least-squares problem is solved using all loop-closures correspondences as constraints. Landmarks are merged if the verified matches indicate that two or more landmarks are actually the same physical landmark. After this process, a joint visual-inertial least-squares is solved to further refine the global map.

### C. Visual-inertial Weighted Least-Squares

A non-linear visual-inertial weighted least-squares (VI-WLS) problem is solved to obtain a consistent global map. We use the Ceres solver \[25\] to minimize the cost \( J(x) \) comprised of visual and inertial error terms. Visual error terms penalize the reprojection error, i.e., the image plane distance between the reprojected 3d landmark position and the measured keypoint location. Inertial error terms penalize the temporal error between two vertices; that is the difference between states of the two vertices and the integrated IMU measurement. The optimization objective is then given by:

\[
J(x) = \sum_{i=1}^{N} \sum_{j=1}^{n(i)} \sum_{k \in \mathcal{K}(i)} e_{i,j,k}^{r} e_{i,j,k}^{T} W_{r,i,j,k} W_{r,i,j,k}^{-1} e_{i,j,k}^{r} e_{i,j,k}^{T} + \sum_{i=1}^{N-1} e_{s}^{T} W_{s} e_{s}^{T} e_{s}
\]

where \( N \) denotes the number of keyframes, \( n(i) \) the number of cameras for the \( i \)-th keyframe, \( \mathcal{K}(i) \) the set of landmarks visible in camera \( j \) of keyframe \( i \), \( e_{i,j,k}^{r} \) the reprojection error of landmark \( k \) in camera \( j \) of keyframe \( i \) and \( e_{i}^{r} \) the temporal IMU error between keyframe \( i \) and \( i+1 \). Terms \( W_{r,i,j,k} \) and \( W_{s} \) denote the weighting information matrices calculated as the inverse of the covariance matrices: keypoint measurement and IMU integration covariance respectively.

### D. Map Summarization

In order to reduce the size of the global localization map we apply a compression and data selection step called summarization. This describes the process of pruning as many landmarks from the map as possible while ensuring sufficient localization coverage over the entire mapped area. Landmark removal is vital, as their 3d positions and especially descriptors constitute the majority of the map size.

To obtain the best possible performance while keeping the computational effort limited, an Integer Linear Programming (ILP) approach is used to select the subset of landmarks which are most informative for localization \[19\]. The result-

\[
\min \; q^T x + \lambda 1^T z
\]

\[
s.t. \; A x + z \geq b, \quad z \in \{0 \} \cup \mathbb{Z}^+ \}
\]

where \( N \) is the initial number of landmarks, \( M \) is the total number of keyframes, \( x \) is a vector of binary switch variables associated with landmarks (1 means the landmark should be retained, 0 otherwise), \( q \) is a vector of scores associated with landmarks (based on the number of observations and the stability of descriptors), \( A \) is a \( M \times N \) visibility matrix, \( b \) is setting a keypoints-per-keyframe threshold, \( z \) is a slack variable and \( n_{desired} \) is the desired number of retained landmarks. This process selects the subset of \( n_{desired} \) landmarks with highest scores, while ensuring localizability of each keyframe, modelled as heuristic constraints (i.e. keyframe observes at least \( b - \zeta \) landmarks).

### III. VISUAL-INERTIAL LOCALIZATION

Due to highly compact maps, often, there is only a very limited number of 2d-3d matches to the localization map available. A global pose estimator solely based on scarce 2d-3d matches would suffer from non-smooth and less accurate estimates. Therefore, we propose to augment the global pose estimation with visual feature tracks and inertial data in the framework of a sliding-window based optimization. This joint estimation improves accuracy and noise characteristics of the global pose when used with highly compact localization maps. Moreover, such a joint estimation of the local VIO and the global pose allows for a seamless switching between localization and exploration mode.

First, we introduce the vision frontend that detects and tracks point-features from image-to-image. Second, we describe the matching stage which establishes correspondences between visual feature tracks and map landmarks. Last, we present the visual-inertial estimator which is used to jointly estimate the robot motion in a local and global frame of reference.

#### A. Vision Frontend

The vision frontend establishes 2d-2d correspondences of keypoints over time. Therefore, we detect AGAST \[26\] keypoints and assign them to bins in a uniform grid on the image plane. In each bin we only retain the \( N \) strongest keypoints according to their detector response. This enforces minimal feature coverage across the image and ensures a good sampling of the observed structure. It is important to note that localization and ego-motion estimation have somewhat different requirements regarding feature selection. The point-based localization relies on a high repeatability of the feature detections whereas the ego-motion estimation requires a uniform sampling of the observed structure. For this reason, a very coarse grid is used when selecting new keypoints for tracking.
The remaining keypoints are tracked between consecutive images using the Lucas-Kanade [27] method. The integrated angular measurements from the gyroscopes are used to predict the keypoint locations between consecutive frames to facilitate and robustify the data association. Finally, a 2-point RANSAC scheme is used to detect and reject outlier matches.

B. Matching Feature Tracks to the Map

All terminated feature tracks (or track above a certain length) are matched against the localization map. In order to find correspondences between 2d observations of the feature tracks to map landmarks, we use an implementation of the method proposed in [14]. To allow matching against a large map, we build a multi-dimensional (product) vocabulary [23] with KD-tree augmentation as an index of all descriptors of the localization map. All descriptors of the feature track are queried for its nearest neighbors to identify landmarks with similar descriptors. These raw matches are clustered based on landmark covisibility [24]. The biggest cluster is then passed to a PnP solver in a RANSAC loop to identify the consistent subset of matches. Finally, the orientation of the recovered camera pose w.r.t. gravity is compared to the current estimate of the VIO. As the localization maps are gravity-aligned, any RANSAC result with a gravity alignment error above a predefined threshold (here: 5 deg) is rejected. A single matched observation of a feature track is used to associate the entire track with this map landmark. Further, the map alignment \( T_{GM} \) obtained in the RANSAC step is used to initialize its linearisation point in the non-linear optimization.

C. Sliding-window Optimization

To allow real-time operation on a robotic platform we solve an approximation of the visual-inertial SLAM problem posed as a non-linear fixed-lag smoother. Therefore, only a fixed number of the most recent visual-inertial keyframes (here: 5) is kept in the optimization window. Older keyframes are marginalized as they are pushed out of the window by new keyframes. The problem is formulated as a factor-graph with the assumption of Gaussian noise. We use GTSAM for solving the resulting non-linear least-squares problem and to marginalize-out old states [28].

At each update, the state is augmented with a new keyframe containing its pose \( T_{MB} \), velocity \( v \) and IMU bias \( b \). The pose is initialized by propagating the pose of the last keyframe using the integrated inertial measurements. We establish an inertial constraint between the current and the last keyframe using the formulation of [29].

All terminated feature tracks obtained by the vision frontend (Section III-A) are separated into odometry and localization tracks. Localization tracks are the feature tracks that were successfully matched to a map landmark (Section III-B). Processing terminated feature tracks ensures good constraints between all keyframes in the window while still being able to sub-select the set of tracks used for the update. To maintain a bounded and nearly constant computational complexity only a limited number of the available tracks are used for the update (here: 50). We use a heuristic score for this selection based on the distance to the triangulated landmark positions and the disparity angle spanned by all observation rays to the landmark.

Processing odometry tracks: A new landmark is initialized for each odometry track by triangulating its position \( M_l \) (Fig. 4). All keypoint measurements of the track are added to the optimization to form constraints between the observing keyframes and the landmark. We use a weighted re-projection error \( e_o \) that takes the following form in the non-linear least-squares problem:

\[
e_o \left( T_{MB}, M_l \right) = \frac{1}{\sigma_o} \left[ f_{proj} \left( T_{MB}^T \cdot M_l \right) - z_i \right]
\]

(3)

where \( T_{MB} \) is the local pose of the body frame \( B \) expressed in the mission frame of reference \( M \), \( M_l \) is the landmark position expressed in the mission frame \( M \), \( f_{proj} \) is the non-linear function projecting a 3d point in the camera frame onto the image plane, \( z_i \) is the keypoint measurement on the image plane, and \( \sigma_o \) the keypoint measurement uncertainty. Additionally, this error is weighted by a Huber loss-function to achieve robustness against errors in tracking and data association.

Processing localization tracks: Each localization track has an associated map landmark that is used to formulate a re-projection error similar to the one used for the odometry tracks but with a known landmark position \( G_l \). We assume a constant and isotropic measurement noise for these reprojection errors (here: 1.0 px). The error \( e_l \) for one observation can be written as:

\[
e_l \left( T_{GM}, T_{MB} \right) = \frac{1}{\sigma_l} \left[ f_{proj} \left( T_{MB}^T \cdot T_{GM}^T \cdot G_l \right) - z_i \right]
\]

(4)

where \( T_{GM} \) is the transformation that aligns the local mission frame \( M \) with the global map frame \( G \). \( G_l \) is the position of the associated map landmark and all other terms are the same as in Eq. (3). Note that this error term constrains the transformation \( T_{GM} \) and \( T_{MB} \); whereas the error term for odometry observations (Eq. (3)) only constrains the local pose \( T_{MB} \). This allows for a tightly-coupled estimation of the local pose \( T_{MB} \) and the global pose \( T_{GB} \) that is given by \( T_{GB} = T_{GM} \cdot T_{MB} \). Special care must be taken to avoid re-use of information. To this end, we need to make sure that each keypoint measurement and map landmark is not processed more than once in a localization update.

The total cost \( J(x) \) of the non-linear least squares problem
where $N$ denotes the number of keyframes in the window, $n(i)$ the number of cameras for the $i$-th keyframe, $\mathcal{K}(i)$ the set of odometry, $\mathcal{M}(i)$ the set of localization landmarks visible in camera $j$ of keyframe $i$, $\mathbf{e}_{i,j,k}$ the reprojection error of odometry landmark $k$ in camera $j$ of keyframe $i$, analogously $\mathbf{e}_{i,j,m}$ the reprojection error of the localization landmark $m$ and $\mathbf{e}_t$ the temporal IMU error between keyframe $i$ and $i+1$. The terms $\mathbf{W}_{o,j,k}, \mathbf{W}_{i,j,k}, \mathbf{W}_l$ denote the inverse of the corresponding measurement covariance matrices. Processed odometry and localization landmarks get marginalized after each update of the factor graph and keyframes once they are pushed out of the window by new keyframes. After marginalization an additional linear term is introduced to Eq. (5) to account for the influence of the removed states at the time of marginalization.

### D. Handling of Degenerate Motion

Since the estimator processes terminated feature tracks, (quasi-)stationary motions need to be handled explicitly as no or very few tracks terminate during such phases. Such phases are detected by heuristics based on inertial measurements, the output of the 2-pt RANSAC used in the vision frontend and the method from [30]. An artificial zero-velocity measurement is added to the inertial states if rotation-only motion was detected during that time. Additionally, a small number of landmarks (here: 15) remain in the state for multiple estimator updates; whereas all other landmarks are marginalized after each update. These local SLAM landmarks are only marginalized once tracking is lost. This allows for a stable ego-motion estimation during slow motion and transitions to and from stationary phases.

### IV. Experimental Evaluation

We evaluate the presented visual-inertial localization method in a collaborative robot scenario involving a micro aerial vehicle (MAV) and an unmanned ground vehicle (UGV) [31]. Both platforms are equipped with a V-Sensor [32], a sensor system containing an IMU and two global shutter cameras.

For simplicity both platforms were manually controlled to follow a path through an indoor environment while recording visual and inertial data. A motion capture system was used to obtain accurate ground-truth data for position and orientation. The poses from the motion capture system were spatially and temporally aligned with the inertial data of the robots using a full-batch maximum-likelihood estimator instead of the filter formulation of [33].

Additionally, our previous work on map summarization [19] is validated under real-world conditions with a team of heterogeneous robots. Furthermore, the influence of parameters in the map summarization is investigated as a trade-off between map-size and localization accuracy.

#### A. Absolute Localization Error

Here, we evaluate the absolute localization error of the presented visual-inertial localization estimator (VIL) against a related sliding-window localization method (SWL) [20]. The SWL method estimates a map alignment transformation over a sliding-window of recent keyframes and associated map landmarks. The keyframe poses and landmark positions are fixed and only the global map alignment is estimated. The VIO estimates (by running the VIL without a map) are used to forward propagate the state from the most recent localization estimate. Therefore, the SWL is a method that utilizes the same information as the VIL (inertial data, feature tracks and 2d-3d matches) but in a loosely-coupled formulation.

In this scenario, a single indoor MAV mission is used to build a map of the environment. Subsequently, the UGV localizes against this map built by the MAV using its on-board sensors. The global position estimates of both methods are evaluated against the motion tracking ground-truth.

#### B. Influence of Map Summarization

This analysis investigates the influence of the localization map size on the global localization error. The setup aims at
Fig. 7: The absolute localization error (left) w.r.t. ground-truth is shown when localizing a UGV mission against a map built from MAV data. This experiment shows an improvement in smoothness and accuracy of the visual-inertial localization (VIL) achieved by jointly estimating the VIO and global poses whereas the sliding-window localization (SWL) and pure visual-inertial odometry (ODO) yield higher variance and mean error. The global poses (left) are compared against ground-truth from motion tracking (REF).

Fig. 8: Statistics of the absolute localization error for different levels of map summarization. The landmarks are reduced to a varying fraction of the initially 200/000 landmarks. The accuracy and smoothness of the VIL estimates remains almost constant up to a summarization level of 5% whereas the SWL shows a significant increase of the median error starting at 25%.

Simulating a typical collaborative mapping and localization scenario between an MAV and a UGV. First, six local maps are built from independent MAV missions using the visual-inertial localization estimator (without a reference map). In a next step, all local maps are merged into a single global map and refined in a visual-inertial least-squares optimization. The resulting global map is then summarized multiple times with increasing levels of landmark removal. A separate UGV dataset is then used to localize against these summarized maps and evaluate the resulting absolute global localization error w.r.t. the motion tracking ground-truth. This is an ideal scenario to assess, not only the visual-inertial localization performance, but also the concept as a whole as it involves all components: local map creation using VIO, map merging, map summarization and finally localization in the global map using a different robotic platform.

Fig. 8 and Table I show that roughly 90 percent of all landmarks can be pruned from the localization map with only a minimal increase of the localization error when using the proposed visual-inertial localization (VIL) method. The loosely-coupled method (SWL), however, shows a significant increase of the mean and the variance of the localization error at the same summarization levels. This study demonstrates the benefit of the proposed joint estimation of VIO and the global pose when used with highly compact maps. Furthermore, it justifies a more aggressive map summarization for the VIL method and thus allows for more compact localization maps while retaining the robustness and accuracy of larger maps.

C. Timings of the VIL-Estimator

Timings are given in Table II for the UGV dataset of Section [V-B] using a localization map containing 30'000 landmarks. Both methods have been evaluated on an Intel i7-3740QM using 2 cores. The sliding-window localization (SWL) runtime includes matching against the map, global pose estimation and visual-inertial odometry (ODO) to propagate global poses between localizations. The visual-inertial estimator (VIL) contains only map matching and estimation as its tightly-coupled formulation jointly estimates local and global pose. The proposed VIL exhibits very similar runtimes compared to the SWL approach.

<table>
<thead>
<tr>
<th>fraction of initial landmark count</th>
<th>[cm]</th>
<th>VIL</th>
<th>SWL</th>
<th>ODO</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>10</td>
<td>5.6</td>
<td>5.6</td>
<td>-</td>
</tr>
<tr>
<td>25%</td>
<td>20</td>
<td>5.6</td>
<td>5.6</td>
<td>-</td>
</tr>
<tr>
<td>15%</td>
<td>30</td>
<td>5.6</td>
<td>5.6</td>
<td>-</td>
</tr>
<tr>
<td>5%</td>
<td>40</td>
<td>5.6</td>
<td>5.6</td>
<td>-</td>
</tr>
<tr>
<td>2.5%</td>
<td>50</td>
<td>5.6</td>
<td>5.6</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>map query</th>
<th>[ms]</th>
<th>VIL</th>
<th>SWL</th>
<th>ODO</th>
</tr>
</thead>
<tbody>
<tr>
<td>matching</td>
<td>7.38</td>
<td>5.6</td>
<td>5.6</td>
<td>-</td>
</tr>
<tr>
<td>RANSAC</td>
<td>2.53</td>
<td>5.6</td>
<td>5.6</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>estimator</th>
<th>[ms]</th>
<th>VIL</th>
<th>SWL</th>
<th>ODO</th>
</tr>
</thead>
<tbody>
<tr>
<td>localization</td>
<td>10.29</td>
<td>5.6</td>
<td>5.6</td>
<td>-</td>
</tr>
<tr>
<td>odometry</td>
<td>34.72</td>
<td>5.6</td>
<td>5.6</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>complete update</th>
<th>[ms]</th>
<th>VIL</th>
<th>SWL</th>
<th>ODO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>54.92</td>
<td>5.6</td>
<td>5.6</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>processing time for one update profiled on an i7-3740QM using 2 cores. The joint optimization of VIO and localization leads to a lower run-time in the VIL as compared to the loosely-coupled SWL.</th>
<th>[ms]</th>
<th>VIL</th>
<th>SWL</th>
<th>ODO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>46.80</td>
<td>54.92</td>
<td>34.72</td>
</tr>
<tr>
<td></td>
<td>±</td>
<td>9.86</td>
<td>15.59</td>
<td>8.22</td>
</tr>
</tbody>
</table>
V. CONCLUSIONS

In this work, we presented a real-time visual-inertial localization method that is particularly suitable for use with highly compact localization maps. Although, the employed map summarization process guarantees a good localization coverage over the mapped space, it generally reduces the number of potential 2d-3d matches during localization. Commonly, a less smooth global pose estimate is expected if it is solely based on these matches. Therefore, we augment the localization problem (based on 2d-3d map matches) with the additional information from visual feature tracks and inertial measurements. The proposed formulation as a sliding-window based optimization jointly estimates the local VIO and the global pose in the map’s frame of reference. Moreover, it is worth noting that this leads to a seamless switching between localization and exploration mode.

A series of experiments with a team of heterogeneous robots validate the proposed visual-inertial localization (VIL) and the concept of summary maps as a whole. When comparing the VIL against a related method that loosely integrates 6-DoF localizations with VIO poses, we can show that in our method smoothness and accuracy of the global pose estimates remain nearly unaffected up to a high level of summarization. The compared loosely-coupled method, however, does not exhibit the same robustness. Thus, the proposed method can tolerate a more aggressive summarization while still maintaining nearly the same performance.

For future work, we plan to perform experiments at a larger scale and investigate possible benefits arising from using projected landmark uncertainties in the localization.

VI. ACKNOWLEDGEMENTS

The authors thank Samuel Bachmann, Michael Burri, Remo Diethelm and Péter Fankhauser for their help with the experiments and Janosch Nikolic for tools to process the motion capture data. The research leading to these results has received funding from Google’s project Tango.

REFERENCES
