LSFB: A Low-cost and Scalable Framework for Building Large-Scale Localization Benchmark

Haomin Liu *, 1, Mingxuan Jiang 1, Zhuang Zhang 1, Xiaopeng Huang 1, Linsheng Zhao 1, Meng Hang 1, Youji Feng 1, Hujun Bao 2 and Guofeng Zhang * 2

1 SenseTime Research
2 State Key Lab of CAD&CG, Zhejiang University

ABSTRACT
With the rapid development of mobile sensor, network infrastructure and cloud computing, the scale of AR application scenario is expanding from small or medium scale to large-scale environments. Localization in the large-scale environment is a critical demand for the AR applications. Most of the commonly used localization techniques require quite a number of data with groundtruth localization for algorithm benchmarking or model training. The existed groundtruth collection methods can only be used in the outdoors, or require quite expensive equipments or special deployments in the environment, thus are not scalable to large-scale environments or to massively produce a large amount of groundtruth data. In this work, we propose LSFB, a novel low-cost and scalable framework to build localization benchmark in large-scale environments with groundtruth poses. The key is to build an accurate HD map of the environment. For each visual-inertial sequence captured in it, the groundtruth poses are obtained by joint optimization taking both the HD map and visual-inertial constraints. The experiments demonstrate the obtained groundtruth poses are accurate enough for AR applications. We use the proposed method to collect a dataset of both mobile phones and AR glass exploring in large-scale environments, and will release the dataset as a new localization benchmark for AR.

* e-mails: liuhaomin@sensetime.com, zhangguofeng@cad.zju.edu.cn.

Index Terms: Human-centered computing—Human Computer Interaction—Interaction Paradigms—Mixed / Augmented Reality; Computing Methodologies—Artificial Intelligence—Computer Vision—Tracking

1 INTRODUCTION
Localization is the core to augmented reality (AR). One of the most popular techniques is visual-inertial SLAM (VI-SLAM), that uses the complementary visual and inertial measurements to localize the mobile device in the unknown environment [27, 30]. The state-of-the-art VI-SLAM has achieved great accuracy and robustness in small and medium scale environments [7], and has been integrated into many AR products like Apple ARKit, Google ARCore and Microsoft Hololens etc. With the rapid development of mobile sensors, network infrastructure and cloud computing, the scale of AR application scenarios is also expanding, from medium-scale rooms to city-scale environments. This trend has received unprecedented attention of both academics and industry, but also posed unprecedented challenges on the accuracy and robustness [3]. To meet the challenges, it is important to have an appropriate benchmark containing a large amount of data with groundtruth localization in the large-scale environments. Existed groundtruth collection methods either require GPS that is not available indoors [11], or require quite expensive equipments [2, 16] or special deployments [5, 28] in the environment, not scalable to large-scale environments or to massively produce a large amount of groundtruth data.

In this work, we propose LSFB, a novel Low-cost and Scalable Framework to build localization Benchmark in large-scale environ-
ments with groundtruth poses. As shown in Fig. 1, our framework is based on building an accurate HD map of the environment. For each visual-inertial sequence captured in the environment, the groundtruth poses are obtained by joint optimization taking both the HD map and visual-inertial constraints. Compared to the existed groundtruth collection methods, our solution is the only one with a combination of attributes including:
- low equipment cost
- low operation cost
- suitable to large environment
- suitable for both indoor and outdoor
- suitable for AR

Detailed comparison is listed in Table 1, and explained in Sect. 2.1.

2 RELATED WORK

In this section, we first review and compare the existed groundtruth collection methods, then review the localization techniques this work focuses on.

2.1 Groundtruth Solution

The very popular KITTI [11] uses a high precision GPS/IMU to produce 6 DOF groundtruth poses in large-scale outdoor environments. To extend to indoor area where GPS is not available, Rawseeds [5], NCLT [4] and Complex Urban [15] propose to use fiducial marker or Lidar scan matching as additional pose constraint in the indoor area. In Rawseeds, the fiducial markers are attached to the moving robot, observed by an external network of pre-calibrated cameras and laser scanner. These external cameras and laser scanners increase the equipment and operation cost which limits the scalability. Opposite to this outside-in solution, both NCLT [4] and Complex Urban [15] use the inside-out measurement of Lidar scan matching, alleviating the cost and scalability limitation. However, they require high precision Lidar on the mobile device, which is infeasible for AR devices. In addition, all the above mentioned methods collect data by a wheeled ground vehicle/robot, whose motion style is quite different from to the handheld/wearable AR devices. By contrast, PennCOSYVIO [28] and ADVIO [6] use handheld VI-sensor as the localized device, thus are suitable for AR and the equipment cost is low. But they respectively require manually measured fiducial markers placed in the environment [28], and manually labeled position to fixes on the floor plan [6], to provide the groundtruth position constraints, thus the operation cost is high and the scalability is limited. EuRoC [2] and ZJU-SenseTime [16] use a VICON motion capture system, which is comprised of VICON trackers deployed in the environment and VICON markers attached to the localized device. Both equipment and deployment cost of VICON trackers are prohibitively high (~$30K dollar for a 20 m² room) to scale to a large environments or outdoors. Aachen Day-Night and CMU Seasons proposed in [32] are two commonly used datasets for large-scale visual localization. Groundtruth is obtained by the Structure-from-Motion (SfM) technique [34]. Considering day-night and across seasons challenges, manually labeled correspondences are required, which increases the operation cost, and the images are not sequential thus not suitable for AR.

Compared to the existed method, LSFB proposed in this work does not require any special deployment in the environments or special equipment on the localized device except a VI-sensor with synchronized camera and IMU, which is commonly available on the current AR device (e.g. mobile phones and AR glasses), thus requires low equipment and operation cost, and suitable for both indoor and outdoor environment for the AR application. The only requirement is that we assume the HD map of the environment has been accurately reconstructed. The groundtruth trajectory for each visual-inertial sequence captured in the environment are obtained by joint optimization taking both the global localization constraint from the HD map and the temporal motion constraint from visual-inertial measurements. Thanks to the recently maturing large-scale 3D reconstruction techniques [13, 41] and the efficient optimization techniques [17, 22], our method is scalable to large-scale environments with a large amount of data.

2.2 Localization

Simultaneous Localization and Mapping (SLAM) is probably the most commonly used localization technique for AR. PTAM [19] is the first visual SLAM (V-SLAM) system that is able to track the camera motion in real-time on mobile devices. Bundle Adjustment (BA) [38] that jointly optimizes keyframe poses and 3D points is performed in a background thread to ensure accuracy without jeopardizing real-time performance. This framework is used by most of the recent V-SLAM systems [23, 26, 37]. Visual-inertial SLAM (VI-SLAM) combines complementary visual and inertial measurements to achieve better accuracy and robustness than V-SLAM, and becomes a standard technique for AR applications. The most successful systems include MSDKF [25], OKVIS [21], VINS [30] etc. Jeffrey Delmerico and Davide Scaramuzza give a detailed evaluation of the state-of-the-art VI-SLAM systems [7].

Visual localization is a promising technique to obtain the global localization in HD map. Traditional methods [14, 31, 33] mainly aim to improve the efficiency and accuracy of the 2D-3D correspondence search. They either use image retrieval [36] to find the correspondence in a coarse-to-fine manner, or employ quantization method [29] to perform direct feature matching. Recent approaches resort to learning-based features (e.g. SuperPoint [8], D2-Net [9], etc) to improve the robustness against view point changes, motion blur, and illumination variations, etc. Localization can also be estimated in an end-to-end manner (e.g. pose-net [18], map-net [1] etc) thanks to the strong capacity of deep neural networks. The main limitation is the generalization issue. The poses would be terribly wrong in some viewpoints that are not covered by the training set. A large amount of training data with groundtruth localization
is required to make these methods work.

3 Method

The framework of the proposed method is illustrated in Fig. 2. We developed a backpack 3D scanner comprised of a Velodyne Ultra Puck VLP-3DC 3D Lidar and a customized Mynt VI-sensor to capture the Lidar point cloud and 360° images of the environment, from which the HD map of the environment is reconstructed (Sect. 3.1). The HD map is used by all the localized devices with VI-sensor (e.g. mobile phone and AR glass) moving in the same environment. Each visual-inertial sequence is localized in the HD map, within which each frame associated with a groundtruth pose (Sect. 3.2).

3.1 HD Map Reconstruction

As shown in the upperleft picture of Fig. 2, a 3D Lidar is placed on top of a fixed structural part containing an array of 4 cameras capturing the 360° images of the environment and an IMU. The camera specification is listed in Table 2. All sensors are hardware synchronized and carefully calibrated.

To guarantee the completeness and accuracy of the final reconstruction, we use the Lidar and IMU measurements to run 3D Lidar SLAM that performs dense mapping in real-time. The reconstruction result is displayed in the handheld screen, guiding the operator to perform a complete and algorithm-friendly scanning. We select Google Cartographer [13] for its robustness, convenience and the ability to close large loops in real-time. Cartographer uses IMU measurements to project Lidar points to the latest Lidar frame at time $i$ (called undistortion) and estimates the optimal Lidar pose $W T_i$ that best aligns the undistorted point cloud in the Lidar coordinate $L_i$ to the global map coordinate $W$. The stream of 360° images, the undistorted Lidar point clouds and the Lidar poses are stored for the final offline reconstruction. The goal is to reconstruct a set of 3D points with the image feature descriptors for the visual localization. To this goal, SIFT features [24] are extracted and matched from 360° images. If the camera poses are known, the 3D position of each feature point can be obtained by triangulation [12]. The camera pose $W T_C$ for each image $i$ could be converted from the Lidar pose by $W T_C = W T_i L T_C$, where $L T_C$ is the pre-calibrated extrinsic between Lidar and camera. However, $W T_L$ from the online Cartographer is inevitably suboptimal. Directly use this pose would result in a very noisy HD map. Existed method for Lidar-based reconstruction [39] mainly focus on improving the consistency between individual Lidar point clouds, rather than improving the consistency between visual-Lidar measurements to obtain an optimal visual map. We propose a novel method taking both the visual and Lidar constraints to obtain this goal.

Given the undistorted Lidar point clouds, we first use LOAM [40] to extract plane and edge features and obtain a set of plane/edge correspondences, denoted as $L_{p} = \{l_{p}^{i} = \{p_{i}^{m}, q_{i}^{m}\}\}$ and $L_{e} = \{l_{e}^{i} = \{p_{i}^{q}, m_{i}\}\}$ respectively, where $p_{i}$ is the feature point in the source point cloud $m_{i}$, $q$ the matched point in the target point could $n_{i}$ with the normal vector $m_{i}$ for plane feature and the direction vector $q_{i}$ for edge feature. The Lidar constraint $L_{p}$ and $L_{e}$ are combined with the set of image feature correspondences to jointly optimize the camera poses $W T_i$ and 3D point $X_j$ of the visual map.

\[
\begin{align*}
\arg\min_{W T_i, X_j} \sum_{l_{p} \in L_{p}} \|n_i^T (l_{p} W T_i p - q_{i}^m)\|^2_2 + \sum_{l_{e} \in L_{e}} \|n_i \times (l_{e} W T_i p - q_{i}^m)\|^2_2 + \sum_{(X_j, X_i)} \|f(W T_i X_j) - X_i\|^2_2 \\
(1)
\end{align*}
\]

where the 3 residuals are Lidar point-to-plane / point-to-edge distance error and image feature reprojection error. $f(T X)$ projects a 3D point $X$ to the image plane by camera pose $T$. $X_i$ is the 2D feature point of the $X_j$ in image $i$. $\Sigma_p$, $\Sigma_e$ and $\Sigma_r$ are the covariance matrices for plane, edge and image feature constraint, which is set to be 0.2 m², 0.5 m² and 2 pixel² respectively. (1) is solved iteratively using the LM algorithm. The reconstructed visual map is visualized in Fig. 3. It can be observed that directly using the initial pose results in very noisy visual map. With the proposed visual-Lidar optimization, the visual map is more accurate.

3.2 Groundtruth Localization

For each input frame, visual localization is performed to find a set of feature correspondences between the 2D feature point in the input image and the 3D feature point in the HD map. We call them global feature correspondences/tracks. The camera pose can be estimated from these 2D-3D correspondences by PnP [20]. However, the result of PnP highly relies on the number of feature correspondences and their distribution in the image. It is inevitable that there are viewpoints that are not covered by the HD map, leading to localization failure in the uncovered area. In addition, the pose obtained by PnP frequently jitters due to lack of temporal constraints, thus is not suitable for AR with strict requirements for smoothness. We propose to use the visual-inertial measurements in the localized sequence as the temporal constraint. To this purpose, we extract and match features for the input image sequence. We call them lo-
Figure 4: The trajectories in the VICON room for (a) iPhone, (b) Android phone and (c) AR glass.

Table 2: Specification of cameras

<table>
<thead>
<tr>
<th>AR device</th>
<th>Backpack 3D scanner</th>
<th>iPhone</th>
<th>AN. phone</th>
<th>AR glass</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Cam.</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Res.</td>
<td>1280×800</td>
<td>640×480</td>
<td>640×480</td>
<td>640×400</td>
</tr>
<tr>
<td>FoV</td>
<td>106°×65°</td>
<td>68°×54°</td>
<td>66°×52°</td>
<td>97°×70°</td>
</tr>
<tr>
<td>Freq.</td>
<td>20 Hz</td>
<td>30 Hz</td>
<td>30 Hz</td>
<td>30 Hz</td>
</tr>
</tbody>
</table>

Table 3: The RMSE comparison

<table>
<thead>
<tr>
<th>AR device</th>
<th>LSFB</th>
<th>LSFB w/Pct. of Vis. Loc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>iPhone</td>
<td>0.037</td>
<td>0.068(65.08%) 0.039 0.036 0.035 0.129</td>
</tr>
<tr>
<td>AN. phone</td>
<td>0.046</td>
<td>0.112(97.55%) 0.044 0.044 0.025 0.104</td>
</tr>
<tr>
<td>AR glass</td>
<td>0.044</td>
<td>0.088(98.43%) 0.037 0.033 0.039 0.133</td>
</tr>
</tbody>
</table>

4 Evaluation

In this section, we first conduct experiments to analyze the accuracy of the proposed method quantitatively, then give the qualitative result of the collected groundtruth dataset which will be released as a new localization benchmark for AR. We use the backpack 3D scanner to reconstruct the HD map of the environments, and use 3 AR devices to collect the visual-inertial data in the environment. The 3 AR devices are: an iPhone XS Max, an Android phone Huawei Mate 30 Pro, and a Shadow Creator AR glass Action One Pro. The camera specifications for the backpack 3D scanner and the AR devices are listed in Table 2.

4.1 Quantitative Result

To the best of our knowledge, there is no other way to evaluate the localization accuracy in large-scale environments for the mobile phones and AR glass. We can only conduct quantitative evaluation in a small room (∼20 m²), using VICON to obtain the localization with millimeter accuracy as the baseline to evaluate our method.

For each of the 3 AR devices, we walked around the room 3 times at the speed of ∼1 m/s. The camera poses estimated by LSFB are compared to that measured by VICON. The translational RMSE (Root Mean Square Error) is calculated for each sequence, and listed in Table 3. We can see that LSFB achieves centimeter-level accuracy for all devices. We also evaluate the camera poses obtained by visual localization. The RMSE are listed in the next column, and the trajectories are shown in Fig. 4. As expected, the poses from visual localization is erroneous and frequently jitters, which is not suitable for AR which has strict requirements for the localization smoothness. We also count the success rate of visual localization in the bracket, which is defined as the percentage of frames that have matched more than 30 features in the HD map. The success rate is high for Android phone and AR glass, but lower for iPhone because the focus is locked, resulting in severe image blur. By the proposed LSFB, all the poses can be accurately recov-
It is easy to fully cover the small room by the 360° images, but not the case for large-scale environments. We conduct an experiment to investigate the effect of visual localization failure. We randomly select a portion of frames and remove all global feature correspondences on them, resulting 60%, 40% and 20% of visual localization. We perform LSFB using the remaining global feature correspondences, and list the RMSE in the last 3 columns in Table 3. We can see that the accuracy is not sensitive to visual localization failure, sometimes even better after removing the erroneous global feature correspondences.

In the last column, we also evaluate the result without HD map (0%), which is actually the traditional VSLAM. The RMSE is several times larger than LSFB, verifying the effectiveness of incorporating the HD map.

4.2 Qualitative Result

We collect groundtruth data in 3 environments as the LSFB benchmark for AR. The 3 environments include an indoor office ~500 m², an outdoor environment around a building and square ~10,000 m² in total, and an environment containing both indoor office building and outdoor garden ~10,000 m² in total. We call them “indoor”, “outdoor” and “indoor-outdoor” respectively. We walked around the environments to collect a set of visual-inertial data using mobile phones and AR glass. We use the proposed LSFB to reconstruct the HD map and localize each sequence in it, and overlay the result together. The results are shown in Fig. 5(a)(b), and Fig. 1 for the 3 environment respectively. The trajectories are globally consistent thanks to the global feature constraints, and temporally smooth thanks to the local visual-inertial constraints.

Since the LSFB is designed for AR, we also show the AR effect using the recovered camera poses. Specifically, for each image, we overlay the future trajectory on the image, as a sequence of spheres sampled every 5 frames in the original sequence. Snapshots are shown in Fig. 6. See the supplementary video for the whole sequence. We can see that the result are smooth and accurate enough for AR applications.

5 Conclusion

In this work, we present LSFB, a novel groundtruth collection method for localization in large-scale environments. To the best of our knowledge, it is the first groundtruth framework with low-cost and the scalability to collect a large amount of data in the large-scale environments. The experiments demonstrate the obtained groundtruth poses are accurate enough for AR applications. We use the proposed method to collect a dataset of both mobile phones and AR glass exploring in large-scale environments, and will release the dataset as a new localization benchmark for AR.

The current solution relies on the precision of visual localization. If the visual localization module returns too many false positive results, the visual-inertial optimization would diverge. Another limitation is the requirement of people to carry a heavy backpack to collect HD maps, so the size of current maps is only building-scale. For the future work, we will address the failure cases, and setup the 3D scanner to a vehicle to enable city-scale HD map reconstruction before collecting the city-scale localization data to further expand our benchmark. We hope that would truly promote the localization techniques and the realization of AR applications in the city-scale environments.

Acknowledgments

This work was partially supported by NSF of China (No. 61672457), and the Fundamental Research Funds for the Central Universities (No. 2019XZZX004-09).

References


