Scale Drift Correction of Camera Geo-Localization using Geo-Tagged Images

Kazuya Iwami[1] [0000−0001−7012−5616], Satoshi Ikehata[2] [0000−0002−6061−7956], and Kiyoharu Aizawa[1] [0000−0003−2146−6275]

[1] The University of Tokyo, Japan
{iwami,aizawa}@hal.t.u-tokyo.ac.jp
[2] National Institute of Informatics, Japan
sikehata@nii.ac.jp

Abstract. Camera geo-localization from a monocular video is a fundamental task for video analysis and autonomous navigation. Although 3D reconstruction is a key technique to obtain camera poses, monocular 3D reconstruction in a large environment tends to result in the accumulation of errors in rotation, translation, and especially in scale: a problem known as scale drift. To overcome these errors, we propose a novel framework that integrates incremental structure from motion (SfM) and a scale drift correction method utilizing geo-tagged images, such as those provided by Google Street View. Our correction method begins by obtaining sparse 6-DoF correspondences between the reconstructed 3D map coordinate system and the world coordinate system, by using geo-tagged images. Then, it corrects scale drift by applying pose graph optimization over Sim(3) constraints and bundle adjustment. Experimental evaluations on large-scale datasets show that the proposed framework not only sufficiently corrects scale drift, but also achieves accurate geo-localization in a kilometer-scale environment.

Keywords: 3D Reconstruction · Localization · Street View

1 Introduction

Camera geo-localization from a monocular video in a kilometer-scale environment is a essential technology for AR, video analysis, and autonomous navigation. To achieve accurate geo-localization, 3D reconstruction from a video is a key technique. Incremental structure from motion (SfM) and visual simultaneous localization and mapping (visual SLAM) achieve large-scale 3D reconstructions by simultaneously localizing camera poses with six degrees-of-freedom (6-DoF) and reconstructing a 3D environment map [7, 19].

Unlike for a stereo camera, an absolute scale of the real world cannot be derived using a single observation from a monocular camera. Although it is possible to estimate an environment's relative scale from a series of monocular observations, errors in the relative scale estimation accumulate over time, and this is referred to as scale drift [6, 22].

For an accurate geo-localization not affected by scale drift ,prior information in a geographic information system (GIS) has been utilized in previous studies. For example, point clouds, 3D models, building footprints, and road maps have been proven to be efficient for correcting reconstructed 3D maps [18, 5, 23, 24, 4]. However, these priors are only available in limited situations, e.g., in an area that is observed in advance, or in an environment consisting of simply-shaped buildings. Therefore, there is a good chance that other GIS information can help to extend the area in which a 3D map can be corrected.

Hence, in this paper, motivated by the recent availability of massive public repositories of geo-tagged images taken all over the world, we propose a novel framework for correcting the scale drift of monocular 3D reconstruction by utilizing geo-tagged images, such as those in Google Street View [1], and achieve accurate camera geo-localization. Owing to the high coverage of Google Street View, our proposal is more scalable than those in previous studies.

The proposed framework integrates incremental SfM and a scale drift correction method utilizing geo-tagged images. Our correction method begins by computing 6-DoF correspondences between the reconstructed 3D map coordinate system and the world coordinate system, by using geo-tagged images. Owing to significant differences in illumination, viewpoint, and the environment resulting from differences in time, it tends to be difficult to acquire correspondences between video frames and geo-tagged images (Fig. 2). Therefore, a new correction method that can deal with the large scale drift of a 3D map using a limited number of correspondences is required. Bundle adjustment with constraints of global position information, which represents one of the most important correction methods, cannot be applied directly. This is because bundle adjustment tends to get stuck in a local minimum when starting from a 3D map including large errors [22]. Hence, the proposed correction method consists of two coarse-to-fine steps: pose graph optimization over Sim(3) constraints, and bundle adjustment. In these steps, our key idea is to extend the pose graph optimization method proposed for the loop closure technique of monocular SLAM [22], such that it incorporates the correspondences between the 3D map coordinate system and the world coordinate system. This step corrects the large errors, and enables bundle adjustment to obtain precise results. After implementing this framework, we conducted experiments to evaluate the proposal.

The contributions of this work are as follows. First, we propose a novel framework for camera geo-localization that can correct scale drift by utilizing geo-tagged images. Second, we extend the pose graph optimization approach to dealing with scale drift using a limited number of correspondences to geo-tags. Finally, we validate the effectiveness of the proposal through experimental evaluations on kilometer-scale datasets.

2 Related Work

2.1 Monocular 3D Reconstruction

Incremental SfM and visual SLAM are important approaches to reconstructing 3D maps from monocular videos. Klein *et al.* proposed PTAM for small AR workspaces [11]. Mur-Artal *et al.* developed ORB-SLAM, which can reconstruct large-scale outdoor environments [19]. For accurate 3D reconstruction, the loop closure technique has commonly been employed in recent SLAM approaches [22, 19]. Loop closure deals with errors that accumulate between two camera poses that occur at the same location, i.e., when the camera trajectory forms a loop. Lu and Milios [16] formulated this technique as a pose graph optimization problem, and Strasdat *et al.* [22] extended pose graph optimization to deal with scale drift for monocular visual SLAM. It is certain that loop closure can significantly improve 3D maps, but this is only effective if a loop exists in the video.

2.2 Geo-registration of Reconstructions

Correcting reconstructed 3D maps by using geo-referenced information has been regarded as a geo-registration problem. Kaminsky *et al.* proposed a method that aligns 3D reconstructions to 2D aerial images [10]. Wendel *et al.* used an overhead digital surface model (DSM) for the geo-registration of 3D maps [26]. Similar to our work, Wang *et al.* used Google Street View geo-tagged images and a Google Earth 3D model for the geo-registration of reconstructed 3D maps [25]. However, because all these methods focus on estimating a best-fitting similarity transformation to geo-referenced information, they only correct the global scale in terms of 3D map correction.

Methods for geo-registration using non-linear transformations have also been proposed. To integrate GPS information, Lhuillier *et al.* proposed incremental SfM using bundle adjustment with constraints from GPS [14], and Rehder *et al.* formulated a global pose estimation problem using stereo visual odometry, inertial measurements, and infrequent GPS information as a 6-DoF pose graph optimization problem [20]. In terms of correcting camera poses using sparse global information, Rehder's method is similar to our pose graph optimization approach. However, our 7-DoF pose graph optimization differs in focusing on scale drift resulting from monocular 3D reconstruction, and in utilizing geo-tagged images. In addition to GPS information, various kinds of reference data have been used for the non-linear geo-registration or geo-localization of a video, such as point clouds [18, 5], 3D models [23], building footprints [24], and road maps [4]. In this paper, we address a method that introduces geo-tagged images to the non-linear geo-registration of 3D maps.

3 Proposed Method

Fig. 1 provides a flowchart of the proposed framework, which is roughly divided into three parts. The first part is incremental SfM, and is described in Sec. 3.2.

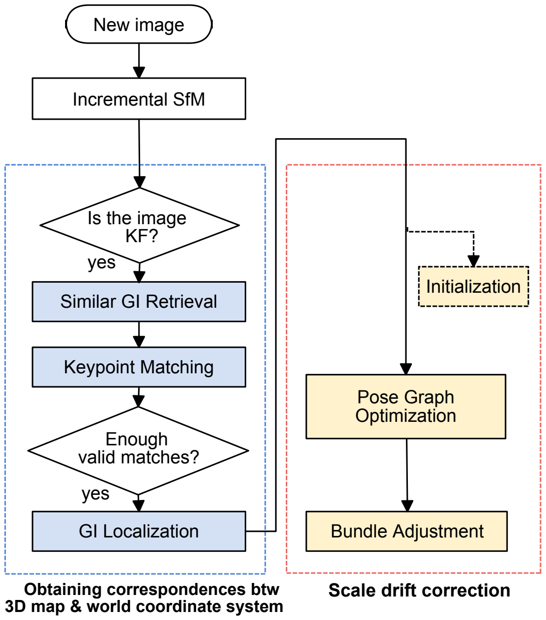


Fig. 1. A flowchart of our proposal. KF and GI denote a keyframe and geo-tagged image, respectively. Initialization is performed only once in a whole reconstruction.

The second part computes 6-DoF correspondences between the 3D map coordinate system and the world coordinate system (as defined below), by making use of geo-tagged images (Sec. 3.3). The third part then uses the correspondences to correct the scale drift of the 3D map, by applying pose graph optimization over Sim(3) constraints (Sec. 3.5) and bundle adjustment (Sec. 3.6) incrementally. The initialization of the scale drift correction method is described in Sec. 3.4.

3.1 World Coordinate System

In this paper, the world coordinates are represented by 3D coordinates (x, y, z), where the xz-plane corresponds to the Universal Transverse Mercator (UTM) coordinate system, which is an orthogonal coordinate system using meters, and y corresponds to the height from the ground in meters. The UTM coordinates can be converted into latitude and longitude if necessary.

3.2 Incremental SfM

As large-scale incremental SfM, we use ORB-SLAM [19] (with no real-time constraints). This is one of the best-performing monocular SLAM systems. Frames that are important for 3D reconstruction are selected as keyframes by ORB-SLAM. Every time a new keyframe is selected, our correction method is performed, and the 3D map reconstructed up to that point is corrected. In the 3D

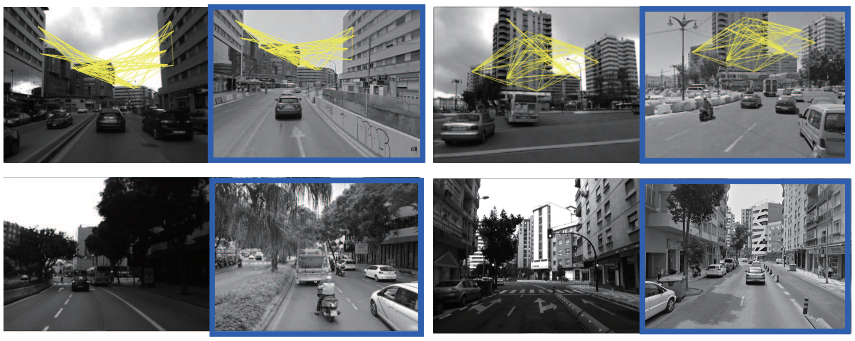


Fig. 2. Examples of keypoint matches between keyframes (without blue squares) and geo-tagged images of Google Street View (with blue squares) after kVLD validation. Top: pairs of images where valid matches are found. Yellow lines denote kVLD graph structures, which are composed of inliers. Bottom: rejected pairs of images where a sufficient number of matches is not found because of differences in illumination, viewpoint, and environment, despite being taken in almost the same location.

reconstruction, we identify 3D map points and their corresponding 2D keypoints in the keyframes (collectively denoted by $C_{\text{map-kf}}$).

Our proposed framework does not depend on a certain 3D reconstruction method, and can be applied to the other monocular 3D reconstruction methods, such as incremental SfM and feature-based visual SLAM.

3.3 Obtaining Correspondences between 3D Map and World Coordinates

Here, we describe the second part of the proposed method, which uses geo-tagged images to compute a 6-DoF correspondence, $C_{\text{map-world}}$, between the 3D map and world coordinate system. For this purpose, we modify Agarwal's method [2] to integrate it into ORB-SLAM. This part consists of the following four steps: geo-tagged image collection, similar geo-tagged image retrieval, keypoint matching, and geo-tagged image localization.

Geo-tagged Image Collection. Google Street View [1] is a browsable street-level GIS, which is one of the largest repositories of global geo-tagged images (i.e., images and their associated geo-tags). All images are high-resolution RGB panorama images, containing highly accurate world positions [12]. We make use of this data by converting each panorama image into eight rectilinear images with the same field-of-view as our input video, with eight horizontal directions. Note that because each geo-tag has a position and rotation in the world coordinates, we can obtain the 6-DoF correspondences between the 3D map coordinate system and world coordinate system if geo-tagged images are localized in the 3D map coordinate system.

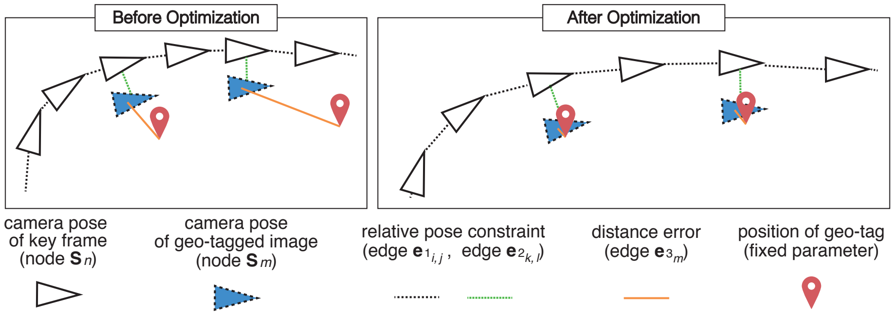

Before Optimization
After Optimization
camera pose of key frame (node S n)
camera pose of geo-tagged image (node S m)
relative pose constraint (edge e1i,j , edge e2k,l)
distance error (edge e3m)
position of geo-tag (fixed parameter)

Fig. 3. An example of the proposed pose graph optimization. This optimization maintains overall relative poses, except for gradual scale changes, and keeps camera poses of geo-tagged images close to the positions of the corresponding geo-tags.

Similar Geo-tagged Image Retrieval. When a new keyframe is selected, we retrieve the top-k similar geo-tagged images. The retrieval system employs a bag-of-words approach based on SIFT descriptors [2].

Keypoint Matching. Given the pairs of keyframes and retrieved geo-tagged images, we detect ORB keypoints [21] from the pairs and perform keypoint matching. Because the matching between video frames and Google Street View images tends to include many outliers [17], we use a virtual line descriptor (kVLD) [15], which can reject outliers by using a graph matching method even when inlier rate is around 10%. All pairs of keyframes and geo-tagged images that have fewer than five keypoint matches are also rejected. Fig. 2 shows examples of valid and rejected pairs. The rejected examples indicate that there are sometimes significant visual differences between an input video frame and a geo-tagged image if they were taken at almost the same location.

Geo-tagged Image Localization. To compute $C_{\text{map-world}}$, we first compute 3D-to-2D correspondences $C_{\text{map-geo}}$ between 3D map points and their corresponding 2D keypoints in geo-tagged images. In particular, we obtain $C_{\text{map-geo}}$ by combining the 2D keypoint matches (computed in the previous step) with the correspondences $C_{\text{map-kf}}$ between 3D map points and their corresponding 2D keypoints in keyframes (computed in 3D reconstruction). Then, we obtain the 6-DoF camera poses of geo-tagged images in the 3D map coordinate system by minimizing the re-projection errors of $C_{\text{map-geo}}$, using the LM algorithm. Finally, we obtain $C_{\text{map-world}}$ by combining the camera poses of geo-tagged images and 6-DoF camera poses of the associated geo-tags.

3.4 Initialization (INIT)

As the initialization, two kinds of linear transformations are performed on the 3D map, because the positions and scales of the 3D map coordinates and world coordinates are significantly different. Initialization is applied once, when the i-th geo-tagged image is localized. We set $i = 4$.

Given the first to i-th $C_{\text{map-world}}$, the first transformation assumes that all camera poses are approximately located in one plane, and rotates the 3D map to align that plane to the world xz-plane. The best-fitting plane can be estimated by a principal component analysis.

Next, we estimate the best-fitting transformation matrix given by Eq. 1, which transforms a point in the 3D map coordinate system $\mathbf{p}_{\text{SLAM},k}$ to be closer to a corresponding point in the world coordinate system $\mathbf{p}_{world,k}$ ($\mathbf{p}_{\text{SLAM},k}$ and $\mathbf{p}_{world,k}$ are denoted using a homogeneous representation):

$$\mathbf{A} = \begin{bmatrix} s * \cos(\theta) & 0 & -s * \sin(\theta) & a \\ 0 & s & 0 & 1 \\ s * \sin(\theta) & 0 & s * \cos(\theta) & b \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1}$$

Using the first to i-th $C_{\text{map-world}}$, we estimate the four matrix parameters $[a, b, s, \theta]$ by minimizing the following cost using RANSAC [8] and the Levenberg-Marquart (LM) algorithm:

$$E = \sum_{k \in 1,2 \ldots i} \|\mathbf{p}_{world,k} - \mathbf{A}\mathbf{p}_{\text{SLAM},k}\|^2 \tag{2}$$

The camera poses of the geo-tagged images in $C_{\text{map-world}}$, keyframes, and 3D map point can then be transformed using the resulting matrix.

3.5 Pose Graph Optimization over Sim(3) Constraints (PGO)

We correct the 3D map focusing on scale drift by using the newest three of $C_{\text{map-world}}$. This correction is performed every time a new $C_{\text{map-world}}$ is found after initialization. Then, we propose a graph-based non-linear optimization method (pose graph optimization) on Lie manifolds, which simultaneously corrects the scale drift and aligns the 3D map with the world coordinates.

Notation. A 3D rigid body transformation $\mathbf{G} \in \mathrm{SE}(3)$ and a 3D similarity transformation $\mathbf{S} \in \mathrm{Sim}(3)$ are defined by Eq. 3, where $\mathbf{R} \in \mathrm{SO}(3)$, $\mathbf{t} \in \mathbb{R}^3$, and $s \in \mathbb{R}^+$. Here, SO(3), SE(3), and Sim(3) are Lie groups, and $\mathfrak{so}(3)$, $\mathfrak{se}(3)$, and $\mathfrak{sim}(3)$ are their corresponding Lie algebras. A Lie group can be transformed into a Lie algebra using its exponential map, and the inverse transformation is defined by the inverse logarithm map. Each Lie algebra is represented by a vector of its coefficients. For example, $\mathfrak{sim}(3)$ is represented as the seven-vector $\boldsymbol{\xi} = (\omega_1, \omega_2, \omega_3, \sigma, \nu_1, \nu_2, \nu_3)^{\mathrm{T}} = (\boldsymbol{\omega}, \sigma, \boldsymbol{\nu})^{\mathrm{T}}$, and the exponential map $\exp_{\mathrm{Sim}(3)}$ and logarithm map $\log_{\mathrm{Sim}(3)}$ are defined as in Eq. 4 and Eq. 5, respectively, where $\mathbf{W}$ is a term similar to Rodriguez's formula. Further details of Sim(3) are given in [22].

$$\mathbf{G} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix} \qquad \mathbf{S} = \begin{bmatrix} s\mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix} \tag{3}$$

$$\exp_{\mathrm{Sim(3)}}(\boldsymbol{\xi}) = \begin{bmatrix} e^{\sigma} \exp_{\mathrm{SO(3)}}(\boldsymbol{\omega}) & \mathbf{W}\boldsymbol{\nu} \\ \mathbf{0} & 1 \end{bmatrix} = \mathbf{S} \tag{4}$$

$$\log_{\mathrm{Sim(3)}}(\mathbf{S}) = {\exp_{\mathrm{Sim(3)}}}^{-1}(\mathbf{S}) = \boldsymbol{\xi} \tag{5}$$

Proposed pose graph optimization. In a general pose graph optimization approach [16, 20], camera poses and relative transformations between two camera poses are represented as elements of SE(3). However, in our approach, 6-DoF camera poses and relative transformations are converted into 7-DoF camera poses, represented by elements of Sim(3). This is achieved by leaving the rotation R and translation $\mathbf{t}$ of a camera pose unchanged, and setting the scale s to 1. The idea that camera poses and relative pose constraints can be handled in Sim(3) was proposed by Strasdat *et al.* [22], for dealing with the scale drift problem in monocular SLAM. In this paper, we introduce 7-DoF pose graph optimization, which has previously only been used in the context of loop closure, to correct 3D reconstruction by utilizing sparse correspondences between two coordinate systems. Our pose graph contains two kinds of nodes and three kinds of edges, as follows (see Fig. 3):

- Node $\mathbf{S}_n \in \mathrm{Sim(3)}$, where $n \in C_1$: the camera pose of the n^{th} keyframe.
- Node $\mathbf{S}_m \in \mathrm{Sim(3)}$, where $m \in C_2$: the camera pose of the m^{th} geo-tagged image.
- Edge $\mathbf{e}_{1_{i,j}}$, where $(i, j) \in C_3$: the relative pose constraint between the i^{th} and j^{th} keyframes. (Eq. 6)
- Edge $\mathbf{e}_{2_{k,l}}$, where $(k, l) \in C_4$: the relative pose constraint between the k^{th} keyframe and the l^{th} geo-tagged image. (Eq. 7)
- Edge $\mathbf{e}_{3_m}$, where $m \in C_2$: the distance error between the position of the m^{th} geo-tagged image and the world position $\mathbf{y}_m$ of the corresponding geo-tag. (Eq. 8)

$$\mathbf{e}_{1_{i,j}} = \log_{\mathrm{Sim(3)}}(\mathit{\Delta}\mathbf{S}_{i,j} \cdot \mathbf{S}_i \cdot \mathbf{S}_j^{-1}) \in \mathbb{R}^7 \tag{6}$$

$$\mathbf{e}_{2_{k,l}} = \log_{\mathrm{Sim(3)}}(\mathit{\Delta}\mathbf{S}_{k,l} \cdot \mathbf{S}_k \cdot \mathbf{S}_l^{-1}) \in \mathbb{R}^7 \tag{7}$$

$$\mathbf{e}_{3_m} = \mathrm{trans}(\mathbf{S}_m) - \mathbf{y}_m \in \mathbb{R}^3 \tag{8}$$

where $\mathrm{trans}(\mathbf{S}) \equiv (\mathbf{S}_{1,4}, \mathbf{S}_{2,4}, \mathbf{S}_{3,4})^{\mathrm{T}}$. Here, N is the total number of keyframes, and M is the total number of geo-tagged images that have correspondences to keyframes. The set C_1 contains all the keyframes positioned between the two that have the newest and the third newest $C_{\text{map-world}}$. The set C_2 contains the newest three of $C_{\text{map-world}}$. The set C_3 contains the pairs of keyframes that observe the same 3D map point in 3D reconstruction, and C_4 contains pairs of keyframes and their corresponding geo-tagged images. Finally, $\varDelta\mathbf{S}_{i,j}$ is the converted Sim(3) relative transformation between $\mathbf{S}_i$ and $\mathbf{S}_j$, which is calculated before the optimization and remains fixed during the optimization.

Note that we newly introduced the nodes $\mathbf{S}_m$, edges $\mathbf{e}_{2_{k,l}}$, and edges $\mathbf{e}_{3_m}$ to Strasdat's pose graph optimization. Minimizing $\mathbf{e}_{1_{i,j}}$ and $\mathbf{e}_{2_{k,l}}$ suppresses changes in the relative transformations between camera poses, with the exception of gradual scale changes. Minimizing $\mathbf{e}_{3_m}$ keeps the positions of the geo-tagged images close to the positions obtained from the associated geo-tags. Our overall cost function E_{PGO} is defined as follows:

$$\begin{aligned} E_{PGO}(\{\mathbf{S}_i\}_{i\in C_1\cup C_2}) = {} & \lambda_1 \sum_{(i,j)\in C_3} \mathbf{e}_{1_{i,j}}^{\mathrm{T}} \mathbf{e}_{1_{i,j}} \\ & + \lambda_2 \sum_{(k,l)\in C_4} \mathbf{e}_{2_{k,l}}^{\mathrm{T}} \mathbf{e}_{2_{k,l}} + \lambda_3 \sum_{m\in C_2} \mathbf{e}_{3_m}^{\mathrm{T}} \mathbf{e}_{3_m} \end{aligned} \tag{9}$$

The corrected camera poses of keyframes $\mathbf{S}_n$ and geo-tagged images $\mathbf{S}_m$ are obtained by minimizing the cost function E_{PGO} on Lie manifolds using the LM algorithm. Following this optimization, we also reflect this correction in the 3D map points, as in [22].

3.6 Bundle Adjustment (BA)

Following the pose graph optimization, we refine the 3D reconstruction by applying bundle adjustment with the constraints of the geo-tagged images. Bundle adjustment is a classic method that jointly refines the 3D structure and camera poses (and camera intrinsic parameters) by minimizing the total re-projection errors. Each re-projection error $\mathbf{r}_{i,j}$ between the i^{th} 3D point and j^{th} camera is defined as:

$$\mathbf{r}_{i,j} = \mathbf{x}_{\mathbf{i}} - \pi(\mathbf{R}_j \mathbf{X}_i + \mathbf{t}_j) \tag{10}$$

$$\pi(\mathbf{p}) = [f_x \frac{\mathbf{p}_x}{\mathbf{p}_z} + c_x,\ f_y \frac{\mathbf{p}_y}{\mathbf{p}_z} + c_y]^{\mathrm{T}} \tag{11}$$

where $\mathbf{X}_i$ is a 3D point and $\mathbf{x}_i$ is the 2D observation of that 3D point; $\mathbf{R}_j$ and $\mathbf{t}_j$ are the rotation and translation of the j^{th} camera pose, respectively; $\mathbf{p} = [\mathbf{p}_x, \mathbf{p}_y, \mathbf{p}_z]^{\mathrm{T}}$ is a 3D point; $\pi(\cdot) : \mathbb{R}^3 \mapsto \mathbb{R}^2$ is the projection function; (f_x, f_y) is the focal length; and (c_x, c_y) is the center of projection.

To incorporate global position information of geo-tagged images with bundle adjustment, we add a penalty term corresponding to the constraint for a geo-tagged image [14]. The total cost function with this constraint is given by:

$$E_{BA}(\{\mathbf{X}_i\}_{i\in C_5}, \{\mathbf{T}_j\}_{j\in C_1}) = \sum_{(i,j)\in C_{\text{map-kf}}} \rho(\mathbf{r}_{i,j}^{\mathrm{T}} \mathbf{r}_{i,j}) + \lambda \sum_{m\in C_3} \|\mathbf{t}_m - \mathbf{y}_m\|^2 \tag{12}$$

where $\mathbf{T}$ is a camera pose of a keyframe represented as an element of SE(3), ρ is the Huber robust cost function, C_5 consists of map points observed by keyframes in C_1, and C_1 and C_3 are defined in Sec. 3.5. Both the positions of 3D points and the camera poses of keyframes are optimized by minimizing the cost function on Lie manifolds using the LM algorithm. This step can potentially correct the 3D map more precisely when it starts from a reasonably good 3D map.

4 Experiments

In this section, we evaluate the proposed method on the Málaga dataset [3], using geo-tagged images obtained from Google Street View. We also investigate the performance of pose graph optimization and bundle adjustment using the KITTI Dataset [9].

4.1 Implementation

We obtained geo-tagged images from Google Street View at intervals of 5 m within the area where the video was captured. We set the cost function weights to $\lambda_1 = \lambda_2 = 1.0 \times 10^5$ and $\lambda_3 = 1.0$, and we employed the g2o library [13] for the implementation of the pose graph optimization and bundle adjustment.

4.2 Performance of the Proposed Method

To verify the practical effectiveness of the proposed method, we evaluate it on the Málaga dataset using geo-tagged images obtained from Google Street View.

The Málaga Stereo and Laser Urban Data Set (the Málaga dataset) [3]—a large-scale video dataset that captures Street-View-usable areas—is employed in this experiment. The Málaga dataset contains a driving video captured at a resolution of 1024×768 at 20 fps in a Spanish urban area. We extracted two video clips (video 1 and video 2) from the video, and used these for the evaluation. The two video clips contain no loops, and their trajectories are over 1 km long. All frames in the videos contain inaccurate GPS positions, which are sometimes confirmed to contain errors of more than 10 m. Because of the inaccuracies, we manually assigned the ground truth positions to some selected keyframes by referring to the videos, inaccurate GPS positions, and Google Street View 3D Map. Fig. 4 presents an example of inaccurate GPS data and our assigned ground truth. Because the ground truth positions are assigned by taking into account the lane from which the video was taken, the errors in the ground truth are considered to be within 2 m, and these errors are sufficiently small for this experiment.

We evaluated the proposed method on the two videos by comparing the proposal and a baseline method that uses a similarity transformation (like a part of [25]). For the baseline method, we apply the initialization (INIT: described in Sec. 3.4) without applying pose graph optimization and bundle adjustment. We did not employ a global similarity transformation as a baseline because it cannot be applied until the end of the whole 3D reconstruction.

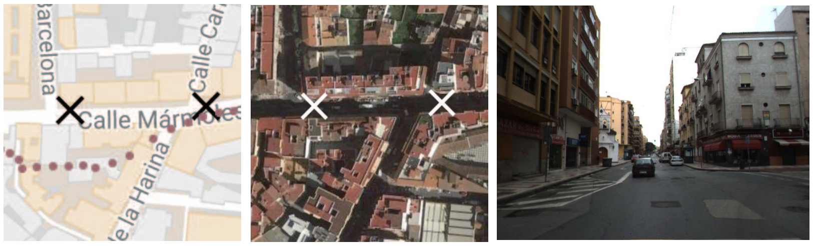


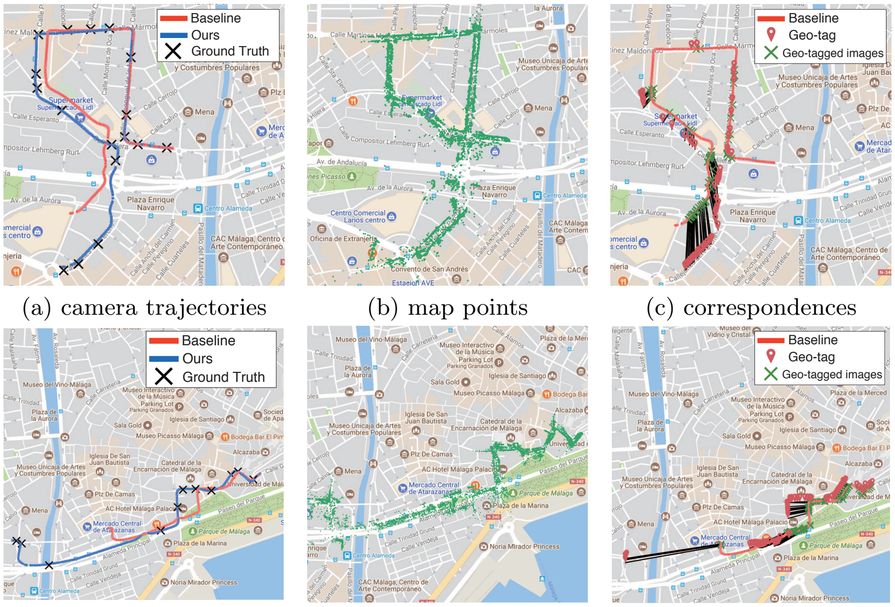
Fig. 4. The left figure shows an example of inaccurate GPS data (brown dots) and manually assigned ground truth positions (back crosses) on Google Maps. Although we use Google Maps to visualize the results clearly, the shapes of roads are not sufficiently accurate. Our ground truth positions are always assigned in the appropriate lane of the road, as seen in the satellite image (white crosses in the center figure). The right figure shows an example of a video frame captured at the left of the two ground truth positions in the left figure.

Table 1. Results of our proposed method on the Málaga dataset using Google Street View.

	video 1		video 2	
	Ave [m]	SD	Ave [m]	SD
Baseline (INIT)	54.8	141.3	142.5	249.8
Ours	6.7	5.6	6.0	3.0

To evaluate the proposed method quantitatively, we considered the average (Ave) and standard deviation (SD) of 2D distances between the ground truth positions and corresponding keyframe positions in the UTM coordinate system (in meters).

Table 1 presents the quantitative results, and Fig. 5 visualizes the results on Google Maps. As is clearly shown in these results, the baseline results accumulate scale errors, resulting in large errors of over 50 m. This is because the trajectories of these videos are long (greater than 1 km) and contain no loops. The proposed method sufficiently corrects scale drift, and significantly improves the 3D map by using geo-tagged images. In (b) and (e) of the visualized results, the 3D map points corrected using the proposed method are projected onto Google Maps, and it is shown that the 3D map points are correctly aligned to the map. To visualize all the correspondences between the 3D map coordinate system and the world coordinate system used in the proposal, we present the correspondences between the positions of geo-tagged images transformed by initialization and the positions of the corresponding geo-tags. These correspondences are employed incrementally for the correction.



(a) camera trajectories (b) map points (c) correspondences

(d) camera trajectories (e) map points (f) correspondences

Fig. 5. Results of our proposed method visualized on Google Maps. Top: results on video 1. Bottom: results on video 2. In (a) and (d), red and blue dots—which appear like lines—indicate the positions of keyframes corrected using a global similarity transformation (INIT) and our proposed method (Ours), respectively. In (b) and (e), 3D map points corrected by our method are depicted by green dots. (c) and (f) show all of the employed correspondences between the positions of geo-tagged images transformed using a global similarity transformation (green crosses) and the positions of the corresponding geo-tags (red pin icons). The correspondences are applied incrementally for scale drift correction in our proposed method.

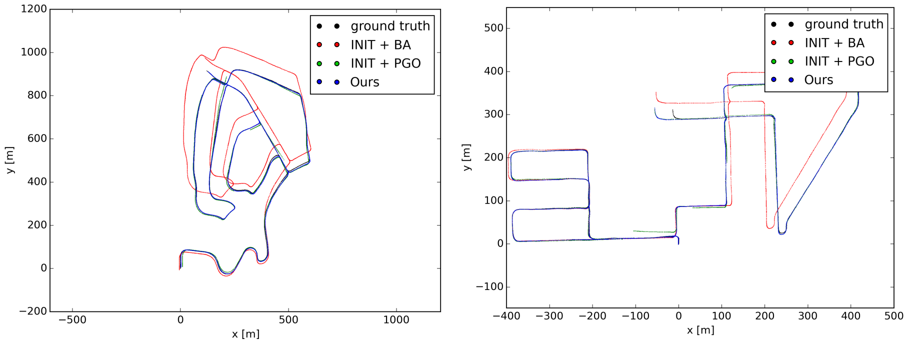
4.3 Performance of PGO and BA

To investigate the performance of the pose graph optimization and the bundle adjustment in our proposed method, we evaluated the performance using different combinations of these when varying the interval of $C_{\text{map-world}}$.

Through the previous experiment, we found that the geo-tag location information of Google Street View and the manually assigned ground truths of the Málaga dataset occasionally had errors of several meters. In this experiment, we control the interval of $C_{\text{map-world}}$, and use high-accuracy ground truths and geo-tags by using the KITTI dataset. The odometry benchmark of KITTI dataset [9] contains 11 sequences of stereo videos and precise location information obtained from RTK-GPS/IMU, and unfortunately Google Street View is not available in Germany where this dataset was captured. The experiment was conducted on two sequences, which include the largest and second-largest errors when applying ORB-SLAM: sequences 02 and 08 (containing 4660 and 4047 frames, respec-

Table 2. Results of the experiments on the KITTI dataset: sequences 02 and 08. Values denote average 2D errors between ground truth positions and the corresponding keyframe positions [m]. Ours consists of INIT, PGO, and BA.

	geotag interval (#02)					geotag interval (#08)				
	100	200	300	400	500	100	200	300	400	500
INIT + BA	1.15	2.22	65.86	164.17	96.66	0.45	1.24	18.43	148.17	52.40
INIT + PGO	4.57	4.26	7.54	10.89	11.96	0.93	2.83	4.64	5.44	9.11
Ours	2.27	2.51	4.87	6.89	12.35	0.50	2.06	2.84	4.19	6.38



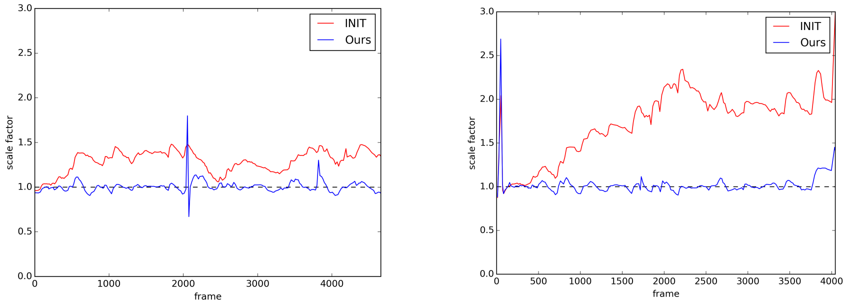
sequence 02

sequence 08

Fig. 6. Results of the experiment on the KITTI dataset when the interval of geo-tagged images is 300 frames. Keyframe trajectories estimated by INIT+BA, INIT+PGO, and Ours are visualized.

tively). The left images of the stereo videos are used as input, and pairs of a right image and location information are identified as geo-tagged images. All the location information associated with keyframes is used as the ground truth. In this experiment with KITTI dataset, we can compare the performances of correction methods accurately for the following reasons: geo-tag information and ground truths are sufficiently precise (open sky localization errors of RTK-GPS/IMU $<$ 5 cm); and errors in geo-tagged image localization are sufficiently small, because keypoint matching between corresponding left and right images performs very well.

For the comparison, we present the results of the methods employing the initialization + the pose graph optimization (INIT+PGO), and initialization + the bundle adjustment (INIT+BA). The correction method of INIT + BA is the same as [14], which is often used with a GPS location information. Ours includes the initialization, the pose graph optimization and the bundle adjustment. We changed the interval of geo-tagged images from 100 frames to 500 frames. For an equal initialization, we set geo-tagged images in the interval of 50 frames from the first to the 200^{th} frame.



sequence 02

sequence 08

Fig. 7. Change in scale factor of the proposed method on the KITTI dataset sequences 02 and 08.

Fig. 6 visualizes the ground truth and keyframe trajectories estimated by INIT+BA, INIT+PGO, and Ours when the interval of geo-tagged images is 300 frames. Table 2 presents the quantitative results of the experiment, where the values represent the average 2D errors between ground truth positions and the corresponding keyframe positions in the UTM coordinate system (in meters). Moreover, we report the errors of the global linear transformation on the sequence 02 and 08 by aligning the keyframe trajectory obtained by ORB-SLAM with ground truths through a similarity transformation: 20.15 and 25.12, respectively. The results show that bundle adjustment with geo-tag constraints, which is typically employed in the fusion of 3D reconstruction and GPS information [14], is not suitable when the interval of $C_{\text{map-world}}$ is large. It can also be seen that Ours (the combination of initialization, pose graph optimization, and bundle adjustment) often estimates the keyframe positions more accurately than any other method.

4.4 Scale Drift Correction

To confirm that scale drift is corrected incrementally, we visualize the change in scale factor of the proposed method on the KITTI dataset sequences 02 and 08. Fig. 7 shows that ORB-SLAM with the initialization accumulates scale errors, and our method can keep the scale factor around 1.

5 Conclusion

In this paper, we propose a novel framework for camera geo-localization that can correct scale drift by utilizing massive public repositories of geo-tagged images, such as those provided by Google Street View. By virtue of the expansion of such repositories, this framework can be applied in many countries around the world, without requiring the user to observe an environment. The framework integrates incremental SfM and a scale drift correction method utilizing geo-tagged images. In the correction method, we first acquire sparse 6-DoF correspondences between the 3D map coordinate system and the world coordinate system by using geo-tagged images. Then, we apply pose graph optimization over Sim(3) constraints and bundle adjustment. Our experiments on large-scale datasets show that the proposed framework sufficiently improves the 3D map by using geo-tagged images.

Note that our framework not only corrects the scale drift of 3D reconstruction, but also accurately geo-localizes a video. Our results are no less accurate than those of mobile devices (between 5 and 8.5 m) that use a cellular network and low-cost GPS [27], and those using monocular video and road network maps [4] (8.1 m in the KITTI sequence 02 and 45 m in sequence 08). This implies that geo-localization using geo-tagged images is sufficiently useful compared with methods using other GIS information.

Acknowledgement

This work was partially supported by VTEC laboratories Inc.

References

1. Google street view, https://www.google.com/streetview/
2. Agarwal, P., Burgard, W., Spinello, L.: Metric localization using google street view. In: Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on. pp. 3111–3118. IEEE (2015)
3. Blanco-Claraco, J.L., Moreno-Dueñas, F.Á., González-Jiménez, J.: The málaga urban dataset: High-rate stereo and LiDAR in a realistic urban scenario. The International Journal of Robotics Research **33**(2), 207–214 (2014)
4. Brubaker, M.A., Geiger, A., Urtasun, R.: Map-based probabilistic visual self-localization. IEEE transactions on pattern analysis and machine intelligence **38**(4), 652–665 (2016)
5. Caselitz, T., Steder, B., Ruhnke, M., Burgard, W.: Monocular camera localization in 3d lidar maps. In: Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on. pp. 1926–1931. IEEE (2016)
6. Clemente, L.A., Davison, A.J., Reid, I.D., Neira, J., Tardós, J.D.: Mapping large loops with a single hand-held camera. In: Robotics: Science and Systems. vol. 2 (2007)
7. Engel, J., Schöps, T., Cremers, D.: LSD-SLAM: Large-scale direct monocular slam. In: European Conference on Computer Vision. pp. 834–849. Springer (2014)
8. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
9. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 3354–3361 (2012)
10. Kaminsky, R.S., Snavely, N., Seitz, S.M., Szeliski, R.: Alignment of 3D point clouds to overhead images. In: Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on. pp. 63–70 (2009)
11. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on. pp. 225–234. IEEE (2007)
12. Klingner, B., Martin, D., Roseborough, J.: Street view motion-from-structure-from-motion. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 953–960 (2013)
13. Kümmerle, R., Grisetti, G., Strasdat, H., Konolige, K., Burgard, W.: g2o: A general framework for graph optimization. In: Robotics and Automation (ICRA), 2011 IEEE International Conference on. pp. 3607–3613. IEEE (2011)
14. Lhuillier, M.: Incremental fusion of structure-from-motion and GPS using constrained bundle adjustments. IEEE transactions on pattern analysis and machine intelligence **34**(12), 2489–2495 (2012)
15. Liu, Z., Marlet, R.: Virtual line descriptor and semi-local matching method for reliable feature correspondence. In: British Machine Vision Conference 2012. pp. 16–1 (2012)
16. Lu, F., Milios, E.: Globally consistent range scan alignment for environment mapping. Autonomous robots **4**(4), 333–349 (1997)
17. Majdik, A.L., Albers-Schoenberg, Y., Scaramuzza, D.: Mav urban localization from google street view data. In: Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on. pp. 3979–3986. IEEE (2013)

18. Middelberg, S., Sattler, T., Untzelmann, O., Kobbelt, L.: Scalable 6-DoF localization on mobile devices. In: European conference on computer vision. pp. 268–283. Springer (2014)
19. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Transactions on Robotics **31**(5), 1147–1163 (2015)
20. Rehder, J., Gupta, K., Nuske, S., Singh, S.: Global pose estimation with limited gps and long range visual odometry. In: Robotics and Automation (ICRA), 2012 IEEE International Conference on. pp. 627–633 (2012)
21. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: An efficient alternative to SIFT or SURF. In: Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 2564–2571. IEEE (2011)
22. Strasdat, H., Montiel, J., Davison, A.J.: Scale drift-aware large scale monocular SLAM. Robotics: Science and Systems VI (2010)
23. Tamaazousti, M., Gay-Bellile, V., Collette, S.N., Bourgeois, S., Dhome, M.: Nonlinear refinement of structure from motion reconstruction by taking advantage of a partial knowledge of the environment. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. pp. 3073–3080. IEEE (2011)
24. Untzelmann, O., Sattler, T., Middelberg, S., Kobbelt, L.: A scalable collaborative online system for city reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 644–651 (2013)
25. Wang, C.P., Wilson, K., Snavely, N.: Accurate georegistration of point clouds using geographic data. In: 3DTV-Conference, 2013 International Conference on. pp. 33–40 (2013)
26. Wendel, A., Irschara, A., Bischof, H.: Automatic alignment of 3D reconstructions using a digital surface model. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on. pp. 29–36. IEEE (2011)
27. Zandbergen, P.A., Barbeau, S.J.: Positional accuracy of assisted GPS data from high-sensitivity GPS-enabled mobile phones. The Journal of Navigation **64**(3), 381–399 (2011)