Bundle adjustment using aerial images with two-stage geometric verification

Hideyuki Kume[a,*], Tomokazu Sato[a], Naokazu Yokoya[a]

[a]*Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara, Japan*

Abstract

In this paper, a new pipeline of Structure-from-Motion for ground-view images is proposed that uses feature points on an aerial image as references for removing accumulative errors. The challenge here is to design a method for discriminating correct matches from unreliable matches between ground-view images and an aerial image. If we depend on only local image features, it is not possible in principle to remove all the incorrect matches, because there frequently exist repetitive and/or similar patterns, such as road signs. In order to overcome this difficulty, we employ geometric consistency-verification of matches using the RANSAC scheme that comprises two stages: (1) sampling-based local verification focusing on the orientation and scale information extracted by a feature descriptor, and (2) global verification using camera poses estimated by the bundle adjustment using sampled matches.

Keywords: Structure-from-motion, bundle adjustment, aerial image, RANSAC

1. Introduction

Structure-from-Motion (SfM) is one of the key techniques developed in the field of computer vision and has been used in many applications, such as three-dimensional reconstruction and image-based rendering. SfM became a widely

*Corresponding author

Email addresses: hideyuki-k@is.naist.jp (Hideyuki Kume), tomoka-s@is.naist.jp (Tomokazu Sato), yokoya@is.naist.jp (Naokazu Yokoya)

used tool after implementations of the state-of-the-art SfM (Bundler [1], VisualSFM [2], etc.) were distributed by their authors. They are very useful for processing a short image sequence. However, one significant problem in SfM, that is, the accumulation of estimation errors in a long image sequence with km order camera movement, remains to be solved. In this paper, to reduce accumulative errors in SfM, we propose a sampling-based bundle adjustment (BA) scheme using the aerial images that are already available for most outdoor scenes as external references.

Although many types of external references, e.g., 3D models [3, 4, 5, 6, 7, 8], GPS [9, 10], and road maps [11], have been used for reducing accumulative errors in SfM, we focus primarily on aerial images owing to their availability for outdoor environments. The existing methods using aerial images [12, 13, 14, 15, 16, 17, 18] are based on feature matching between given aerial images and ground-view images taken by standard cameras. Unfortunately, existing methods can handle only a short image sequence that does not include difficult situations. In this paper, we tackle more difficult situations where a large number of similar/repetitive patterns exist and/or only a few texture patterns are available in a long image sequence, e.g., uniformly tiled ground or a road environment, where most of the available feature points are on uniform road signs drawn on the ground surface. Even if we can approximately limit the search area of feature points by using GPS, which is also commonly used as an external reference in studies in the literature, if we depend on a local consistency check in the feature matching stage, in principle, it is not possible to remove all incorrect matches for a long image sequence because of the existence of repetitive and/or similar patterns.

In order to overcome this problem, we remove incorrect matches caused by repetitive/similar patterns by introducing a RANSAC framework [19] into both the feature matching and BA stages that verifies the local and the global consistencies among estimated camera poses and matched features. Figure 1 shows the flow of the proposed method. For local feature matching, in this study, we assume that we can approximately limit the area used for feature matching, by

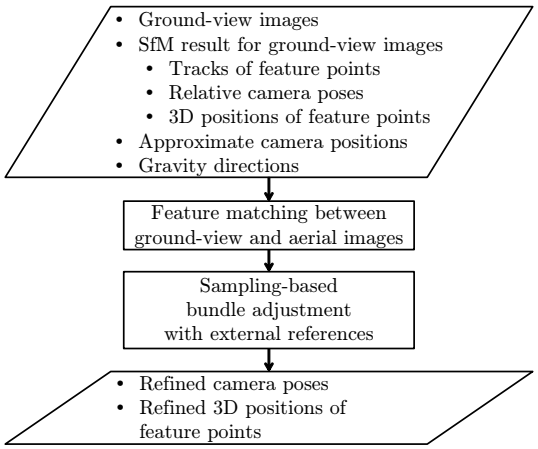
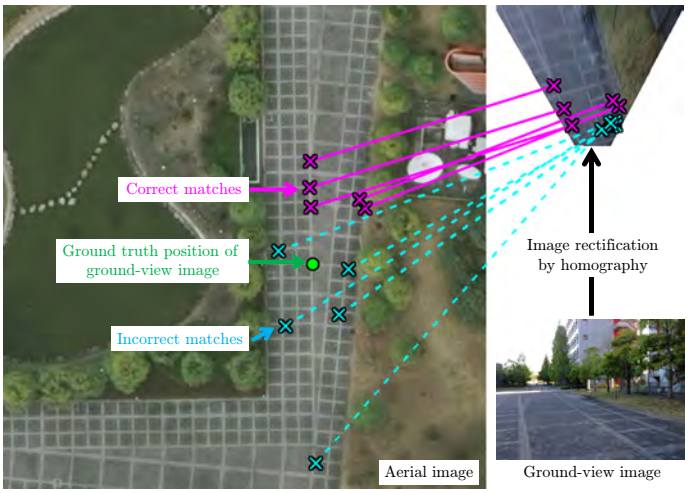


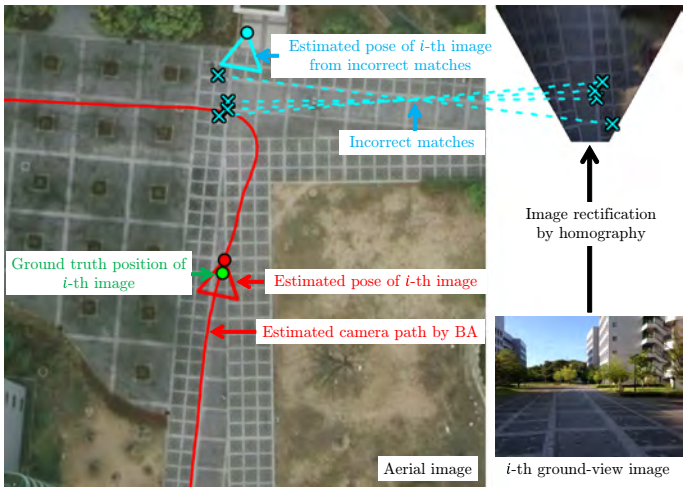
Figure 1: Flow of the proposed method.

using, e.g., GPS embedded in mobile devices. Figure 2(a) shows an example of a conventional feature matching result obtained by using a common combination of SIFT [20] and RANSAC for an aerial image containing many repetitive patterns. Even after limiting the search area and rectifying the ground-view image to facilitate matching, incorrect matches (blue dashed lines in the figure) are often erroneously determined to be inliers. In this scene, SIFT finds 158 tentative matches of which only 5 are correct. In order to successfully determine the incorrect matches as outliers, we modify the verification process of RANSAC so that it additionally checks the consistencies of matches according to the scale and orientation information from feature detectors and descriptors.

Although it is expected that most of the incorrect matches will be determined as outliers by this local verification method, some incorrect matches may simultaneously satisfy the consistency in the position, scale, and orientation, because the ground textures have similar structures. Figure 2(b) shows an example in which incorrect matches (blue dashed lines in the figure) are found even when the consistency check for scale and orientation is used. To remove these remaining incorrect matches, in the BA stage, we verify the global consistency of matches and poses for all the images using the RANSAC framework. More precisely, camera poses are first estimated by the BA scheme using sam-



(a) Local geometric verification. In this example, 158 tentative matches, of which only 5 are correct, are found by SIFT [20]. Blue dashed lines are incorrect matches selected by conventional RANSAC. Magenta solid lines are correct matches selected by RANSAC with a scale and orientation check.



(b) Global geometric verification. Blue dashed lines are incorrect matches obtained by RANSAC with a scale and orientation check. The red line and triangle represent the camera path and pose estimated by BA using sampled matches, respectively.

Figure 2: Two-stage geometric verification by RANSAC.

pled matches as external references (red line and triangle in the figure), and the consistency between the estimated poses and each match is then checked. After iterating sampling and estimation, the best samples that maximize the number of consistent frames are selected as inlier frames in which it is expected that incorrect matches will be excluded. When good feature matches have been obtained through the two-stage verification, the camera poses are refined by the BA scheme using the feature matches as external references to remove accumulative errors.

It should be noted that the proposed method assumes that an SfM result for ground-view images is given as an initial guess for the BA. The camera model for an aerial image can be approximated by the orthographic camera model, and its image plane is perpendicular to the gravity direction. In addition, approximate positions and gravity directions of the ground-view images are given by external sensors. An easy method for obtaining this information, which is employed in this study, is to use the GPS and gyroscope sensors embedded in most recent smartphones. It should be noted that this paper is an extended version of a previous conference paper [21]. We have added experiments using a roadway and an in-depth discussion in this version.

2. Related Work and our Contributions

To reduce accumulative errors in SfM, loop closing techniques [22, 23, 24] are sometimes employed. When the loops have been detected, the accumulative errors can be reduced in the BA stage. In an approach related to loop closing, Cohen et al. [25] exploited symmetries, which often exist in man-made structures, instead of loops. Although these techniques are effective for some applications, it is essentially difficult for these techniques to remove accumulative errors for a general image sequence without either loops or symmetry.

To reduce accumulative errors in an image sequence captured by a moving camera, several kinds of external references have been used. These external references can be classified into 3D models [3, 4, 5, 6, 7, 8], GPS [9, 10], road

maps [11], and aerial images [12]. In studies in the literature, many types of 3D models, including 3D points [3, 4], wire-frame models [5], plane-based models [6, 7], textured 3D models [8], and digital elevation models (DEM) [7] were employed as references. One disadvantage of 3D-model-based methods for large outdoor environments is that time consuming manual intervention is required to create the 3D models. Although some models are already available in the GIS database [6, 7], the available areas are still limited to large cities. One method to create 3D models without much manual intervention is to use 3D reconstruction techniques, e.g., SfM and multi-view stereo [26]. However, the reconstructed models are also affected by accumulative errors caused by SfM itself.

In contrast to 3D models, GPS, road maps, and aerial images are already available for most outdoor scenes around the world. Yokochi et al. [9] and Lhuillier [10] proposed extended-BA using GPS that minimizes the energy function defined as the sum of reprojection errors and a penalty term of GPS. This method can globally optimize camera poses and reduce accumulative errors by updating poses so as to minimize the energy function. However, the accuracy of this method is directly affected by errors in GPS positioning, which easily grow to several tens of meters in urban areas when using the GPS embedded in smartphones. Brubaker et al. [11] proposed a method that uses community developed road maps. This method can reduce accumulative errors by matching the trajectory from SfM to road maps, unless there are ambiguities in the matched trajectories (e.g., straight roads and Manhattan worlds). Pink et al. [12] fused sparsely obtained camera poses from aerial images into SfM by using the Kalman filter. However, unlike BA-based fusion, global optimization is difficult in the Kalman filter-based approaches.

There exist several methods that use aerial images as one of input data of SfM [27] or RGBD-SLAM [28]. Shan et al. [27] employed oblique aerial images as additional inputs of SfM for reconstructing the regions that are not covered by ground-view images. One challenge in our case is the employment of top-view aerial images as an external reference in which common feature

matching method used in [27] cannot give reasonable matches. For obtaining matches between top-view aerial images and ground-view images, Forster et al. [28] utilized dense depth maps obtained from perspective aerial images in the feature matching stage. Although feature matching methods from widely different viewpoints [29, 30] also work in the case depth/3D information is available, unfortunately, the information is not always easy to be obtained from commonly available orthographic aerial images.

On the other hand, some methods estimate camera poses directly from aerial images [13, 14, 15, 16, 17, 18]. There are two types of aerial images: perspective and orthographic. Bansal et al. [13] proposed a method for estimating camera poses by matching façades in the ground-view input image with perspective aerial images. Although perspective aerial images are available on Google Maps and Microsoft Bing Maps, the available areas are still limited to large cities.

Most methods using aerial images employ the orthographic aerial images. These methods can be classified into learning- [14] and feature-matching-based [15, 16, 17, 18]. Lin et al. [14] proposed a method based on the relationship of the appearance between ground-view and aerial images learned through community photos with position information. Although this method estimates camera positions from large regions (1,600 [km^2] in their experiments), camera positions can be estimated only approximately. Other methods match the building edges [15, 16] or feature points [17, 18] of ground-view and aerial images. One of the difficulties of this approach is finding good matches for all the images of a video sequence under severe conditions for feature matching. Toriya et al. [17] and Noda et al. [18] relaxed the problem by stitching multiple ground images for feature matching. Toriya et al. [17] also proposed a robust feature matching procedure that compares the orientation and scale of each match with the *dominant* orientation and scale. Unfortunately, existing feature matching-based methods, which are expected to achieve highly accurate pose estimation, do not have the capability to handle a long image sequence because of the strong dependence on matching information given only for a local region. To the best of our knowledge, no method exists that handles feature matches between an aerial

image and ground-view images in the global optimization stage (BA stage).

As mentioned in the previous section, we tackle difficult situations for feature matching, where a large number of similar patterns exist in a large-scale outdoor environment. The main contributions of this paper are summarized as follows:

- BA-based global optimization that uses feature matches between ground-view and aerial images.
- Two-stage geometric verification for removing incorrect matches
  - Local verification that focuses on the transformation between aerial image and each ground-view image and considers in particular the orientation and the scale information extracted by a feature descriptor,
  - Global verification that focuses on camera poses estimated using the BA scheme with sampled matches.

3. Feature Matching between Ground-view and Aerial Images

This section describes a method to find good matches between an aerial image and each ground-view image with local verification. As shown in Fig. 3, the method is composed of three processes: (1) ground-view image rectification, (2) feature matching, and (3) local geometric verification by RANSAC.

3.1. Ground-view image rectification and feature matching

Before finding matches, as in existing methods [17, 18], we rectify the ground-view images so that the texture patterns are similar to those of the aerial image. To achieve this, we use homography calculated from the gravity direction in the camera coordinate system. More precisely, we map the ground image to a plane that is perpendicular to the gravity direction. To estimate the gravity direction, the vanishing points of parallel lines [31] or a gyroscope sensor can be used. Since even a cheap gyroscope sensor provides an accurate gravity direction, we used a gyroscope embedded in a smartphone in the experiment described below. Even

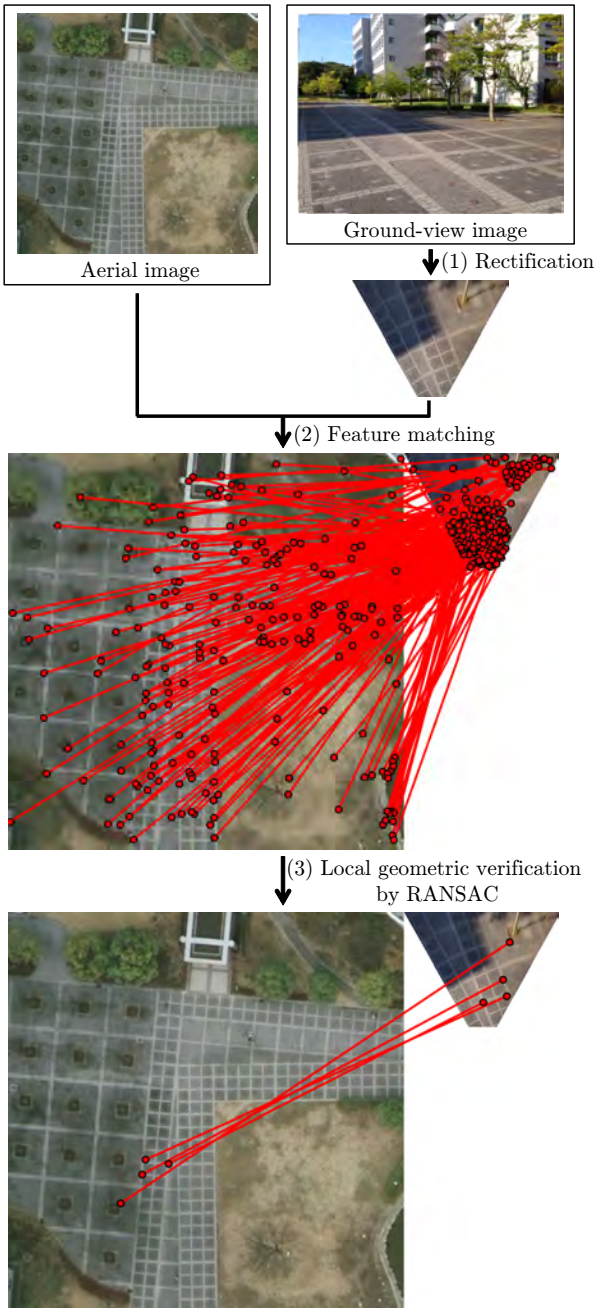


Figure 3: Flow of the feature matching.

if the patterns cannot be perfectly rectified because of the irregularity of the ground plane, it is expected that the chance of obtaining correct matches will be increased by using this rectification process.

A region in an aerial image for feature matching is then determined. We first determine a certain size of the region the center of which is the GPS position, which includes measurement errors. In the experiment, the size is set to 50 [m] $\times$ 50 [m]. Tentative matches are then found between the rectified ground image and the limited region in the aerial image using a feature detector and a descriptor. Although we employed SIFT [20] in the experiment because of its robustness, any feature operators that output scale and orientation information can be employed in our framework. It should also be noted that a large GPS error may result in correct feature points outside the limited region. Even in this case, incorrect matches are automatically excluded by applying two-stage RANSAC with a geometric consistency check.

3.2. Local geometric verification

Tentative matches often include many incorrect matches. The rate of incorrect matches sometimes reached over 95% in our experiment, even if the search range for matching was correctly set. In order to decrease the number of incorrect matches included in the tentative matches, we apply local geometric verification by using RANSAC with a consistency check of the orientation and scale of texture patterns, as shown in Fig. 4.

Although final camera poses are estimated in 6-DOF with BA, to achieve stable matching between rectified ground-view images and the aerial image, we use a 3-DOF similarity transform, which is composed of scale s, rotation θ, and translation $\boldsymbol{\tau}$, as the model in RANSAC. In a standard RANSAC procedure, tentative matches of minimum number required for estimating the similarity transform $(s, \theta, \boldsymbol{\tau})$, which are two matches in this case, are randomly selected first. The number of inlier matches that satisfy the following condition is then counted.

$$|\boldsymbol{a}_k - (s\mathrm{R}(\theta)\boldsymbol{g}_k + \boldsymbol{\tau})| < d_{\mathrm{th}}, \tag{1}$$

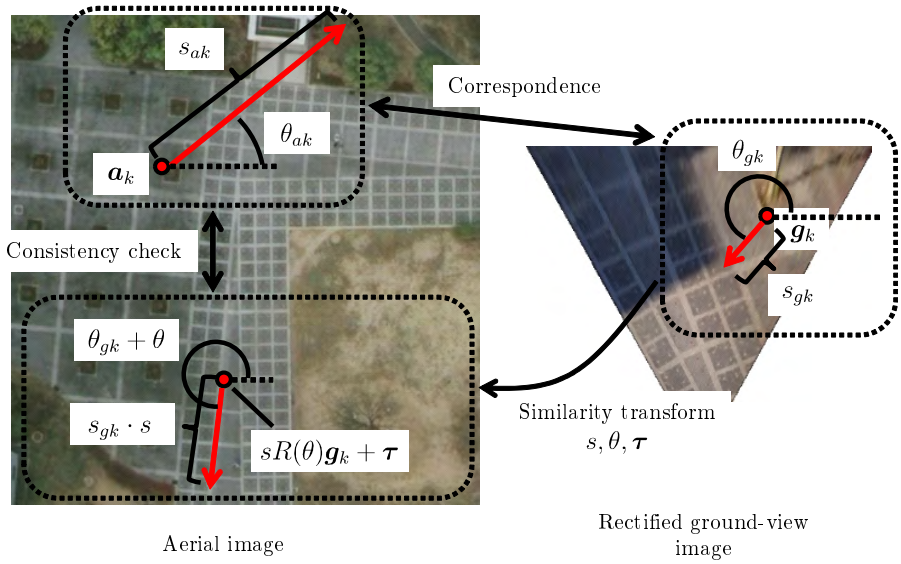


Figure 4: Criteria used in local geometric verification.

where $\boldsymbol{a}_k$ and $\boldsymbol{g}_k$ are the 2D positions of the k-th match in the aerial image and the rectified ground-view image, respectively. R(θ) is the 2D rotation matrix with rotation angle θ, and d_{th} is a threshold. After iterating the random sampling process, the trial with the largest number of inlier matches is selected.

The problem here is that the distance-based single criterion described above cannot successfully find correct matches when there exists a huge number of incorrect matches. In order to achieve more robust matching, we modify the criterion commonly used in RANSAC by adding a consistency check for orientation and scale information extracted from a feature descriptor. More precisely, we select the matches that simultaneously satisfy Equation (1) and the following two conditions as inliers in RANSAC procedure.

$$\max\left(\frac{s_{\text{gk}}\cdot s}{s_{\text{ak}}},\frac{s_{\text{ak}}}{s_{\text{gk}}\cdot s}\right)<s_{\text{th}}, \tag{2}$$

$$\text{aad}(\theta_{\text{gk}}+\theta,\theta_{\text{ak}})<\theta_{\text{th}}, \tag{3}$$

where $(s_{\text{ak}},s_{\text{gk}})$ and $(\theta_{\text{ak}},\theta_{\text{gk}})$ represent the scale and orientation of feature points for the k-th match on the aerial image and the rectified ground-view

image, respectively. The function 'aad' returns the absolute angle difference in the domain $[0°, 180.0°]$. s_{th} and θ_{th} are the thresholds for scale and angle, respectively. By using the additional consistency check, the feature matches are strictly verified, and it is expected that most of the incorrect matches will be removed as outliers. It should be noted that even though we employ a 3-DOF model in RANSAC in this stage, as shown in the experiment in Section 5, feature points on slanted ground that violate the 3-DOF model are successfully matched, since we can relax each threshold by simultaneously checking three criteria in this stage.

4. Sampling-based Bundle Adjustment

As shown in Fig. 2(b), some frames contain incorrect matches even after local geometric verification because of repetitive similar patterns. In this study, as a global verification stage, we propose a new sampling-based BA scheme to find the frames that contain incorrect matches.

4.1. Definition of energy function

In order to use the matches between ground-view and aerial images in BA as external references, as shown in Fig. 5, the energy function E is newly defined for this problem as the sum of reprojection errors for both ground-view (perspective) images Φ and the aerial (orthographic) image Ψ:

$$\mathrm{E}(\{\boldsymbol{R}_i, \boldsymbol{t}_i\}_{i=1}^{I}, \{\boldsymbol{p}_j\}_{j=1}^{J}) = \Phi(\{\boldsymbol{R}_i, \boldsymbol{t}_i\}_{i=1}^{I}, \{\boldsymbol{p}_j\}_{j=1}^{J}) + \omega\Psi(\{\boldsymbol{p}_j\}_{j=1}^{J}), \quad (4)$$

where $\boldsymbol{R}_i$ and $\boldsymbol{t}_i$ represent 3D rotation and translation from the world coordinate system to the camera coordinate system for the i-th frame, respectively. $\boldsymbol{p}_j$ is a 3D position of the j-th feature point, I and J are the number of frames and feature points, respectively, and ω is a weighting coefficient that balances Φ and Ψ. Since the energy function is non-linearly minimized in BA, good initial values of parameters are required to avoid local minima. Before minimizing the energy function, we fit the parameters estimated by SfM to the GPS positions

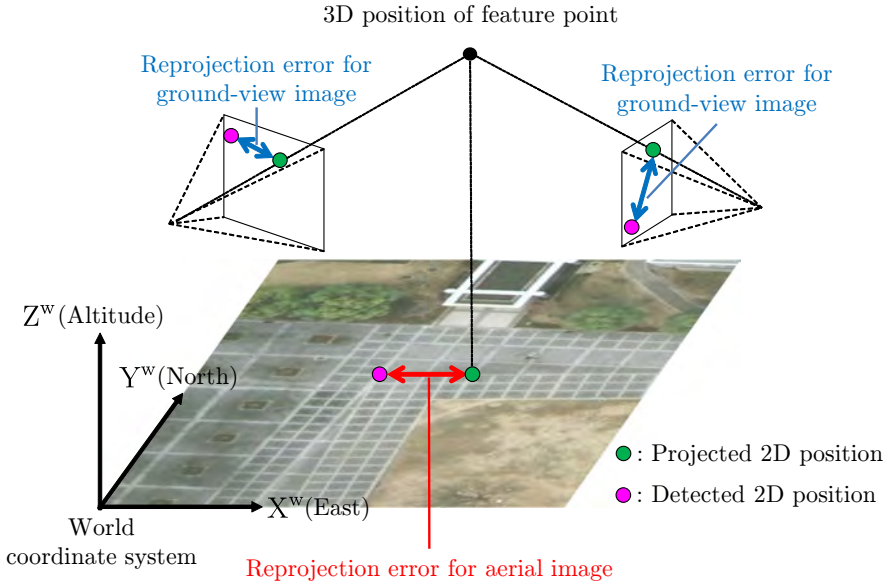


Figure 5: Reprojection errors for ground-view (perspective) images and an aerial (orthographic) image.

using a 3D similarity transform. In the following, two energy terms associated with reprojection errors Φ and Ψ are given in detail.

4.1.1. Reprojection errors for ground-view images

In our method, camera poses and 3D positions of the feature points estimated by BA dynamically move in the world coordinate system, which is set on the aerial image coordinate, because of the tension from the external references (matches on the aerial image). Because of this dynamic camera movement, the 3D positions of the reference points on an aerial image frequently go behind the camera. However, the commonly used reprojection errors for the pinhole camera model cannot deal with projections from behind the camera. In this study, instead of the commonly used squared distance errors on the image plane, we employ the following angular reprojection error that is employed in SfM for

omnidirectional cameras.

$$\Phi(\{\boldsymbol{R}_i, \boldsymbol{t}_i\}_{i=1}^{I}, \{\boldsymbol{p}_j\}_{j=1}^{J}) = \frac{1}{\sum_{i=1}^{I} |\boldsymbol{P}_i|} \sum_{i=1}^{I} \sum_{j \in \boldsymbol{P}_i} \Phi_{ij}, \tag{5}$$

$$\Phi_{ij} = \angle \left(\begin{pmatrix} x_{ij} \\ f_i \end{pmatrix}, \begin{pmatrix} X_{ij} \\ Z_{ij} \end{pmatrix} \right)^2 + \angle \left(\begin{pmatrix} y_{ij} \\ f_i \end{pmatrix}, \begin{pmatrix} Y_{ij} \\ Z_{ij} \end{pmatrix} \right)^2, \tag{6}$$

$$(X_{ij}, Y_{ij}, Z_{ij})^{\mathrm{T}} = \boldsymbol{R}_i \boldsymbol{p}_j + \boldsymbol{t}_i, \tag{7}$$

where $\boldsymbol{P}_i$ is a set of feature points detected in the i-th frame. Function $\angle$ returns an angle between two vectors, $(x_{ij}, y_{ij})^{\mathrm{T}}$ is a detected 2D position of the j-th feature points in the i-th frame, and f_i is the focal length of the i-th frame. By this definition, the energy becomes large when projections behind the camera occur.

Here, as mentioned in [32], the convergence of energy is very poor with an angular reprojection error $\hat{\Phi}_{ij} = \angle((x_{ij}, y_{ij}, f_i)^{\mathrm{T}}, (X_{ij}, Y_{ij}, Z_{ij})^{\mathrm{T}})^2$. We then split the angular reprojection error into xz and yz components in order to simplify the Jacobian matrix of E required by non-linear least squares methods, such as the Levenberg-Marquardt method. The first and second terms of Φ_{ij} do not depend on the y and x components of $\boldsymbol{t}_i$ in this definition. We experimentally confirmed that this splitting largely affects the convergence performance.

4.1.2. Reprojection errors for aerial image

The reprojection errors for the aerial (orthographic) image are defined as

$$\Psi(\{\boldsymbol{p}_j\}_{j=1}^{J}) = \frac{1}{\sum_{i \in \boldsymbol{M}} |\boldsymbol{A}_i|} \sum_{i \in \boldsymbol{M}} \sum_{j \in \boldsymbol{A}_i} |\boldsymbol{a}_j - \mathrm{pr}(\boldsymbol{p}_j)|^2, \tag{8}$$

where $\boldsymbol{M}$ is a set of frames in which matches between ground-view and aerial images are found, $\boldsymbol{A}_i$ is a set of feature points that are matched to the aerial image in the i-th frame, and $\boldsymbol{a}_j$ is the 2D position of the j-th feature point in the aerial image. The function 'pr' projects a 3D point onto the xy plane (aerial image coordinate system). Although the height of the 3D points is not affected by this term, the remaining 2D positions are constrained to their positions on the

aerial image. These constraints are effective for reducing the accumulative errors through simultaneously minimizing both the perspective and orthographic reprojection errors in the BA.

4.2. Global geometric verification

This section describes a RANSAC scheme introduced into BA for global geometric verification. Since the matches remaining after local verification are consistent in each frame, we judge inliers in a frame-wise manner. First, we randomly sample n frames from the frames that passed local geometric verification and execute BA using the matches in the sampled frames, i.e., using a set of sampled frames $\boldsymbol{M}'$ instead of $\boldsymbol{M}$ in Equation (8). We then check the consistency between the camera poses obtained by BA and each frame that includes feature matches. More precisely, we count the number of inlier frames that satisfy the condition

$$\underset{j \in \boldsymbol{A}_i}{\text{average}}(\alpha_{ij}) < \alpha_{\text{th}}, \tag{9}$$

where α_{ij} is an angular reprojection error of the j-th feature point on the aerial image coordinate system, as shown in Fig. 6, and α_{th} is a threshold. Here, α_{ij} is computed as

$$\alpha_{ij} = \angle\left(\boldsymbol{a}_j - \text{pr}(-\boldsymbol{R}_i^{\text{T}}\boldsymbol{t}_i), \text{pr}\left(\boldsymbol{R}_i^{\text{T}}(x_{ij}, y_{ij}, f_i)^{\text{T}}\right)\right). \tag{10}$$

After iterating the random sampling process at given times, the trial that has the largest support is selected. Finally, camera poses are refined by executing BA again using the feature matches in the selected inlier frames.

In the experiments described below, the threshold α_{th} is experimentally determined. It should also be noted that biased sampling, where samples are close to each other, frequently yields an unstable result in RANSAC. Thus, we modify the random sampling process of frames so that the distances between the average positions of matches on an aerial image are larger than threshold l_{th}.

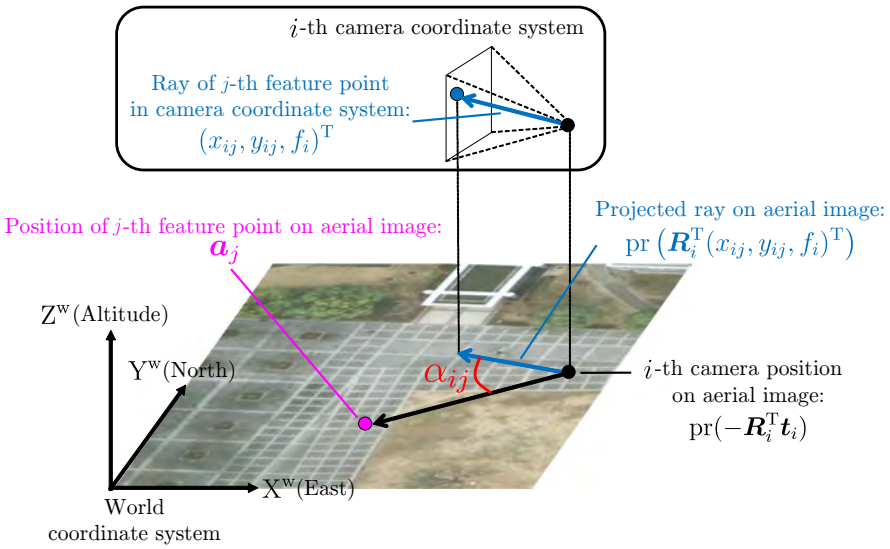


Figure 6: Criterion used in global geometric verification.

5. Experiments

To validate the effectiveness of the proposed method, we quantitatively evaluated the performance of the proposed BA with two-stage geometric verification using two datasets: (1) data captured by a hand-held sensor unit on textured ground for Experiment 1, and (2) data captured by a car-mounted sensor unit on a roadway for Experiment 2. In the following, we first describe the setup used for both the experiments. The results of each experiment are then detailed.

5.1. Experimental setup

We used an iPhone 5 (Apple) as a sensor unit including a camera, GPS, and a gyroscope. The GPS and gyroscope measured the position at 1 [Hz] and the direction of gravity for every frame, respectively. We also used an RTK-GPS (Topcon GR-3, 1 [Hz]; horizontal positioning accuracy in the specification sheet is 0.01 [m]) to obtain the ground truth positions. The positions obtained from the GPS data were assigned temporally to the nearest frame. As external references, we downloaded the aerial images covering the area used in the

experiments from Google Maps [maps.google.com], whose coordinate system is associated with the metric scale.

To obtain the initial values for the BA, we employed VisualSFM [2] as a state-of-the-art SfM implementation. For non-linear optimization, we used Ceres-Solver [33]. We experimentally set $d_{\mathrm{th}} = 2$ [pixel], $s_{\mathrm{th}} = 2$ and $\theta_{\mathrm{th}} = 40$ [°] for the feature matching, and $\omega = 10^{-5}$ and $\alpha_{\mathrm{th}} = 5.0$ [°] for the BA.

5.2. Quantitative evaluation using data captured on textured ground (Experiment 1)

In this experiment, we used video images (640 [pixel] × 480 [pixel], 2,471 frames, 494 [s]) captured by a hand-held sensor unit on a textured ground. As shown in Fig. 2, a large number of similar patterns exist on the ground in this environment. Figure 7 shows an aerial image covering the area used in the experiments (approximately 1 [pixel] = 5.2 [cm]).

5.2.1. Effect of local verification

In this experiment, we first evaluated the effectiveness of the proposed feature matching process including local geometric verification by RANSAC using the scale and orientation check described in Section 3. Here, we tested local verification with variable thresholds s_{th} and θ_{th}. To count the number of correctly matched frames, we first selected frames that had four or more inlier matches after local verification. From these frames, we manually selected the frames whose matches were correct.

Figure 8 shows the rate and the number of frames in which all the selected matches were correct. It should be noted that $s_{\mathrm{th}} = \infty$ and $\theta_{\mathrm{th}} = 180.0$ [°], which means that the orientation check and scale check were disabled, respectively. The results indicate that the rate was significantly improved through the scale and orientation check. We can also confirm that small values of s_{th} and θ_{th} tend to increase this rate. However, the number of correctly matched frames, which is important for optimizing camera poses using BA, was decreased when

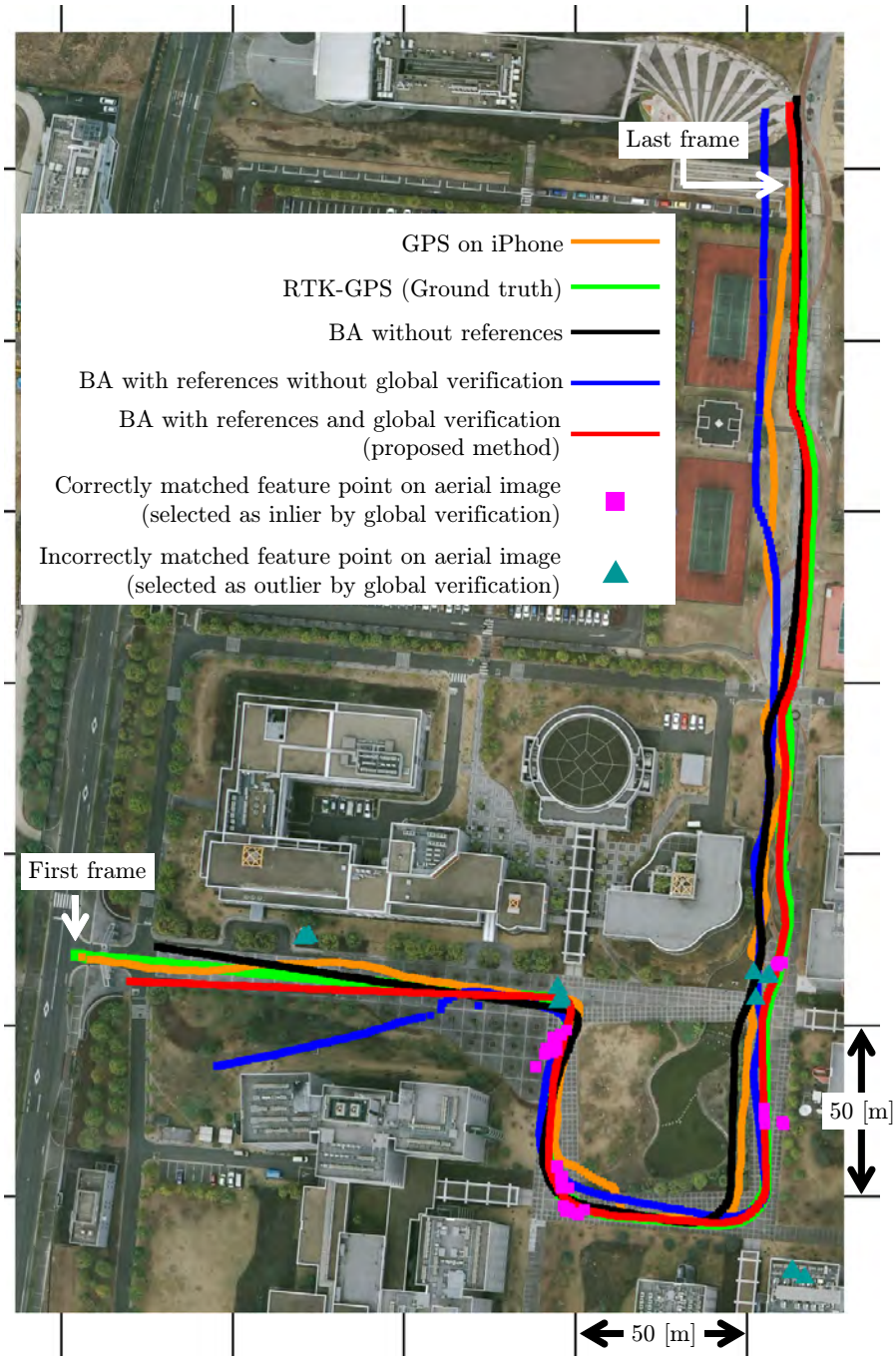
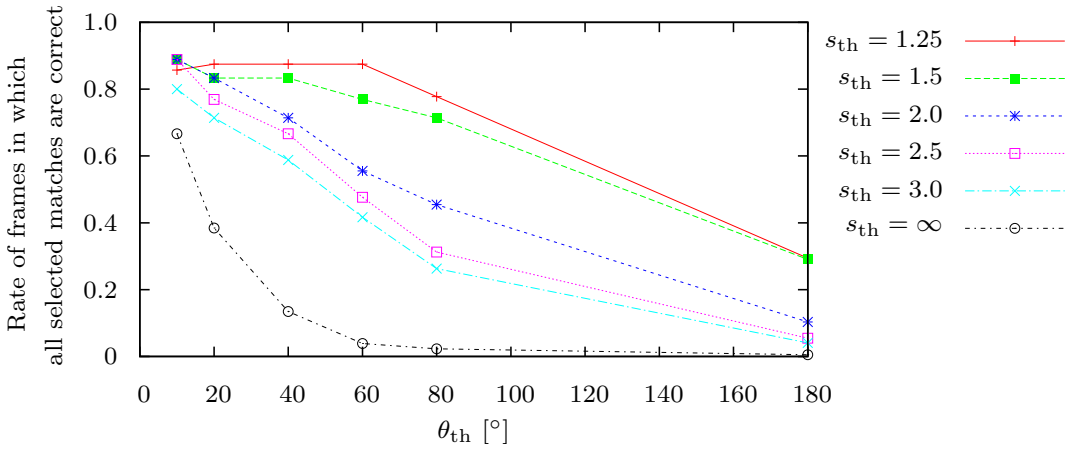
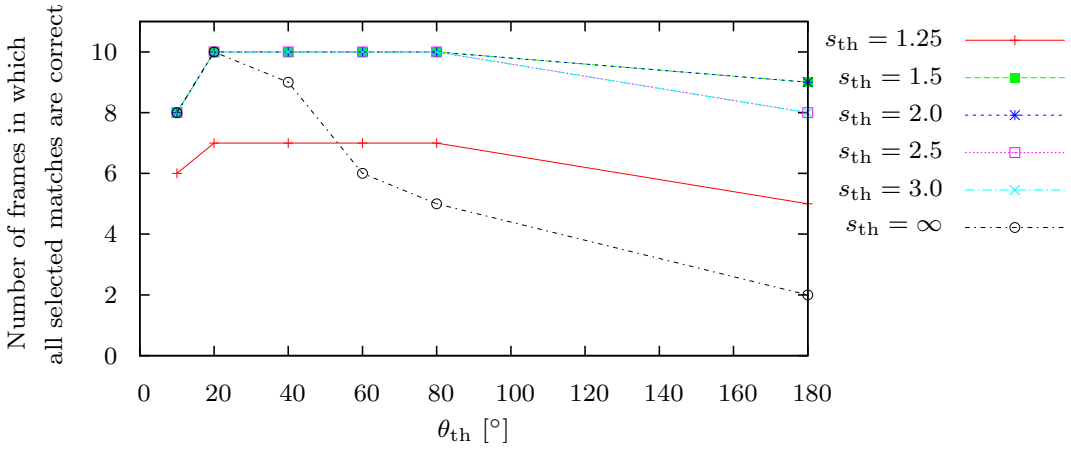


Figure 7: Experimental environment and results (Experiment 1).



(a) Rates of frames in which all selected matches are correct



(b) Numbers of frames in which all selected matches are correct

Figure 8: Rates and numbers of frames in which all selected matches are correct (Experiment 1).

using small thresholds. In the following experiments, we then employed feature matches with $s_{\mathrm{th}} = 2$ and $\theta_{\mathrm{th}} = 40$ [°].

Figure 9 shows the effects of the scale and orientation check for two sampled images. In both cases, local verification without a scale and orientation check could not select any correct matches, whereas the proposed local verification with a scale and orientation check was able to do so. However, as shown in Fig. 10, incorrect matches still remained even when we used both the scale and orientation check, because similar patterns exist.

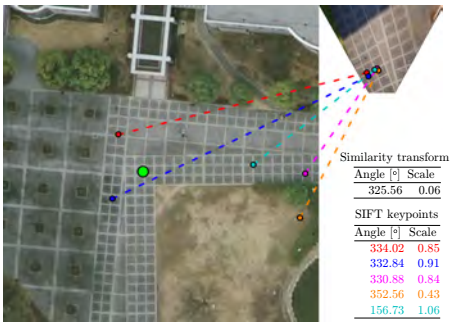
5.2.2. Effect of global verification

We then evaluated the effectiveness of global geometric verification by sampling-based BA, as described in Section 4. In this stage, frames with GPS data were sampled (650 out of 2,471 frames) and used for the BA to reduce the computational time. As external references, we used the frames and feature matches selected through the orientation and scale check described in the previous section. Here, 10 out of 14 frames had correct matches.

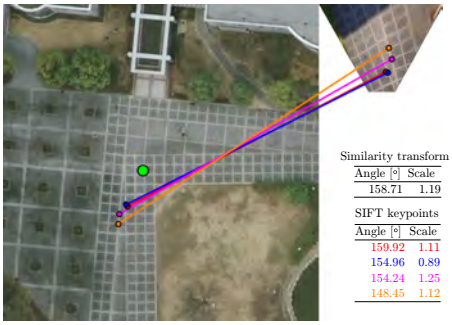
We first investigated the influence of weight ω for balancing two types of reprojection errors in the energy function of the BA. Figure 11 shows the average position errors produced by the BA with variable weight ω using all the correctly matched frames. This result demonstrates that position errors did not largely depend on weight ω, except when small values were applied. In the experiments described next, we employed $\omega = 10^{-5}$.

We next evaluated the proposed sampling-based global verification in terms of its capability to select frames with correct matches. Here, we experimentally set $n = 4$ and $l_{\mathrm{th}} = 25$ [m], and tested 100 trials. Figure 12 shows the number of inlier frames produced by global verification with variable threshold α_{th}. The results demonstrate that incorrectly matched frames were selected as inliers when large values of α_{th} were used and that the number of correctly matched frames decreased when small values of α_{th} were used. In the experiments described next, we employed $\alpha_{\mathrm{th}} = 5.0$ [°].

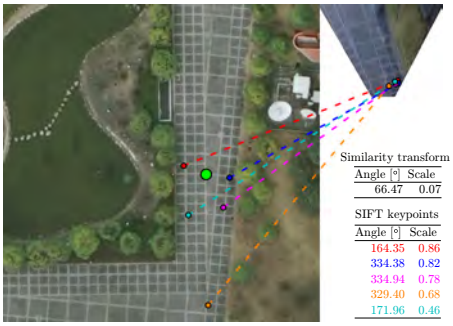
We also checked the number of inlier frames selected in each trial with $\alpha_{\mathrm{th}} =$



(a) Without orientation and scale check



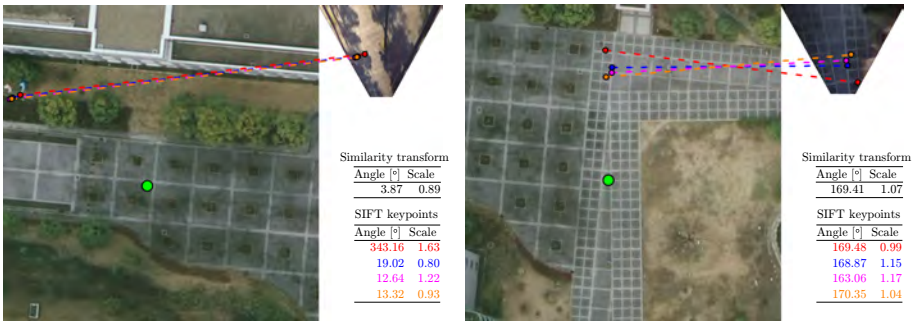
(b) With scale check, with/without orientation check



(c) Without orientation and scale check

(d) With orientation check, with/without scale check

Figure 9: Selected inliers for example images (Experiment 1). The solid and dashed lines represent correct and incorrect matches, respectively. The relative angle and scale of the matched feature points are shown in the bottom right hand table together with the corresponding line colors. The green points are the ground truths of the camera positions. Note that local verification with/without the orientation check for (b) and scale check for (d) gave the same results.



(a) (b)

Figure 10: Examples of incorrect matches by local verification using orientation and scale check (Experiment 1). The interpretations of the symbols are the same as in Fig. 9.

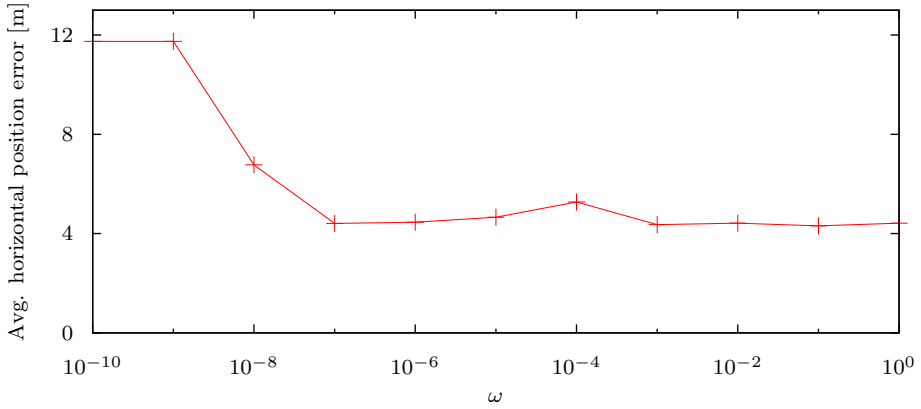


Figure 11: Relationship between weight ω and average horizontal position error (Experiment 1).

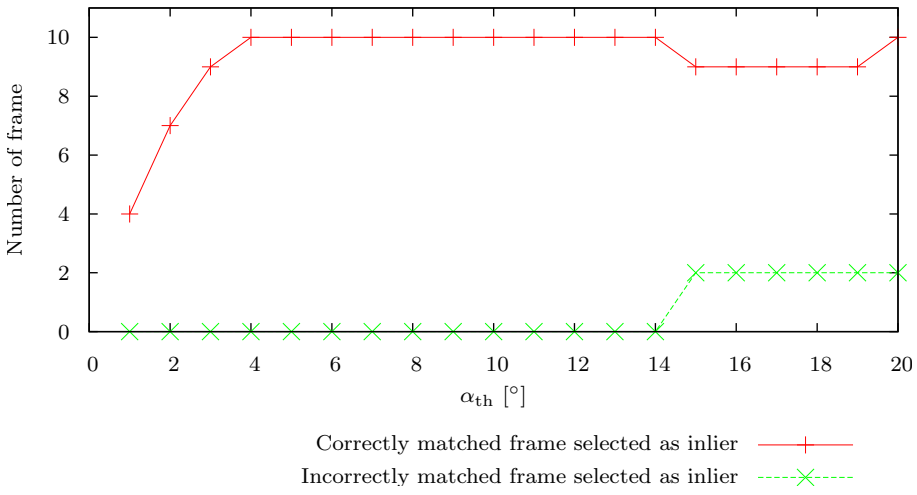


Figure 12: Number of inlier frames with variable threshold α_{th} (Experiment 1).

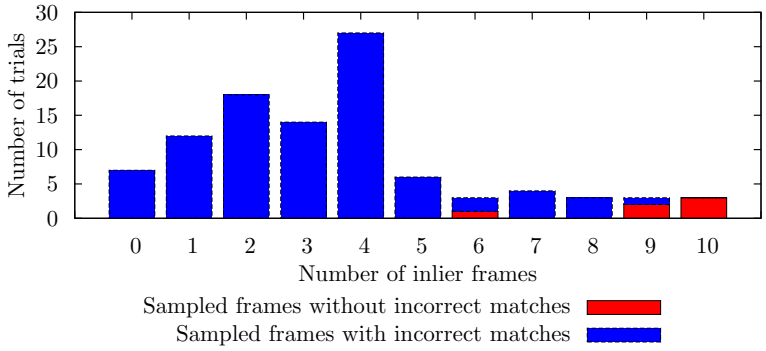


Figure 13: Number of trials and inlier frames derived by each trial (Experiment 1).

5.0 [°]. Figure 13 shows the number of trials and inlier frames derived by each trial. In this figure, it can be seen that the sampled frames without incorrect matches tend to increase the number of inlier frames. This result demonstrates that the criterion of global verification is effective. We also confirmed that the trials that derived the largest number of inlier frames successfully selected all of the correct matches.

In order to validate the effectiveness of the global verification and the use of external references on an aerial image, the results of the following three methods were compared.

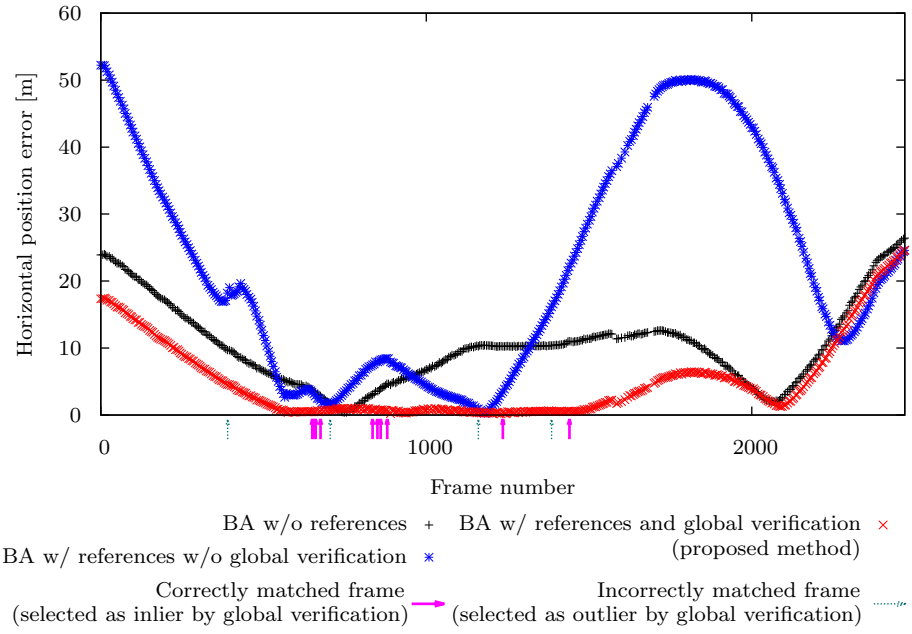


Figure 14: Horizontal position error in each frame (Experiment 1).

- BA without references [2]
- BA with references without global verification
- BA with references and global verification (proposed method).

Figures 7 and 14 show the estimated camera positions and horizontal position errors for each frame, respectively. Since the BA without references cannot estimate absolute camera poses, we fitted the camera positions estimated using SfM to the ground truths through a similarity transform. These results demonstrate that the camera positions estimated through the BA without references were affected by the accumulative errors. The BA without global verification was affected by the incorrect matches. The proposed BA with global verification reduced the accumulative errors. It should be noted that, at the end of the sequence, the accumulative errors still remained, because the ground was not level and thus no matches were found.

5.3. Quantitative evaluation using data captured on roadways (Experiment 2)

In this experiment, we used video images (640 [pixel] × 480 [pixel], 7,698 frames, 396 [s]) captured by a car-mounted sensor unit on a roadway. Figure 15 shows an aerial image covering the area used in the experiments (approximately 1 [pixel] = 4.5 [cm]). It should be noted that we manually excluded frames captured when the car was stopped at a traffic light.

We first applied the feature matching process, including local verification with a scale and orientation check. After selecting the frames with 4 or more inlier matches, we obtained 37 frames (28 frames without and 9 frames with incorrect matches). We then applied global verification using frames with GPS data (739 out of 7,698 frames). Here, we experimentally set $n = 7$ and $l_{\mathrm{th}} = 100$ [m]. After 100 trials, the trial that derived the largest number of inlier frames selected 22 frames as inliers (19 frames without and 3 frames with incorrect matches) and 15 frames as outliers (9 frames without and 6 frames with incorrect matches). Figures 16 and 17 show example frames selected as inliers and outliers, respectively. As shown in Figure 16, the frames with incorrect matches were selected as inlier frames by global verification because the positions of the incorrect matches on the aerial image were close to the correct positions. Figures 15 and 18 show the estimated camera positions and horizontal position errors for each frame, respectively. Although the frames with incorrect matches still remained even when using two-stage geometric verification, the proposed method clearly reduced the accumulative errors. However, as can be seen in Fig. 18, the accumulative errors are still large around the 6,000th frame because there was only a small number of matches.

6. Discussion and Limitations

This section discusses the way of parameter setting and the limitations in the feature matching process. The proposed method has some parameters which should be manually determined. Generally, it is not easy to find the best parameters for individual dataset. However, it should be noted that, as shown

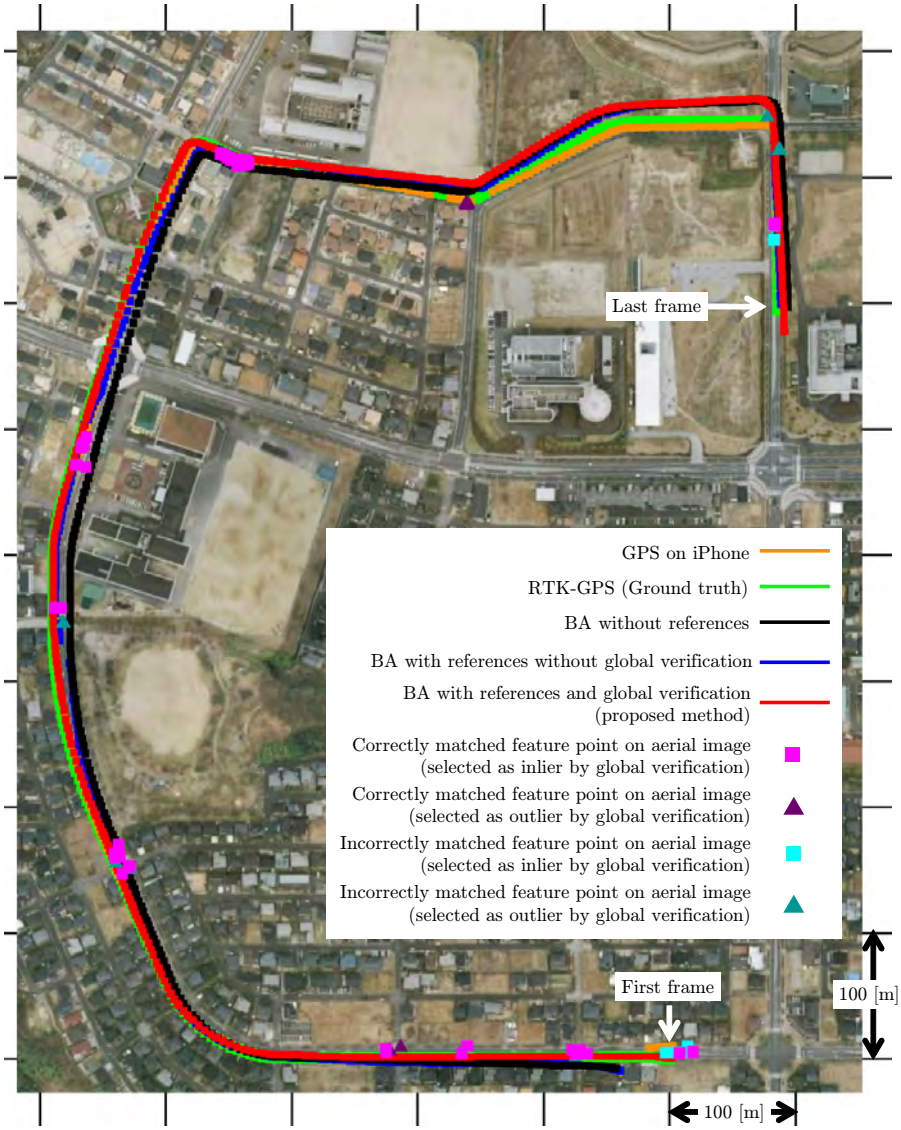
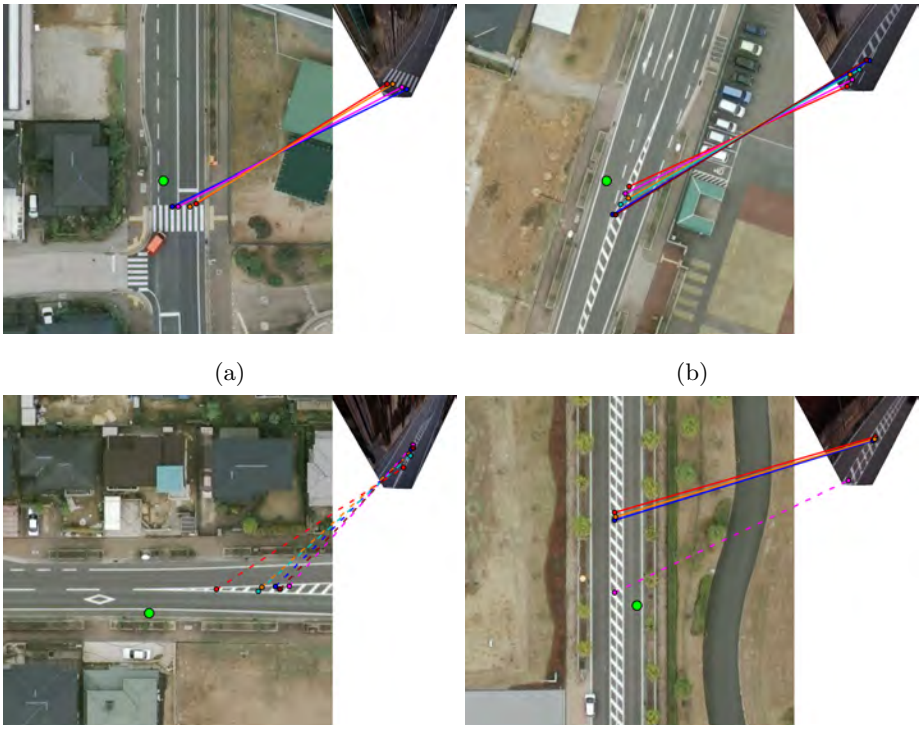


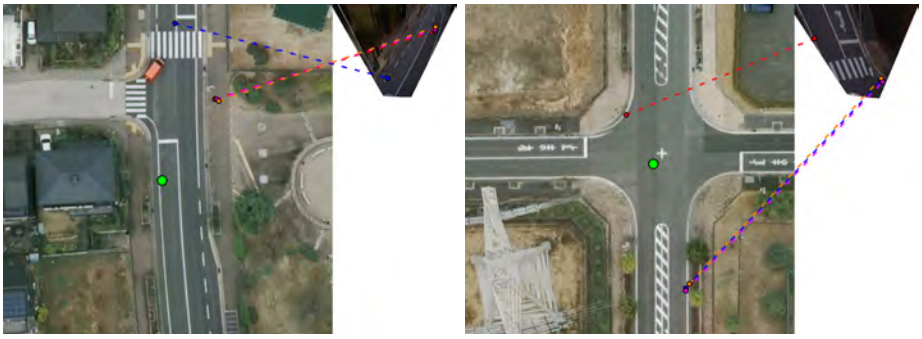
Figure 15: Experimental environment and results (Experiment 2).



(a) (b)

(c) (d)

Figure 16: Examples of frames selected as inliers by global verification (Experiment 2). The solid and dashed lines represent correct and incorrect matches, respectively.



(a) (b)

Figure 17: Examples of frames selected as outliers by global verification (Experiment 2). The dashed lines represent incorrect matches.

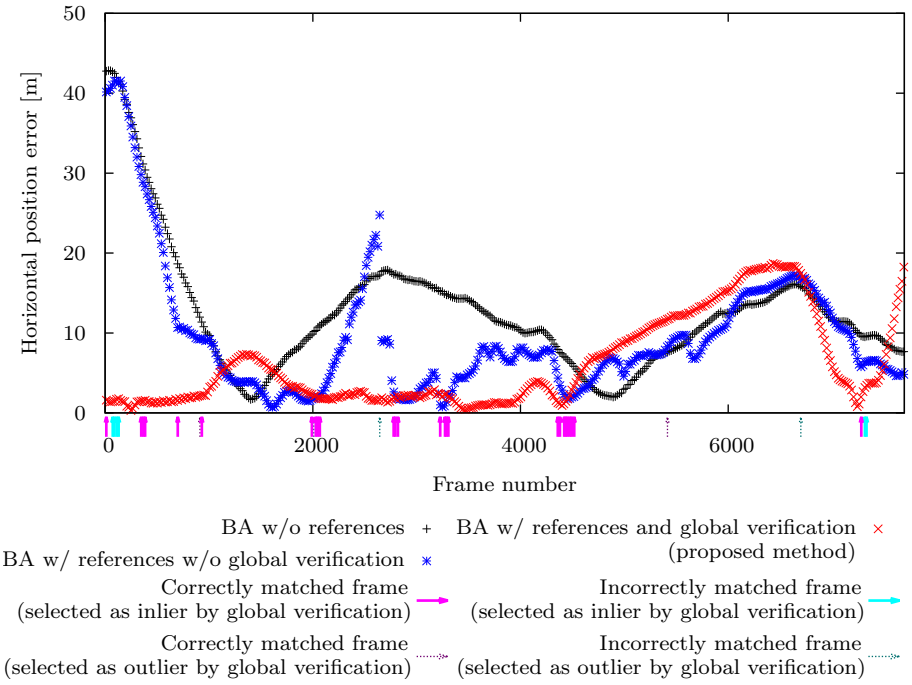


Figure 18: Horizontal position error in each frame (Experiment 2).

in Figs. 8, 11 and 12, most of parameters in the proposed pipeline have wide range of sweet spots where sub-optimal results can be obtained. We thus have employed common values for most of parameters in two different experiments despite characteristics of the datasets are quite different. These results imply that parameter values used in the experiments may be valid for other datasets. Parameters for which we did not use common values in the experiments are the number of samples n and the minimum distance between samples l_{th} used in the global geometric verification. These parameters depend on the number of matches, the ratio of incorrect matches, and the scale of input data, e.g. moving distance, distance between camera and feature point, and number of images. With analyzing various experimental data, automatic determination way for these two parameters is expected to be developed in the future.

One major limitation of the proposed method exists in the feature matching process. Since the proposed method projects a ground image to a plane that is

perpendicular to the gravity direction, it is not easy to find correct matches in situations where the ground is not level or not a plane. In fact, there exist a long slope and steps around the last frame of the dataset 1 (top right of Fig. 7) and thus our method could not find correct matches for these regions. Matches also cannot be obtained from texture-less ground such as roadways shown in Section 5.3. However, as shown in these experiments, even if there exist regions where the proposed method cannot obtain good matches, the proposed method can successfully reduce the accumulative errors only if there exist several regions where correct matches can be found. To find more good matches even for non-level ground, affine and/or perspective invariant features, such as ASIFT [34] and Ferns [35], can be used with finding better homography parameters in local verification process. However, it should be noted that simply applying these feature operators to our pipeline will drop the accuracy for many cases due to the increment of degrees of freedom in feature matching process. If the scene is expected to be level for most regions, standard feature matching operators should give better result. The development of effective way for employing affine/perspective invariant feature operators into the proposed pipeline is our next challenge for increasing the practicality.

7. Conclusion

In this paper, we proposed a method for removing accumulative errors in SfM using aerial images that are already available for many places around the world as external references. To achieve this, we proposed a new BA scheme that uses feature matches between the ground-view and aerial images. In order to discriminate the correct matches among unreliable matches, we introduced local and global geometric verification procedures provided by RANSAC. The local verification focuses on transformation between the aerial image and each ground-view image, considering in particular the orientation and the scale information extracted by a feature descriptor. The global verification focuses on the consistency of matches and poses for all the images through a sampling-based

BA. To the best of our knowledge, ours is the first method that uses aerial images as external references in BA. We confirmed experimentally that the proposed method is effective for estimating the camera poses of real video sequences taken in outdoor environments. However, the accumulative errors still remain when there are no available matches during a long period of time. To find matches in situations where the ground is not level, affine and/or perspective invariant features, such as ASIFT [34] and Ferns [35], can be used with homography as a geometric transformation in local verification. The proposed method requires several tens of hours for global geometric verification. To reduce computational time, BA with incorrect matches should be determined and discontinued in an early step of non-linear optimization.

Acknowledgements

This research was partially supported by JSPS Grant-in-Aid for Scientific Research Nos. 23240024 and 26330193.

References

[1] N. Snavely, S. M. Seitz, R. Szeliski, Modeling the World from Internet Photo Collections, Int. J. of Computer Vision 80 (2) (2008) 189–210.

[2] C. Wu, VisualSFM: A Visual Structure from Motion System, `http://ccwu.me/vsfm/` (2013).

[3] T. Sato, S. Ikeda, N. Yokoya, Extrinsic Camera Parameter Recovery from Multiple Image Sequences Captured by an Omni-directional Multi-camera System, in: Proc. European Conf. on Computer Vision, 2004, pp. 326–340.

[4] N. Fioraio, L. D. Stefano, Joint Detection, Tracking and Mapping by Semantic Bundle Adjustment, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2013, pp. 1538–1545.

[5] G. Bleser, H. Wuest, D. Stricker, Online Camera Pose Estimation in Partially Known and Dynamic Scenes, in: Proc. IEEE and ACM Int. Symp. on Mixed and Augmented Reality, 2006, pp. 56–65.

[6] P. Lothe, S. Bourgeois, E. Royer, M. Dhome, S. Naudet-Collette, Real-time Vehicle Global Localisation with a Single Camera in Dense Urban Areas: Exploitation of Coarse 3D City Models, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2010, pp. 863–870.

[7] D. Larnaout, S. Bourgeois, V. Gay-Bellile, M. Dhome, Towards Bundle Adjustment with GIS Constraints for Online Geo-localization of a Vehicle in Urban Center, in: Proc. Int. Conf. on 3D Imaging, Modeling, Processing, Visualization and Transmission, 2012, pp. 348–355.

[8] M. Tamaazousti, V. Gay-Bellile, S. Naudet Collette, S. Bourgeois, M. Dhome, NonLinear Refinement of Structure from Motion Reconstruction by Taking Advantage of a Partial Knowledge of the Environment, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2011, pp. 3073–3080.

[9] Y. Yokochi, S. Ikeda, T. Sato, N. Yokoya, Extrinsic Camera Parameter Estimation Based-on Feature Tracking and GPS Data, in: Proc. Asian Conf. on Computer Vision, 2006, pp. 369–378.

[10] M. Lhuillier, Incremental Fusion of Structure-from-Motion and GPS using Constrained Bundle Adjustments, IEEE Trans. on Pattern Analysis and Machine Intelligence 34 (12) (2012) 2489–2495.

[11] M. A. Brubaker, A. Geiger, R. Urtasun, Lost! Leveraging the Crowd for Probabilistic Visual Self-localization, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2013, pp. 3057–3064.

[12] O. Pink, F. Moosmann, A. Bachmann, Visual Features for Vehicle Localization and Ego-motion Estimation, in: Proc. IEEE Intelligent Vehicles Symposium, 2009, pp. 254–260.

[13] M. Bansal, K. Daniilidis, H. Sawhney, Ultra-wide Baseline Facade Matching for Geo-localization, in: Proc. European Conf. on Computer Vision, 2012, pp. 175–186.

[14] T.-Y. Lin, S. Belongie, J. Hays, Cross-view Image Geolocalization, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2013, pp. 891–898.

[15] S. Kim, S. DiVerdi, J. S. Chang, T. Kang, R. Iltis, T. Höllerer, Implicit 3D Modeling and Tracking for Anywhere Augmentation, in: Proc. ACM Symp. on Virtual Reality Software and Technology, 2007, pp. 19–28.

[16] K. Y. K. Leung, C. M. Clark, J. P. Huissoon, Localization in Urban Environments by Matching Ground Level Video Images with an Aerial Image, in: Proc. IEEE Int. Conf. on Robotics and Automation, 2008, pp. 551–556.

[17] H. Toriya, I. Kitahara, Y. Ohta, A Mobile Camera Localization Method using Aerial-view Images, in: Proc. IAPR Asian Conf. on Pattern Recognition, 2013, pp. 49–53.

[18] M. Noda, T. Takahashi, D. Deguchi, I. Ide, H. Murase, Y. Kojima, T. Naito, Vehicle Ego-localization by Matching In-vehicle Camera Images to an Aerial Image, in: Proc. Computer Vision in Vehicle Technology: From Earth to Mars, 2010, 10 pages.

[19] M. A. Fischler, R. C. Bolles, Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography, Communications of the ACM 24 (6) (1981) 381–395.

[20] D. G. Lowe, Distinctive Image Features from Scale-invariant Keypoints, Int. J. of Computer Vision 60 (2) (2004) 91–110.

[21] H. Kume, T. Sato, N. Yokoya, Sampling based Bundle Adjustment using Feature Matches between Ground-view and Aerial Images, in: Proc. Int. Conf. on Computer Vision Theory and Applications, Vol. 3, 2014, pp. 692–698.

[22] A. Angeli, D. Filliat, S. Doncieux, J.-A. Meyer, Fast and Incremental Method for Loop-closure Detection using Bags of Visual Words, IEEE Trans. on Robotics 24 (5) (2008) 1027–1037.

[23] H. Strasdat, A. J. Davison, J. Montiel, K. Konolige, Double Window Optimisation for Constant Time Visual SLAM, in: Proc. IEEE Int. Conf. on Computer Vision, 2011, pp. 2352–2359.

[24] B. Williams, M. Cummins, J. Neira, P. Newman, I. Reid, J. Tardós, A Comparison of Loop Closing Techniques in Monocular SLAM, Robotics and Autonomous Systems 57 (12) (2009) 1188–1197.

[25] A. Cohen, C. Zach, S. N. Sinha, M. Pollefeys, Discovering and Exploiting 3D Symmetries in Structure from Motion, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2012, pp. 1514–1521.

[26] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, R. Szeliski, Building Rome in a Day, in: Proc. IEEE Int. Conf. on Computer Vision, 2009, pp. 72–79.

[27] Q. Shan, R. Adams, B. Curless, Y. Furukawa, S. M. Seitz, The Visual Turing Test for Scene Reconstruction, in: Proc. Int. Conf. on 3D Vision, 2013, pp. 25–32.

[28] C. Forster, M. Pizzoli, D. Scaramuzza, Air-ground Localization and Map Augmentation Using Monocular Dense Reconstruction, in: Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 2013, pp. 3971–3978.

[29] C. Wu, B. Clipp, X. Li, J.-M. Frahm, M. Pollefeys, 3D Model Matching with Viewpoint-invariant Patches (VIP), in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2008, 8 pages.

[30] B. Zeisl, K. Köser, M. Pollefeys, Automatic Registration of RGB-D Scans via Salient Directions, in: Proc. IEEE Int. Conf. on Computer Vision, 2013, 8 pages.

[31] A. Criminisi, I. Reid, A. Zisserman, Single View Metrology, Int. J. of Computer Vision 40 (2) (2000) 123–148.

[32] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, P. Sayd, Generic and Real-time Structure from Motion using Local Bundle Adjustment, Image and Vision Computing 27 (8) (2009) 1178–1193.

[33] S. Agarwal, K. Mierle, Others, Ceres Solver, `https://code.google.com/p/ceres-solver/` (2013).

[34] J.-M. Morel, G. Yu, ASIFT: A New Framework for Fully Affine Invariant Image Comparison, SIAM J. on Imaging Sciences 2 (2) (2009) 438–469.

[35] M. Özuysal, M. Calonder, V. Lepetit, P. Fua, Fast Keypoint Recognition using Random Ferns, IEEE Trans. on Pattern Analysis and Machine Intelligence 32 (3) (2010) 448–461.