Sampling based Bundle Adjustment using Feature Matches between Ground-view and Aerial Images

Hideyuki Kume, Tomokazu Sato and Naokazu Yokoya
Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara, Japan

Keywords: Structure-from-Motion, Bundle Adjustment, Aerial Image, RANSAC.

Abstract: This paper proposes a new pipeline of Structure-from-Motion that uses feature matches between ground-view and aerial images for removing accumulative errors. In order to find good matches from unreliable matches, we newly propose RANSAC based outlier elimination methods in both feature matching and bundle adjustment stages. To this end, in the feature matching stage, the consistency of orientation and scale extracted from images by a feature descriptor is checked. In the bundle adjustment stage, we focus on the consistency between estimated geometry and matches. In experiments, we quantitatively evaluate performances of the proposed feature matching and bundle adjustment.

1 INTRODUCTION

Structure-from-Motion (SfM) is one of key technologies developed in computer vision field, and SfM has been used to achieve great works such as 'building Rome in a day' (Agarwal et al., 2009) and 'PTAM' (Klein and Murray, 2007). In SfM, one essential problem is accumulation of estimation errors for a long image sequence. Although many kinds of methods that reduce accumulative errors are proposed, SfM methods essentially cannot be free from accumulative errors unless some external references (e.g., like GPS, aerial images, and feature landmarks) are given.

The techniques that remove accumulative errors in the bundle adjustment (BA) stage of SfM using some external references are called as extended bundle adjustment (extended-BA), and earlier researches of extended-BA mainly focused on the combination of GPS and SfM (Lhuillier, 2012). In this paper, we employ the framework of extended-BA to the SfM problem that uses an aerial image as an external reference. In order to successfully use an aerial image as a reference of SfM, successful matching between aerial image and ground-view image is very important. To find good matches from unreliable matches, in addition to the use of GPS and gyroscope sensors embedded in most of recent smartphones, we newly use two methods: (1) RANSAC (Fischler and Bolles, 1981) based outlier elimination in the feature matching stage by focusing on consistency of orientation and scale extracted from images by feature descriptor like SIFT (Lowe, 2004) and (2) RANSAC based outlier elimination in the BA stage using consistency of estimated geometry and matches.

2 RELATED WORK

In order to reduce accumulative errors in SfM, loop closing techniques (Williams et al., 2009) are sometimes employed with the BA. Although these techniques can reduce accumulative errors, it is essentially difficult for the techniques that rely on only images to remove accumulative errors for a long sequence without the loop.

In order to reduce accumulative errors in general movement of camera, several kinds of external references such as GPS (Lhuillier, 2012) and road maps (Brubaker et al., 2013) are used with SfM. Lhuillier (Lhuillier, 2012) 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 parameters and reduce accumulative errors by updating parameters so as to minimize the energy function. However, the accuracy of this method is directly affected by errors of GPS positioning, which easily grow to the 10m level in urban areas. Brubaker et al. (Brubaker et al., 2013) proposed the method that uses community developed road maps. Although this

Kume H., Sato T. and Yokoya N..
Sampling based Bundle Adjustment using Feature Matches between Ground-view and Aerial Images.
DOI: 10.5220/0004850306920698
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 692-698
ISBN: 978-989-758-009-3

method can reduce accumulative errors by matching trajectory from SfM to road maps, there are ambiguities for some scenes such as straight roads or Manhattan worlds.

On the other hand, in order to estimate absolute camera positions and postures, some methods estimate camera parameters directly from references without SfM (Pink et al., 2009; Noda et al., 2010). Pink et al. (Pink et al., 2009) used aerial images as references and estimated camera parameters based on feature matching between input images and aerial images. However, it is not easy to find good matches for all the images of a long video sequence especially for the scenes where unique landmarks cannot be observed. Although Mills (Mills, 2013) proposed a robust feature matching procedure that compares orientation and scale of each matches with dominant orientation and scale identified by histogram analysis, it cannot work well when there exist a huge number of outliers. Noda et al. (Noda et al., 2010) relaxed the problem by generating mosaic images of the ground from multiple images for feature matching. However, accumulative errors are not considered in this work. Unlike previous works, we estimate relative camera poses for all the frames using SfM, and we remove accumulative errors by selecting correct matches between aerial image and ground-view image from candidates by employing a multi-stage RANSAC scheme.

3 FEATURE MATCHING BETWEEN GROUND-VIEW AND AERIAL IMAGES

In this section, we propose a robust method to obtain feature matches between ground-view and aerial images. As shown in Figure 1, the method is composed of (1) ground-view image rectification by Homography, (2) feature matching, and (3) RANSAC. Here, in order to achieve robust matching, we propose new criteria for RANSAC with consistency check of orientation and scale from a feature descriptor. It should be noted that matching for all the input frames are not necessary in our pipeline. Even if we can find only several candidates of matched frames, they can be effectively used as references in the BA stage.

3.1 Image Rectification by Homography

Before calculating feature matches using a feature detector and a descriptor, as shown in Figure 1, we rectify ground-view images so that texture patterns are similar to those of the aerial image. In most cases, aerial images are taken very far away from the ground and thus they are assumed to be captured by an orthographic camera whose optical axis is directed to gravity direction. In order to rectify ground-view images, we also assume that the ground-view images contain the ground plane whose normal vector is directed to gravity direction. Then, we compute Homography matrix using the gravity direction in camera coordinate system which can be estimated from the vanishing points of parallel lines or a gyroscope.

3.2 Feature Matching

Feature matches between rectified ground-view images and the aerial image are calculated. Here, we use GPS data corresponding to the ground-view images to limit searching area in the aerial image. More concretely, we select the region whose center is GPS position and its size is $l \times l$. In the experiment described later, l is set to 50 [m]. Feature matches are then calculated by a feature detector and a descriptor. We employ SIFT (Lowe, 2004) in the experiment because of its robustness for changes in scale, rotation and illumination.

3.3 RANSAC with Orientation and Scale Check

As shown in Figure 1, tentative matches often include many outliers. In order to remove outliers, we use RANSAC with consistency check of orientation and scale parameters.

For matches between rectified ground-view images and the aerial image, we can use the similarity transform which is composed of scale s, rotation θ and translation $\boldsymbol{\tau}$. In RANSAC procedure, we randomly sample two matches (minimum number to estimate similarity transform) to compute the similarity transform $(s, \theta, \boldsymbol{\tau})$. Here we count the number of inliers which satisfy

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

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. $\mathrm{R}(\theta)$ is the 2D rotation matrix with rotation angle θ and d_{th} is a threshold. After repeating random sampling process, the sampled matches with the largest number of inliers are selected.

The problem here is that the distance-based criterion above cannot successfully find correct matches when there exist a huge number of outliers. In order to achieve more robust matching, we modify the criterion of RANSAC by checking the consistency of

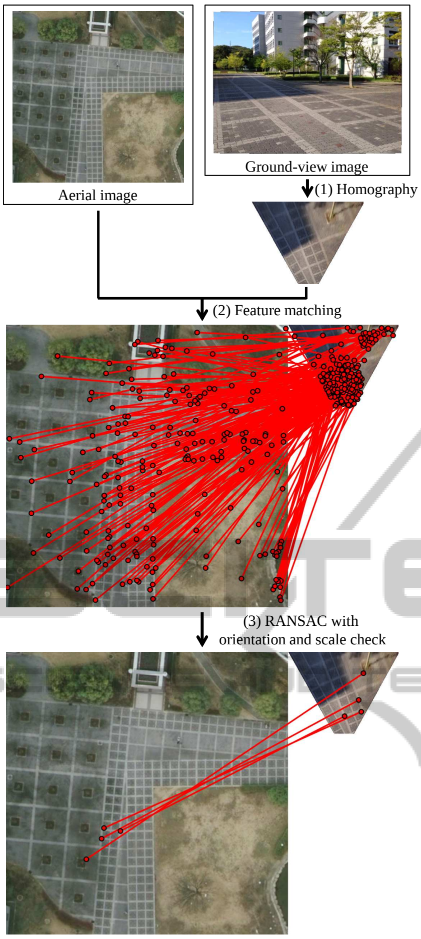


Figure 1: Flow of feature matching.

orientation and scale from a feature descriptor. Concretely, we count the number of inliers which simultaneously satisfy Equation (1) and the following two conditions.

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

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

where (s_{ak}, s_{gk}) and $(\theta_{ak}, \theta_{gk})$ are 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 thresholds for scale and angle, respectively.

4 SAMPLING BASED BUNDLE ADJUSTMENT

Even by the modified RANSAC proposed in previous section, it is not possible to remove all the incorrect matches in principle because there may exist repetitive and similar patterns, e.g., road signs in real environments. In order to overcome this difficulty, we also employ RANSAC for the BA stage by focusing on consistency between feature matches and estimated camera poses from images.

4.1 Definition of Energy Function

In order to consider the matches between ground-view and aerial images, here, an energy function is defined and minimized. The energy function E is defined by using reprojection errors for ground-view (perspective) images Φ and the aerial (orthographic) image Ψ as follows:

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

where $\boldsymbol{R}_i$ and $\boldsymbol{t}_i$ represent rotation and translation from world coordinate system to 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 weight 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 positions from GPS by 3D similarity transform. In the following, the energy associated with reprojection errors Φ and Ψ are detailed.

4.1.1 Reprojection Errors for Ground-view Images

The commonly used reprojection errors employ the pinhole camera model which cannot deal with projections from behind the camera. Projections from behind the camera often occur in BA with references due to dynamic movement of camera parameters by references. Here, instead of common squared distance errors on image plane, we employ the reprojection error by using angle of rays as follows:

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

$$\Phi_{ij} = \angle\left(\left(\frac{x_{ij}}{f_i}\right), \left(\frac{X_{ij}}{Z_{ij}}\right)\right)^2 + \angle\left(\left(\frac{y_{ij}}{f_i}\right), \left(\frac{Y_{ij}}{Z_{ij}}\right)\right)^2, \quad (6)$$

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

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 camera.

Here, we split the angular reprojection error into xz component and yz component because Jacobian matrix of E required by non-liner least squares method such as Levenberg-Marquardt method becomes simple. Especially, the first term of Φ_{ij} does not depend on the y component of $\boldsymbol{t}_i$ and the second term does not depend on the x component of $\boldsymbol{t}_i$ in this definition. We have experimentally confirmed that this splitting largely affects the performance of convergence.

4.1.2 Reprojection Errors for Aerial Image

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

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

where $\boldsymbol{M}$ is a set of frames in which feature matches between ground-view and aerial images are obtained. $\boldsymbol{A}_i$ is a set of feature points which are matched to the aerial image in the i-th frame. $\boldsymbol{a}_j$ is the 2D position of the j-th feature point in the aerial image. $\mathrm{pr}_{\mathrm{xy}}$ is a function that projects a 3D point onto xy plane (aerial image coordinate system).

4.2 RANSAC for Bundle Adjustment

RANSAC scheme is introduced in BA by using the consistency between feature matches and estimated camera poses from images. First, we randomly sample n frames from the candidates of matched frames and do BA using feature matches included in sampled frames, i.e., using a set of selected frames $\boldsymbol{M}'$ instead of $\boldsymbol{M}$ in Equation (8). Then we count the number of inlier frames which satisfy the following two conditions.

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

$$\angle\left(\boldsymbol{R}_i^{\mathrm{T}}(0,0,1)^{\mathrm{T}}, \boldsymbol{o}_i\right) < \beta_{\mathrm{th}}, \tag{10}$$

where α_{ij} is an angular reprojection error of the j-th feature point on aerial image coordinate system. $\boldsymbol{o}_i$ is the direction of optical axis in world coordinate system calculated from Homography and similarity transform estimated in the feature matching stage. α_{th} and β_{th} are thresholds. Here, α_{ij} is computed as follows:

$$\alpha_{ij} = \angle\left(\boldsymbol{a}_j - \mathrm{pr}_{\mathrm{xy}}(\boldsymbol{R}_i^{\mathrm{T}}\boldsymbol{t}_i), \mathrm{pr}_{\mathrm{xy}}\left(\boldsymbol{R}_i^{\mathrm{T}}(x_{ij}, y_{ij}, f_i)^{\mathrm{T}}\right)\right). \tag{11}$$

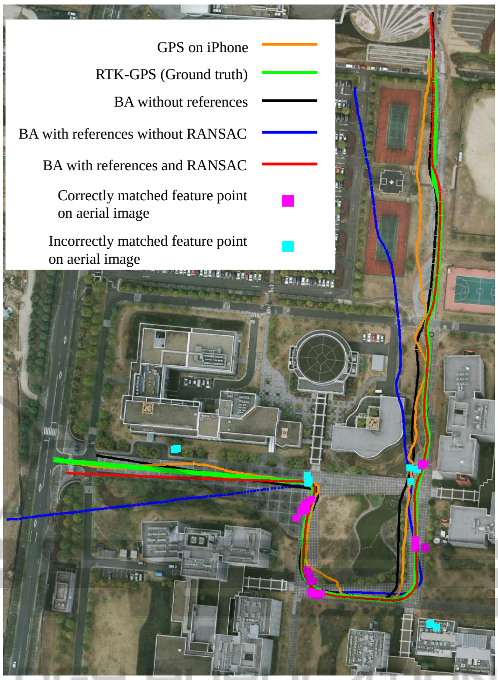

GPS on iPhone
RTK-GPS (Ground truth)
BA without references
BA with references without RANSAC
BA with references and RANSAC
Correctly matched feature point on aerial image
Incorrectly matched feature point on aerial image

Figure 2: Experimental environment and results.

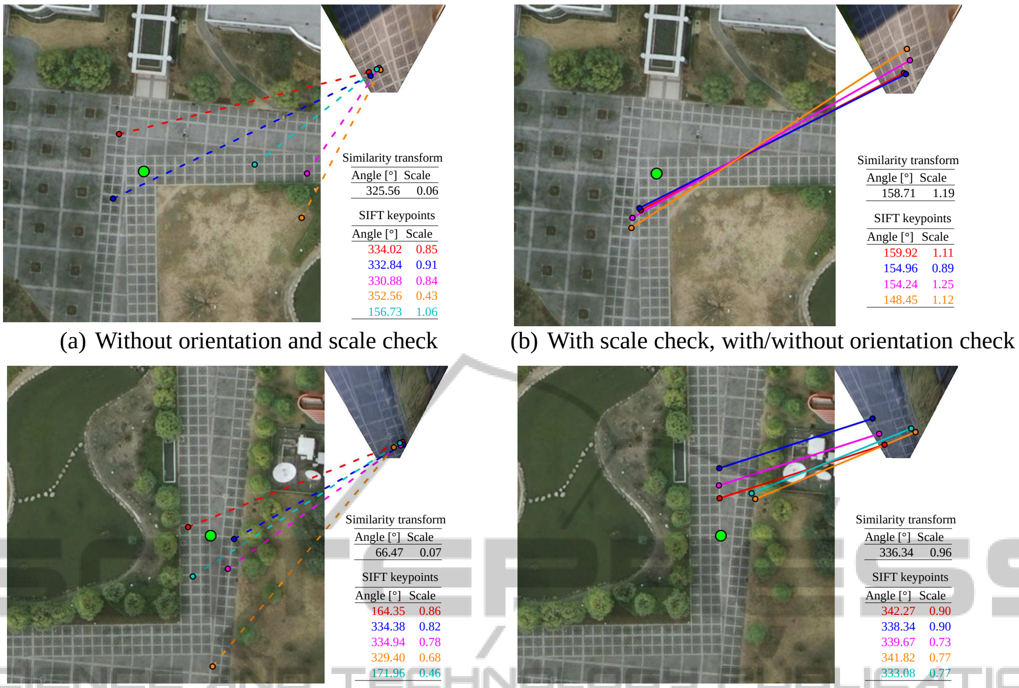
After repeating the random sampling process at given times, sampled frames with the largest number of inlier frames are selected. Finally, camera poses are refined by redoing BA using feature matches with selected inlier frames as references.

5 EXPERIMENTS

In order to validate the effectiveness of the proposed method, we quantitatively evaluated performances of sampling based BA as well as the feature matching process.

5.1 Experimental Setup

We used iPhone 5 (Apple) as a sensor unit including a camera, GPS and gyroscope. The camera captured video images (640 × 480 pixels, 2471 frames, 494 seconds). GPS and gyroscope measured position at 1 Hz and gravity direction for every frame, respectively. We also used RTK-GPS (Topcon GR-3, 1 Hz, accuracy of horizontal positioning is 0.03 [m]) to obtain the ground truth positions. Positions from GPS data were assigned temporally to the nearest frame. As the external reference, we downloaded the aerial image covering the area used in this experiment from Google Maps [maps.google.com] whose coordinate system is associated with the metric scale, i.e., 19.2 [pixel] = 1 [m]. Figure 2 shows an aerial image and GPS positions.



(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 3: Selected inliers for example images. Solid and dashed lines are correct and incorrect matches, respectively. Relative angle and scale of matched feature points are shown in bottom right table with corresponding lines' colors. Green points are ground truths of camera positions. RANSAC with/without orientation check for (b) and scale check for (d) gave the same results.

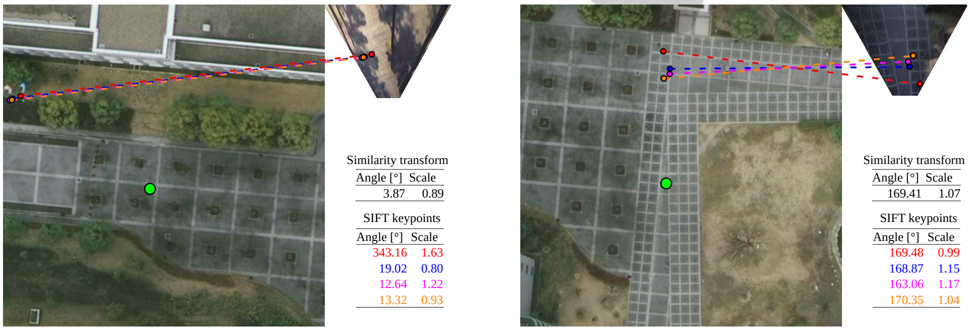


Figure 4: Examples of incorrect matches by RANSAC with orientation and scale check. The interpretations of symbols are the same as Figure 3.

In order to obtain initial values for BA, we employed VisualSFM (Wu, 2011) as a state-of-the-art implementation of SfM. For non-linear minimization, we used ceres-solver (Agarwal et al., 2013).

5.2 Result of Feature Matching

In this experiment, we evaluate the effectiveness of proposed feature matching process including RANSAC with scale and orientation check described in Section 3. Here, we have compared four types of RANSAC by enabling and disabling scale and orientation check. In order to count the number of correctly matched frames, we first selected frames which have four or more inlier matches after RANSAC. From these frames, we manually counted frames whose matches are correct. Here, we set $d_{\text{th}} = 2$ [pixel], $s_{\text{th}} = 2$ and $\theta_{\text{th}} = 40$ [°].

Table 1 shows rates of frames in which all the selected matches are correct. From this table, we can confirm that the rates are significantly improved by scale and orientation check. Figure 3 shows effects of scale and orientation check for sampled two images. In both cases, RANSAC without scale and orientation check could not select correct matches and proposed RANSAC with scale and orientation check could select correct matches. However, as shown in Figure 4, incorrect matches still remain even if we use both scale and orientation check.

5.3 Result of Bundle Adjustment

In this experiment, we evaluate the effectiveness of BA with RANSAC described in Section 4. In this stage, the frames with GPS data were sampled (650 of 2471 frames) and used in order to reduce computational time. As external references, we used frames and feature matches selected with orientation and scale check described in previous section. Here, we set $\omega = 10^{-5}$, $\alpha_{th} = 10.0$ [°], $\beta_{th} = 150.0$ [°], and $n = 2$.

We first evaluate the proposed RANSAC in terms of capability to select frames whose matches are correct. Here, as shown in Table 1, 10 of 14 frames have correct matches. We tested all the pairs of 14 frames as samples of RANSAC, and the number of inlier frames that are selected in each trial is checked. Figure 5 shows the number of trials and inlier frames derived by each trial. From this figure, we can see that the sampled frames without incorrect matches tend to increase the number of inlier frames. We also confirmed that the trials which derive the largest number of inlier frames successfully selected all of collect matches.

Next, we evaluate the accuracy of the proposed method by comparing the following methods.

- BA without references (Wu, 2011),
- BA with references without RANSAC which uses all the matches obtained by feature matching process,
- BA with references and RANSAC.

Since the BA without reference cannot estimate the metric scale, we fitted the camera positions estimated by SfM to the ground truths by similarity transform. Figures 2 and 6 show the estimated camera positions and horizontal position errors for each frame, respectively. From these results, it is confirmed that estimated camera positions by BA without references are affected by accumulative errors. The BA without RANSAC is affected by incorrect matches. The proposed BA with RANSAC can reduce the accumulative errors. It should be noted that, in the end of the sequence, the accumulative errors are still remained because there are no available matches.

Table 1: Rates of frames in which all the selected matches are correct. Number of frames in which all the selected matches are correct / number of frames which have four or more inlier matches are shown in bracket.

	w/ orientation check	w/o orientation check
w/ scale check	0.714 (10 / 14)	0.103 (9 / 87)
w/o scale check	0.134 (9 / 67)	0.005 (2 / 380)

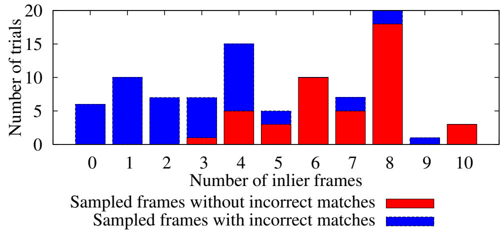


Figure 5: Number of trials and inlier frames derived by each trial.

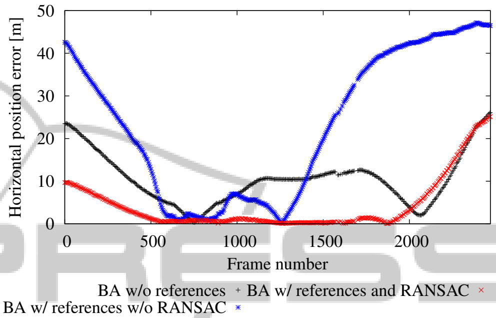


Figure 6: Horizontal position error in each frame.

6 CONCLUSIONS

In this paper, we have proposed a method to remove accumulative errors of SfM by using aerial images as external references that already exist for many places in the world. To this end, we have proposed SfM method that uses feature matches between ground-view and aerial images. In order to find correct matches from unreliable matches, we have introduced the new RANSAC schemes to both feature matching and bundle adjustment stages. In experiments, we have confirmed that the proposed method is effective for estimating camera poses of real video sequence taken in an outdoor environment. In the future, we will test the effectiveness of the proposed method in various environments including roadways.

ACKNOWLEDGEMENTS

This research was partially supported by JSPS Grant-in-Aid for Scientific Research No. 23240024 and “Ambient Intelligence” project granted from MEXT.

REFERENCES

Agarwal, S., Mierle, K., and Others (2013). Ceres solver. https://code.google.com/p/ceres-solver.

Agarwal, S., Snavely, N., Simon, I., Seitz, S. M., and Szeliski, R. (2009). Building Rome in a day. In *Proc. Int. Conf. on Computer Vision*, pages 72–79.

Brubaker, M. A., Geiger, A., and Urtasun, R. (2013). Lost! leveraging the crowd for probabilistic visual self-localization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3057–3064.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.

Klein, G. and Murray, D. (2007). Parallel tracking and mapping for small AR workspaces. In *Proc. Int. Symp. on Mixed and Augmented Reality*, pages 225–234.

Lhuillier, M. (2012). Incremental fusion of structure-from-motion and GPS using constrained bundle adjustments. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(12):2489–2495.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Computer Vision*, 60(2):91–110.

Mills, S. (2013). Relative orientation and scale for improved feature matching. In *Proc. IEEE Int. Conf. on Image Processing*, pages 3484–3488.

Noda, M., Takahashi, T., Deguchi, D., Ide, I., Murase, H., Kojima, Y., and Naito, T. (2010). Vehicle ego-localization by matching in-vehicle camera images to an aerial image. In *Proc. ACCV2010 Workshop on Computer Vision in Vehicle Technology: From Earth to Mars*, pages 1–10.

Pink, O., Moosmann, F., and Bachmann, A. (2009). Visual features for vehicle localization and ego-motion estimation. In *Proc. IEEE Intelligent Vehicles Symposium*, pages 254–260.

Williams, B., Cummins, M., Neira, J., Newman, P., Reid, I., and Tardós, J. (2009). A comparison of loop closing techniques in monocular SLAM. *Robotics and Autonomous Systems*, 57(12):1188 – 1197.

Wu, C. (2011). VisualSFM: A visual structure from motion system. http://ccwu.me/vsfm/.