IPSJ Transactions on Computer Vision and Applications Vol.5 134–142 (Aug. 2013)

[DOI: 10.2197/ipsjtcva.5.134]

Research Paper

6-DOF Camera Pose Estimation Using Reference Points on an Aerial Image without Altitude Information

Taiki Sekii [1,†1,a)] Tomokazu Sato [1,b)] Hideyuki Kume [1,c)] Naokazu Yokoya [1,d)]

Received: October 10, 2012, Accepted: April 11, 2013

Abstract: A new method for estimating a six-degrees-of-freedom camera pose for a ground-view image using reference points on an aerial image is presented. Unlike typical P*n*P problems, altitude information is not available for the reference points in our case. The camera pose is estimated by minimizing a cost function defined as the sum of squared distances between observed 2D positions of reference points on a ground-view image and corresponding lines that are projections of 3D vertical lines passing through 2D reference points on an aerial image. The accuracy of the proposed method is evaluated quantitatively in both simulation and real environments. The availability of the proposed method is demonstrated by generating AR images from aerial and ground-view images downloaded from Google Maps and Flickr.

Keywords: Extrinsic camera parameter estimation, perspective-*n*-points problem, Augmented Reality

1. Introduction

The estimation of a six-degrees-of-freedom (6-DOF) camera pose (position and posture) from a still image using reference points of known 3D positions is called a perspective-*n*-points (P*n*P) problem and solvers for the problem [1], [2], [3], [4], [5], [6], [7] are very useful for many types of applications such as image-based 3D reconstruction and augmented reality (AR). In order to estimate the camera pose in large outdoor environments, several kinds of references (pre-knowledge), e.g. street-view images [8], 3D CAD models [9], [10], feature landmark databases [11], [12], aerial images [13], [14], [15], have been used. In this paper, we focus on aerial images that already exist for many places in the world.

Aerial image-based methods [13], [14], [15] estimate the camera pose of a ground-view image from correspondences of reference points or lines on a ground-view image and aerial image. In most cases, aerial images are taken very far away from the ground and thus they are assumed to be captured with orthographic projection. Unlike other types of references, unfortunately, standard aerial images have no accurate altitude (height) information; thus, the common solvers for P*n*P problems [1], [3], [4], [5], [6], [7], which require 3D reference points, are not applicable. Common epipolar geometry estimators [4] for perspective imagery are also not useful in this case because of the combination of orthographic projection (aerial images) and perspective projection (ground-view images).

Most conventional works that use aerial images estimate camera poses by assuming that all of the altitudes of the reference points are the same or by reducing DOF to three when estimating the camera pose. Noda et al. [13] estimated the 3-DOF pose of a camera mounted on a vehicle in an environment where the ground surface was assumed to be flat plane that is parallel to the aerial image plane. They estimated the 2D camera position and direction by using the homography computed for an aerial image plane and ground plane in an input image. By combining the homography parameters for multiple images, they successfully determined the camera position even though there were few observable feature points from a single viewpoint. Although it is possible to decompose the homography parameters to a 6-DOF camera pose, the problem of assuming a flat and level ground still exists. Cham et al. [14] estimated the 3-DOF pose for an omni-directional camera by using the boundary of buildings extracted from a 2D map. In their work, the normal vectors of building sides were first computed by using the vanishing points detected from vertical edges of the buildings. A 3-DOF camera pose was then computed by using the geometric relationship of corresponding normal directions of the building sides on the 2D map.

A simple idea for 6-DOF pose estimation is to use P*n*P solvers, which can handle points on a plane [1], [3], [4], [5], [6], [7]. By using these solvers with known intrinsic camera parameters, a unique solution can be linearly computed by assuming that all of the altitudes of the reference points are the same, which means that the reference points are on a flat plane that is parallel to an aerial image plane. Although most common methods are useful for certain applications, e.g. vision-based robot localization in a flat environment, a method that can estimate camera poses even in non-flat environments is preferable. Yet another type of P*n*P solver estimates a 6-DOF camera pose using 3D reference

[1] Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara, Japan
[†1] Presently with Presently with Panasonic System Networks R&D Lab. Co., Ltd.
[a)] sekii.taiki@jp.panasonic.com
[b)] tomoka-s@is.naist.jp
[c)] hideyuki-k@is.naist.jp
[d)] yokoya@is.naist.jp

lines, e.g. wire frame models [16], [17], [18]. This approach is closely related to our approach because 2D reference points with arbitrary altitudes on an aerial image can be treated as 3D reference lines. Dhome et al. [16] proposed PnP solvers using lines under the condition that all of their 3D reference lines are not parallel to each other. Ramalingam et al. [18] provided minimal solutions for this problem. Such solvers are effective in the environments where many lines can be observed. On the other hand, Zhang et al. [19] proposed the epipolar geometry estimator for the combination of orthographic projection and weak-perspective projection. Although a 6-DOF camera pose can be linearly computed by minimizing object space errors derived from the epipolar constraint in this method, the performance of this method for the combination of strong-perspective (ground view) and orthographic (aerial) images has not been evaluated. In the experiment of this paper, we will show that our method, which minimizes the errors on the image space, gives more accurate camera pose than the method [19] for the combination of strong perspective and orthographic projection images.

In order to achieve 6-DOF camera pose estimation for a strong-perspective ground-view image using 2D reference points distributed on non-flat sloping ground on an aerial image, we newly define a cost function for this case. More concretely, camera parameters are determined to minimize the cost function, which is defined as the sum of new reprojection errors. We first estimate the initial pose by quasi-linearly minimizing the approximated reprojection errors, and newly defined reprojection errors are then minimized nonlinearly. Provided that intrinsic camera parameters are known, we can estimate the 2-DOF absolute camera position, 3-DOF absolute camera posture in the world coordinate system as defined in an aerial image, and 1-DOF relative height of the camera from reference points. In order to achieve more accurate and robust estimation in many real situations, we also examine the case in which the gravity direction is given, e.g. from vanishing points of parallel lines [20] or a gyro sensor. Although intrinsic camera parameters are assumed to be known in this article, we will show that we can generate good AR images when using Internet photographs for which there are no accurate intrinsic camera parameters available. It should also be noted that, although we do not care about the method for finding good corresponding pairs of feature points between images in this article, we will show that the proposed method can remove outliers by applying RANSAC scheme in the experiment.

2. Camera pose estimation for a ground-view image using reference points on an aerial image

As shown in Fig. 1, we define the cost function in this article as the sum of reprojection errors: these are distances between the observed 2D position $\mathbf{m}_i$ of reference points i on a ground-view image and the corresponding line g_i, which is the projection of the 3D vertical line passing through the 2D position $\mathbf{P}_i$ of reference point i in an aerial image. Although a 6-DOF camera pose can be estimated by being minimized, a good initial pose is required because this is a nonlinear least squares minimization problem. In order to estimate an initial pose, we also define another cost function: the sum of approximated reprojection errors on an aerial image, which can be minimized quasi-linearly.

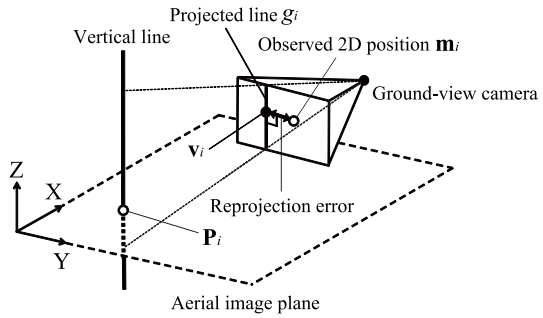


Fig. 1 Reprojection error for reference point without altitude information.

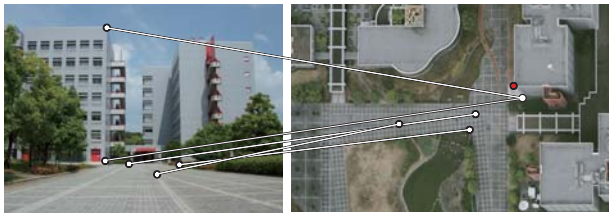


Fig. 2 Corresponding reference points between a ground-view image (left) and an aerial image (right).

A camera pose for ground view image is estimated by the following steps.

(1) Reference points on a ground-view image and aerial image are matched.
(2) An initial 5-DOF camera pose excluding the altitude is estimated by quasi-linear minimization of the approximated reprojection errors.
(3) The 5-DOF camera pose is refined by nonlinear minimization of the reprojection errors.
(4) The relative altitude for each reference point (remaining 1-DOF) is computed from the estimated 5-DOF camera pose.

In this article, we define the $X-Y$ plane in the world coordinate system as being on the horizontal plane of the aerial image as shown in Fig. 1, and the Z axis as the altitude direction. The intrinsic camera parameters are assumed to be known, and the corresponding pairs of reference points in step 1 are assumed to be given. Fig. 2 shows an example of manually matched points between a ground-view image and aerial image. We assume that the aerial image is taken very far away from the ground. It provides an orthographic top-down view of a scene where the camera model can be an affine projection, and it is generated by projecting the 3D scene onto the horizontal ground surface. Points that do not exist on the ground surface are obliquely projected onto the aerial image as shown in Fig. 2 (right). Such points, e.g. on top of a building in a ground-view image are assumed to match the corresponding positions on the ground surface in the aerial image in our case. For example, in Fig. 2, the reference point at the top of the building in the ground-view image (left) should not be matched to the red point on the aerial image (right) but to the point at the bottom of the building in the aerial image. Through this approach, even if an aerial image is captured obliquely, corresponding points that are not on the ground can also be used as reference points. In the following sections, two cost functions are

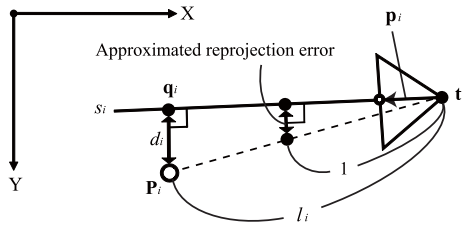


Fig. 3 Parameters for an aerial image plane.

defined, and their minimization processes are then detailed.

2.1 Cost functions

2.1.1 Reprojection error

As shown in Fig. 1, we define the cost function E_v, which is a sum of reprojection errors, as follows:

$$E_v = \sum_{i=1}^{n} |\mathbf{m}_i - \mathbf{v}_i|^2, \tag{1}$$

where $\mathbf{m}_i$ is the observed 2D position of reference point i on the ground-view image, $\mathbf{v}_i$ is the point closest to $\mathbf{m}_i$ on the projected line g_i, and n is the number of corresponding pairs.

2.1.2 Approximated reprojection error

In order to estimate an initial pose, we define the additional cost function E_a which is the sum of approximated reprojection errors on an aerial image. As shown in Fig. 3, E_a is defined as the weighted sum of squares of d_i representing the distance between a 2D position $\mathbf{P}_i = (X_i, Y_i)^{\mathrm{T}}$ of reference point i and a 2D line s_i, which is a projection of the ray passing through the projection center $\mathbf{t}$ of the ground-view camera and observed position of reference point i on the ground-view image. E_a is defined as follows:

$$E_a = \sum_{i=1}^{n} a_i^2,\ a_i^2 = \frac{d_i^2}{l_i^2},\ d_i = |\mathbf{P}_i - \mathbf{q}_i|,\ l_i = |\mathbf{P}_i - \mathbf{t}|, \tag{2}$$

where $\mathbf{q}_i$ is the closest 2D position of $\mathbf{P}_i$ on s_i. Supposing that the rotation matrix from the camera coordinate system (the 3D coordinate system which has its origin at the projection center) to the world coordinate system is defined as $\mathbf{R} = \left(\mathbf{r}_1^{\mathrm{T}}, \mathbf{r}_2^{\mathrm{T}}, \mathbf{r}_3^{\mathrm{T}}\right)^{\mathrm{T}}$, and the camera position on the $X-Y$ plane is $\mathbf{t} = (t_1, t_2)^{\mathrm{T}}$, we obtain the following equations based on the orthogonality and collinearity conditions of vectors.

$$(\mathbf{q}_i - \mathbf{P}_i) \cdot (\mathbf{q}_i - \mathbf{t}) = 0, \tag{3}$$

$$(\mathbf{q}_i - \mathbf{t}) = \lambda_i (\mathbf{p}_i \cdot \mathbf{r}_1, \mathbf{p}_i \cdot \mathbf{r}_2)^{\mathrm{T}}, \tag{4}$$

where $\cdot$ indicates the inner product of two vectors, λ_i is a parameter for each i, $\mathbf{p}_i$ is a 3D vector from the projection center to the observed position of the reference point i on the ground-view image in the camera coordinate system, and $(\mathbf{p}_i \cdot \mathbf{r}_1, \mathbf{p}_i \cdot \mathbf{r}_2)^{\mathrm{T}}$ is the projected 2D vector onto the $X-Y$ plane of the 3D vector $\mathbf{p}_i$. From Eqs. (2) to (4), we obtain

$$a_i = l_i^{-1} w_i \{\mathbf{p}_i \cdot (t_2 \mathbf{r}_1 - t_1 \mathbf{r}_2) - Y_i (\mathbf{p}_i \cdot \mathbf{r}_1) + X_i (\mathbf{p}_i \cdot \mathbf{r}_2)\}, \tag{5}$$

$$w_i = \left\{(\mathbf{p}_i \cdot \mathbf{r}_1)^2 + (\mathbf{p}_i \cdot \mathbf{r}_2)^2\right\}^{-\frac{1}{2}}. \tag{6}$$

2.2 Estimation of 6-DOF camera pose

In this section, we first detail the minimization process of each cost function to estimate a 5-DOF pose. We then describe the method for determining the relative altitude of each reference point.

2.2.1 Estimation of initial 5-DOF camera pose

The cost function E_a is minimized to obtain an initial 5-DOF camera pose, which is used to minimize E_v. First, Eq. (5) is unified about n corresponding pairs as follows:

$$(a_1, a_2, \cdots, a_n)^{\mathrm{T}} = \mathbf{W}\mathbf{A}\mathbf{x}, \tag{7}$$

$$\mathbf{W} = \begin{pmatrix} l_1^{-1} w_1 & \cdots & 0 \\ \vdots & l_2^{-1} w_2 & \vdots \\ & & \ddots \\ 0 & \cdots & l_n^{-1} w_n \end{pmatrix}, \tag{8}$$

$$\mathbf{A} = \begin{pmatrix} \mathbf{p}_1^{\mathrm{T}} & -Y_1 \mathbf{p}_1^{\mathrm{T}} & X_1 \mathbf{p}_1^{\mathrm{T}} \\ \mathbf{p}_2^{\mathrm{T}} & -Y_2 \mathbf{p}_2^{\mathrm{T}} & X_2 \mathbf{p}_2^{\mathrm{T}} \\ \vdots & \vdots & \vdots \\ \mathbf{p}_n^{\mathrm{T}} & -Y_n \mathbf{p}_n^{\mathrm{T}} & X_n \mathbf{p}_n^{\mathrm{T}} \end{pmatrix}, \tag{9}$$

$$\mathbf{x} = (t_2 \mathbf{r}_1 - t_1 \mathbf{r}_2, \mathbf{r}_1, \mathbf{r}_2)^{\mathrm{T}}, \tag{10}$$

where $\mathbf{W} \in \mathbb{R}^{n \times n}$ in Eq. (8) is a diagonal weighting matrix, $\mathbf{A} \in \mathbb{R}^{n \times 9}$ in Eq. (9) is composed of known parameters, and $\mathbf{x} \in \mathbb{R}^{9 \times 1}$ is an unknown parameter vector. From Eq. (7), $E_a = |\mathbf{W}\mathbf{A}\mathbf{x}|^2$. If all of the parameters of $\mathbf{W}$ are known and $n \geq 8$, $\mathbf{x}$ can be determined by linearly minimizing $|\mathbf{W}\mathbf{A}\mathbf{x}|^2$.

However, there are two problems with minimizing $|\mathbf{W}\mathbf{A}\mathbf{x}|^2$. First, $\mathbf{W}$ is comprised of unknown parameters. Second, $\mathbf{x}$ is comprised of the products of the unknown parameters. For the former problem, we first initialize $\mathbf{W}$ as an identity matrix and iteratively update it by solving $\mathbf{x}$. For the latter problem, after determining $\mathbf{r}_1$ and $\mathbf{r}_2$ by orthonormalizing them, t_1 and t_2 are determined by treating $\mathbf{r}_1$ and $\mathbf{r}_2$ as constants. By iterating these steps, $(\mathbf{r}_1, \mathbf{r}_2, t_1, t_2)$ are determined quasi-linearly. After a certain number of iterations, $\mathbf{r}_3$ is computed from $\mathbf{r}_1$ and $\mathbf{r}_2$ based on the orthonormality condition of the basis vectors. In this minimization, an outlier detection scheme, e.g. RANSAC [1], can be incorporated.

It should be noted that when all of the reference points are observed on the same line in a ground-view image, we cannot estimate the camera pose due to the rank-deficiency of the matrix $\mathbf{A}$. In Section 3, we examine another scenario where the gravity direction is assumed to be known to avoid this critical condition.

2.2.2 Refinement of 5-DOF camera pose

From the initial 5-DOF camera pose estimated in the previous step, the cost function E_a is first minimized nonlinearly using the Levenberg-Marquardt algorithm because all of the camera parameters are not simultaneously optimized in the previous step. Next, the reprojection error E_v defined in Eq. (1) is minimized to estimate a 5-DOF camera pose in the same manner.

2.2.3 Computation of relative altitude

Relative altitude $\hat{t}_{3i}$ for each reference point i from the camera (remaining 1-DOF) is calculated from the estimated 5-DOF camera pose as

$$\hat{t}_{3i} = \lambda_i(\mathbf{p}_i \cdot \mathbf{r}_3), \tag{11}$$

where λ_i is computed using Eqs. (2) to (4) as follows:

$$\lambda_i = w_i^2 \left|(\mathbf{p}_i \cdot \mathbf{r}_1)(X_i - t_1) + (\mathbf{p}_i \cdot \mathbf{r}_2)(Y_i - t_2)\right|. \tag{12}$$

3. Camera pose estimation using a given gravity direction

As the distribution of observed positions of reference points on a ground-view image approaches linear distribution (discussed in Section 2.2), the pose estimation using the method described in Section 2.2 becomes unstable. In practical situations, observed points on the ground surface are often distributed linearly and horizontally on a ground-view image and come close to the critical condition. In this case, especially, the gravity direction of the camera becomes unstable because the reprojection error becomes insensitive to the changes of the camera posture except for the yaw-angle. If the gravity direction is given from other sources, the estimation can be kept stable even in the critical condition. The following sections describe the method for camera pose estimation where the gravity direction is given.

3.1 Gravity direction estimation

The gravity direction can be assumed to be known in some cases, e.g. using the vanishing points of parallel lines [20] or a gyro sensor. Image-based estimation using vanishing points can be used for an image taken in an environment where vertical lines on building structures are visible. For the other approach, most recent smartphones with camera units have gyro sensors embedded inside that provide accurate gravity direction. Even a cheap gyro sensor provides the direction of gravity with an accuracy of about 0.5°. In the experiments presented in Section 4, we will show that camera pose can be accurately estimated by using an estimated gravity direction from an image.

3.2 6-DOF camera pose estimation using a given gravity direction

When the gravity direction is given, 2-DOF (pitch and roll) of the camera posture can be determined. In this case, the number of unknown parameters for cost functions E_a and E_v is reduced to three. We estimate a 3-DOF camera pose in the same manner as the method described in Section 2.

We first redefine the parameters required to compute the cost function E_a using the given unit vector $\mathbf{g}_c$ of gravity direction in the camera coordinate system:

$$\mathbf{g}_c = \mathbf{R}^{\mathrm{T}}\mathbf{g}, \tag{13}$$

where $\mathbf{g} = (0, 0, -1)^{\mathrm{T}}$ is a unit vector of gravity direction in the world coordinate system. From Eq. (13), we obtain

$$\mathbf{r}_3 = -\mathbf{g}_c^{\mathrm{T}}. \tag{14}$$

From the orthonormality condition of the rotation matrix $\mathbf{R}$, $\mathbf{r}_2$ is defined as follows:

$$\mathbf{r}_2 = \mathbf{r}_3 \times \mathbf{r}_1. \tag{15}$$

Substituting Eqs. (14) and (15) into Eq. (5), the parameters to compute E_a are redefined as follows:

$$a_i = l_i^{-1} w_i \{\mathbf{p}_i \cdot (t_2\mathbf{r}_1 - t_1(\mathbf{r}_3 \times \mathbf{r}_1)) + \mathbf{b}_i \cdot \mathbf{r}_1\}, \tag{16}$$

$$\mathbf{b}_i = \begin{pmatrix} X_i(r_{32} + q_i r_{33}) - Y_i p_i \\ X_i(p_i r_{33} - r_{31}) - Y_i q_i \\ X_i(q_i r_{31} - p_i r_{32}) - Y_i \end{pmatrix}^{\mathrm{T}}, \tag{17}$$

$$\mathbf{p}_i = (p_i, q_i, 1)^{\mathrm{T}}, \tag{18}$$

$$\mathbf{A} = \begin{pmatrix} \mathbf{p}_1^{\mathrm{T}} & \mathbf{b}_1 \\ \mathbf{p}_2^{\mathrm{T}} & \mathbf{b}_2 \\ \vdots & \vdots \\ \mathbf{p}_n^{\mathrm{T}} & \mathbf{b}_n \end{pmatrix}, \tag{19}$$

$$\mathbf{x} = (t_2\mathbf{r}_1 - t_1(\mathbf{r}_3 \times \mathbf{r}_1), \mathbf{r}_1)^{\mathrm{T}}. \tag{20}$$

When n is five or more, the initial 5-DOF camera pose can be determined by the method detailed in Section 2.2.1 using these equations. Then, E_a and E_v are minimized nonlinearly from the initial camera pose. The relative altitude can also be computed from the estimated 3-DOF and given 2-DOF.

4. Experiments

In order to evaluate the performance of the proposed methods in non-flat environments, the accuracies of the following four methods are compared quantitatively in both simulation and real environments.

P4P2D: Conventional PnP solver for a planar surface.
EPI2D: Zhang's epipolar geometry estimator [19] that uses weak-perspective assumption.
P8P2D: Proposed method with unknown gravity direction.
P5P2D: Proposed method with known gravity direction.

4.1 Quantitative evaluation in simulation environment

4.1.1 Setting

In order to compare the accuracies of the methods in various situations, we have used the simulation environment shown in Fig. 4, where reference points are randomly spread inside a cuboid using variable parameters α, β and γ, which are the height, the depth and the width of the cuboid, respectively. The camera (Image resolution: 1280×865 (px), focal length: 885 (px), FoV: $71.7° \times 44.3°$) is set at the position shown in Fig. 4, and the directions of basis vectors are set to be the same as those of the world coordinate system. In this experiment, 12 reference points are randomly spread inside the cuboid, and reference points are projected onto the image plane with quantization errors and the Gaussian noise of the standard deviation ω (px). The projected positions of the reference points on X-Y plane (positions on aerial image) and the perspectively projected positions of the reference points on the image with Gaussian noise (positions on ground-view image) are used as inputs for each method. For P4P2D, all of the altitudes of the 2D reference points are treated as 0 for input. The gravity direction for P5P2D is generated from the ground truth by adding the Gaussian noise of the standard deviation $\sigma = 0.0, 1.0$ (degree). It should be noted that α and β in this experiment denote the level of discrepancy from the flat ground assumption, and the weak-perspective assumption, respectively.

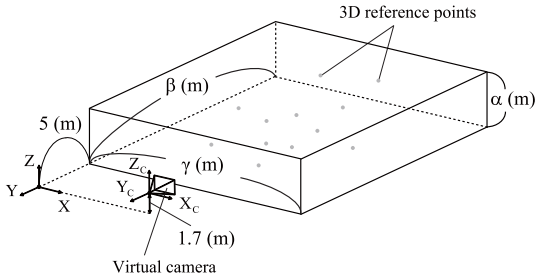
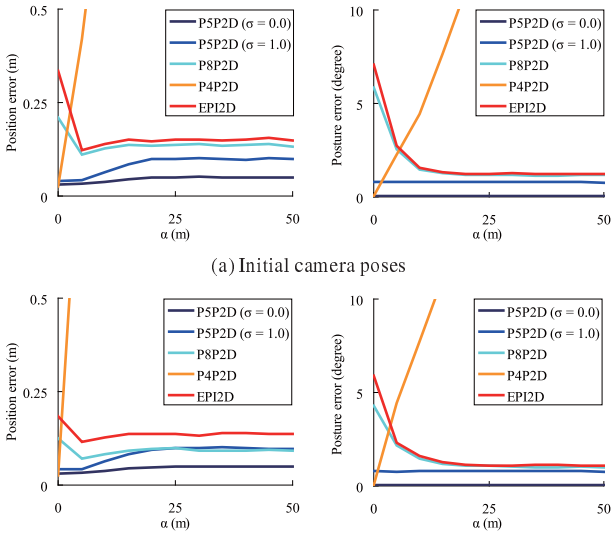


Fig. 4 A simulation environment.



(a) Initial camera poses

(b) Refined camera poses

Fig. 5 Average pose errors for variable α ($\beta = 50, \gamma = 20, \omega = 1.0$).

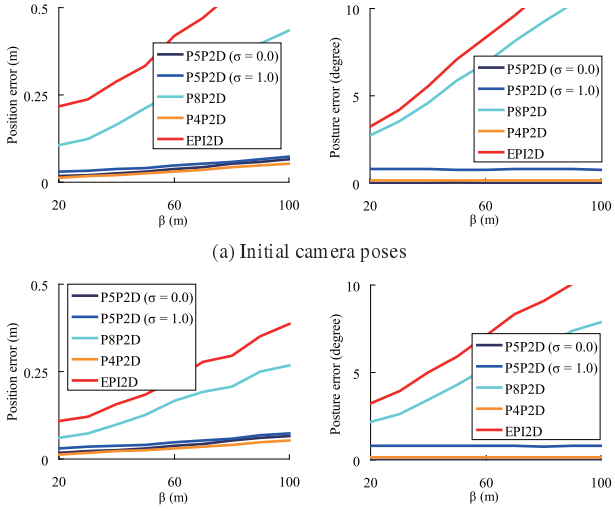
4.1.2 Results

We have evaluated the accuracy for both (a) initial camera poses without minimization of reprojection errors and RANSAC, and (b) refined camera poses with RANSAC. Fig. 5, 6, 7 and 8 show the position and posture errors for variable α, β, γ, ω, respectively. These errors are computed by comparing them with the ground truth over 1000 trials. Here, the position error of the camera is computed as the 2D distance between camera position and the ground truth on the aerial image plane because absolute altitude is not available in this experiment. The posture error indicates the angle of the Y_C axis between that estimated and the ground truth.

Fig. 5 shows the result for variable α (height). Here, $\alpha = 0$ means that all of the observed points exist on a flat surface ($X - Y$ plane). As we can see in this figure, errors of P4P2D are drastically increased as α becomes higher. In the case α is large, the flat surface assumption used in P4P2D is violated.

Fig. 6 shows the result for variable β (depth). We can confirm that P8P2D gives us better results than that by EPI2D when the distribution of the depth of the reference points becomes large. Fig. 7 shows the results for variable γ (width). We can see that wider point distribution is better for pose estimation by E8P2D and EPI2D.

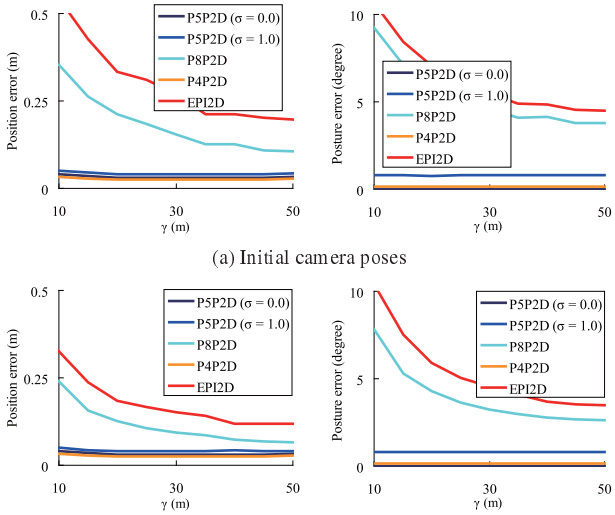
Fig 8 shows the result for variable noise level ω. Even for larger noise level, the proposed method achieves better performance than EPI2D. From all the results, it can be confirmed that P5P2D provides the most stable and accurate camera poses com-



(a) Initial camera poses

(b) Refined camera poses

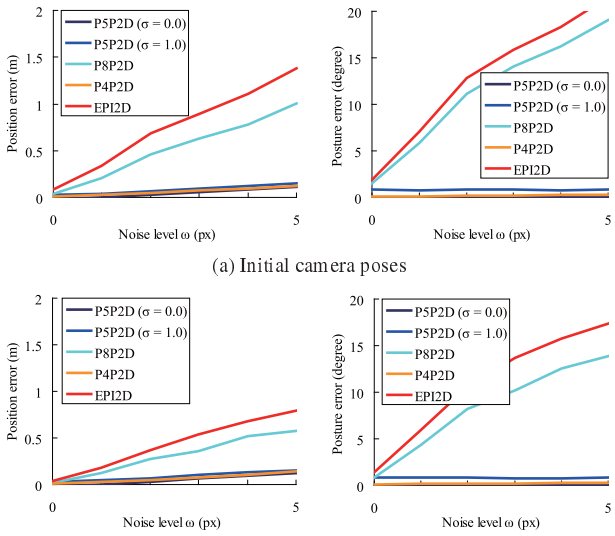
Fig. 6 Average pose errors for variable β ($\alpha = 0, \gamma = 20, \omega = 1.0$).



(a) Initial camera poses

(b) Refined camera poses

Fig. 7 Average pose errors for variable γ ($\alpha = 0, \beta = 50, \omega = 1.0$).



(a) Initial camera poses

(b) Refined camera poses

Fig. 8 Average pose errors for variable ω ($\alpha = 0, \beta = 50, \gamma = 20$).

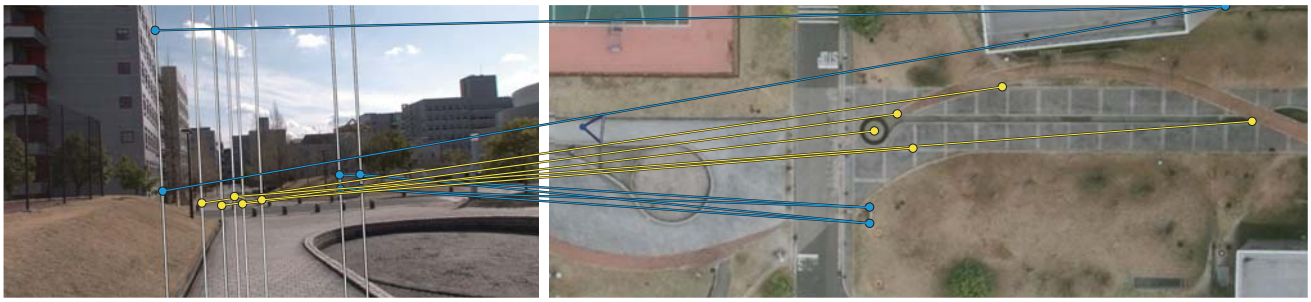


Fig. 9 An example of given point correspondences between a ground-view image (left) and an aerial image (right). The yellow and blue points indicate the points in S_{2D} and S_{3D}, respectively. The blue triangle indicates a camera pose estimated by P5P2D using given points. White lines in the left image are projected lines of reference points with arbitrary altitudes using the estimated camera pose.

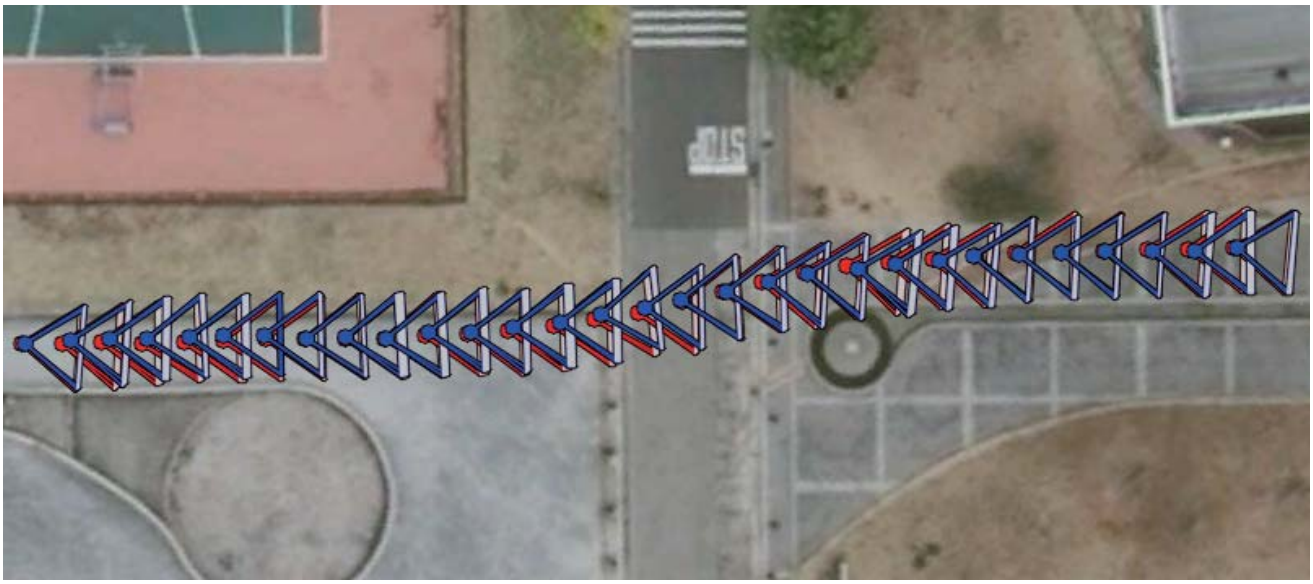


Fig. 10 Camera poses estimated by P5P2D (blue) and ground truth (red).

pared to those produced by others for various point distributions of reference points and noise level. From these observations, we can conclude that our method has a clear advantage for camera pose estimation especially when reference points are widely distributed in a 3D space and they are detected with large errors.

4.2 Quantitative evaluation in real environment

4.2.1 Setting

In this experiment, we have evaluated performance of the compared methods using 30 ground-view images from a dataset of Campus Package 02 on TrakMark [21] [trakmark.net] which is publicly available on the Internet, as input images of a real environment. The reference camera pose of each image, which are included in this package, is used as the ground truth for this experiment. In addition to this dataset, we have downloaded the aerial image covering the area of this dataset from Google Maps [maps.google.com]. In order to evaluate estimated camera poses quantitatively, we have aligned the coordinates of the aerial image and the reference points in the dataset using a transformation matrix computed by manually specifying reference points on the aerial image. For this dataset, we first manually found corresponding points between the ground-view images and 2D reference points on the aerial image. We then classified reference points into two groups: S_{2D} includes points on the ground surface, and S_{3D} includes the other points. Each method is tested using both S_{2D} and $S_{2D} \cup S_{3D}$ as input. Fig. 9 shows an example of the given corresponding points between the ground-view and aerial images.

In this experiment, gravity directions for P5P2D are estimated from the images using the method proposed by Criminisi et al. [20]: the vertical lines of building edges are specified manually for every ground-view image. The average error of the estimated gravity directions (from ground truth) was 0.36°. We have also tested one extension here for more stable estimation: when two or more points are given on the same line of the vertical edge in the image, we automatically generate additional points along this line where all points correspond to the one reference point on the aerial image. Here, RANSAC and the refinement process are used to remove outliers and to minimize reprojection errors for all the compared methods.

4.2.2 Results

Fig. 10 compares the ground truth and camera poses estimated by P5P2D with S_{2D} drawn on the aerial image. We can confirm that the camera poses by P5P2D are correctly estimated for this

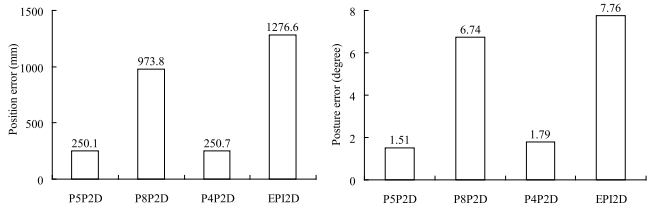


Fig. 11 Comparison of average errors of estimated position and posture for the dataset S_{2D}.

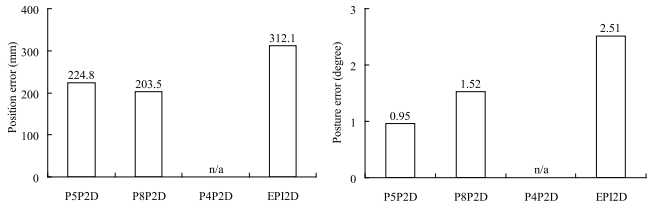


Fig. 12 Comparison of average errors of estimated position and posture for the dataset $S_{2D} \cup S_{3D}$.

dataset. The average errors of the position and posture estimated by each method are shown in Fig. 11 and 12 for the dataset S_{2D} and $S_{2D} \cup S_{3D}$, respectively. Although P4P2D provides a reasonable estimate for the camera position of dataset S_{2D}, P5P2D gives us more accurate results due to the slightly sloped ground of the target environment. For the dataset $S_{2D} \cup S_{3D}$, P4P2D could not provide valid results due to the violation of the flat surface assumption. Although P8P2D is not accurate for S_{2D}, P8P2D obtains more accurate camera poses for $S_{2D} \cup S_{3D}$, where the points are widely distributed on the image. P5P2D which uses estimated gravity direction gives stable and accurate camera poses for both datasets. In this experiment for EPI2D, we found that the RANSAC automatically removed sets of randomly selected reference points in which points are widely distributed for depth direction. This means that selected feature points were adjusted so that EPI2D can work with weak perspective assumption. However, the number of available feature points is reduced in this case, and it results in worse accuracy than that produced by the P8P2D.

In this experiment, average computational costs of P8P2D and EPI2D with PC(Core i5-2467M 2.3GHz) were 30.9 (ms/image) and 16.0 (ms/image), respectively. Although our method needs slightly more time for camera pose estimation than EPI2D, this cost will not be a bottleneck of actual AR applications.

5. Demonstration

In this section, we demonstrate the availability of the proposed method for landscape simulation. First, we have downloaded aerial images from Google Maps and Internet photographs from Flickr [flickr.com] and Google [google.com]. CG objects aligned to the aerial image were downloaded from Google 3D-Galerie [sketchup.google.com/3dwarehouse]. We have estimated the camera pose for each photograph by P8P2D using manually matched reference points, and the AR images are generated using estimated camera poses. Here, the relative altitude between the camera and CG object is estimated by manually selecting a reference point existing on the ground surface. Because accurate intrinsic camera parameters of each photograph are not available, we estimate the field of view and projection center by the exhaustive search for these parameters and selecting the one with the smallest reprojection error. In these landscape simulations, lens distortion is not considered. As shown in Fig. 13, we can successfully generate AR images for many places in the world using only images.

6. Conclusion

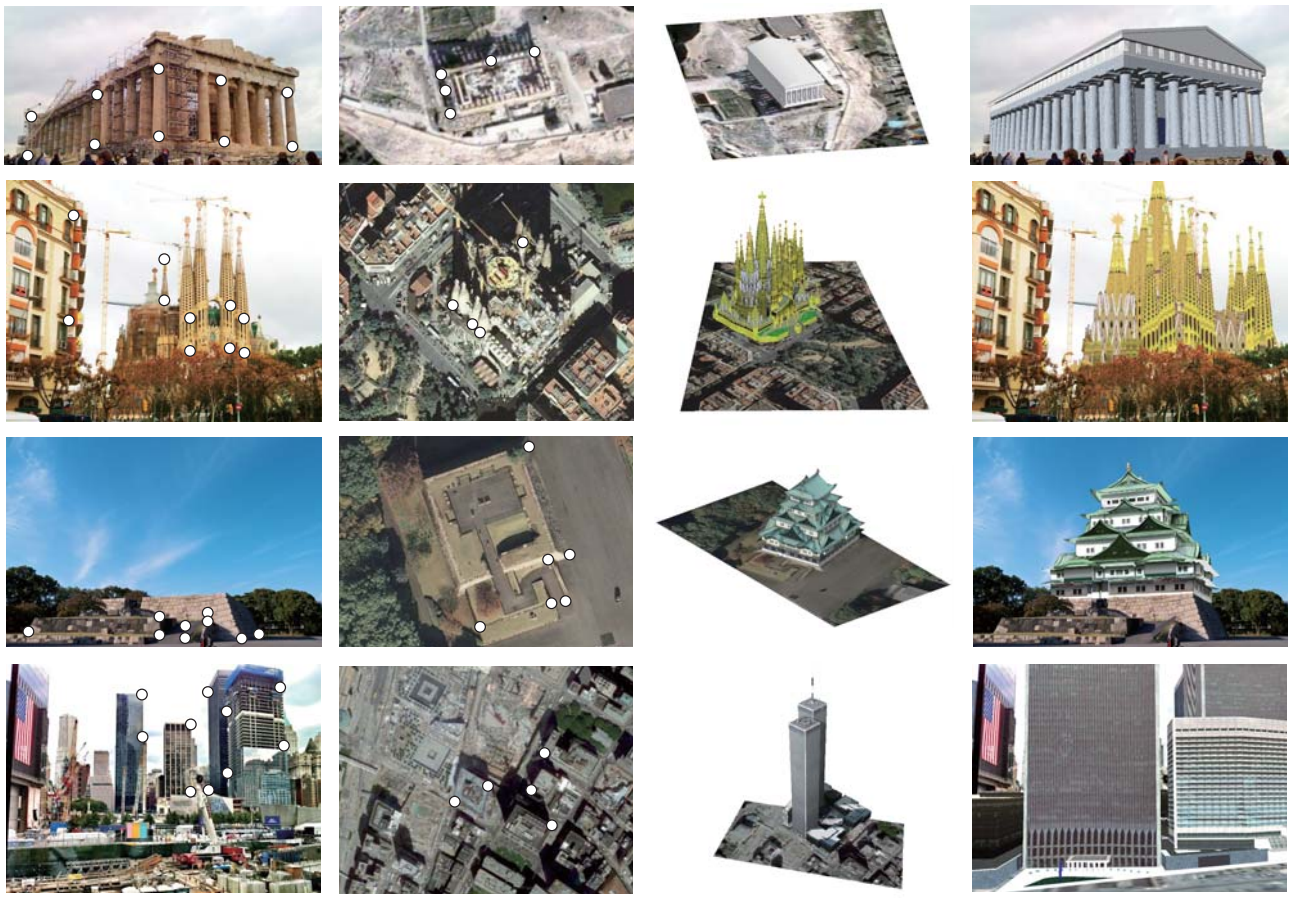
We have proposed a method for estimating a 6-DOF camera pose for a ground-view image using reference points on an aerial image. In order to estimate a 6-DOF camera pose in non-flat environments without altitude information, a cost function is newly defined as the sum of reprojection errors for lines and points. In the experiments, we have confirmed that the proposed method accurately estimates camera poses compared to conventional solvers that assume a planar surface and weak-perspective projection. The availability of the proposed method is demonstrated by generating augmented reality (AR) images from aerial and ground-view images downloaded from the Internet.

At this moment, our method is useful for the application of offline land scape simulation in which corresponding pairs of feature points can be manually given by users. However, in order to generate an AR image on a real site without manual operation, an automatic method for feature point matching is necessary. Although feature point matching between an aerial image and a ground view image is unfortunately still an open problem in this field, we will investigate the solution by combining the state-of-the-art techniques including automatic and global image-rectification techniques [22], [23] and locally distortion-resistant feature operators [24], [25] with roughly limited searching parameter space (e.g. by GPS, gyro and compass).

Acknowledgments This work was partially supported by JSPS KAKENHI Nos. 23240024 and 23700208.

References

[1] Fischler, M. A. and Bolles, R. C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, *Communications of ACM*, Vol. 24, No. 6, pp. 381–395 (1981).
[2] Liu, Y., Huang, T. S. and Faugeras, O. D.: Determination of camera location from 2-D to 3-D line and point correspondences, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 82–88 (1988).
[3] Abidi, M. A. and Chandra, T.: A new efficient and direct solution for pose estimation using quadrangular targets: Algorithm and evaluation, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 17, No. 5, pp. 534–538 (1995).
[4] Hartley, R. I. and Zisserman, A.: *Multiple View Geometry in Computer Vision*, Cambridge University Press, second edition (2004).
[5] Schweighofer, G. and Pinz, A.: Robust pose estimation from a planar target, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 28, No. 12, pp. 2024–2030 (2006).
[6] Lepetit, V., Moreno-Noguer, F. and Fua, P.: EP*n*P: An accurate $O(n)$ solution to the P*n*P Problem, *Int. J. of Computer Vision*, Vol. 81, No. 2, pp. 155–166 (2009).
[7] Hesch, J. A. and Roumeliotis, S. I.: A direct least-squares (DLS) method for PnP, *Proc. Int. Conf. on Computer Vision*, pp. 383–390 (2011).
[8] Torii, A., Sivic, J. and Pajdla, T.: Visual localization by linear combination of image descriptors, *Proc. Int. Conf. on Computer Vision Workshops*, pp. 102–109 (2011).
[9] Comport, A., Marchand, E., Pressigout, M. and Chaumette, F.: Real-time markerless tracking for augmented reality: the virtual visual servoing framework, *IEEE Trans. on Visualization and Computer Graphics*, Vol. 12, No. 4, pp. 615–628 (2006).
[10] Reitmayr, G. and Drummond, T. W.: Going out: robust model-based tracking for outdoor augmented reality, *Proc. Int. Symp. on Mixed and*

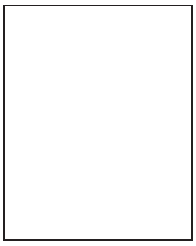


(a) (b) (c) (d)

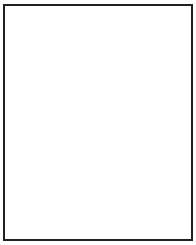
Fig. 13 AR landscape simulations: (a) Internet photographs from Flickr and Google, (b) aerial images from Google Maps, (c) CG objects drawn on aerial images, (d) generated AR images. White points are manually given reference points.

Augmented Reality., pp. 109–118 (2006).
[11] Arth, C., Wagner, D., Klopschitz, M., Irschara, A. and Schmalstieg, D.: Wide area localization on mobile phones, *Proc. Int. Symp. on Mixed and Augmented Reality*, pp. 73–82 (2009).
[12] Taketomi, T., Sato, T. and Yokoya, N.: Real-time and accurate extrinsic camera parameter estimation using feature landmark database for augmented reality, *Int. J. of Computers and Graphics*, Vol. 35, No. 4, pp. 768–777 (2011).
[13] Noda, M., Takahashi, T., Deguchi, D., Ide, I., Murase, H., Kojima, Y. and Naito, T.: Vehicle ego-localization by matching in-vehicle camera images to an aerial image, *Proc. ACCV2010 Workshop on Computer Vision in Vehicle Technology: From Earth to Mars*, pp. 1–10 (2010).
[14] Cham, T. J., Ciptadi, A., Tan, W. C., Pham, M. T. and Chia, L. T.: Estimating camera pose from a single urban ground-view omnidirectional image and a 2D building outline map, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 366–373 (2010).
[15] Bansal, M., Sawhney, H. S., Cheng, H. and Daniilidis, K.: Geo-localization of street views with aerial image databases, *Proc. ACM Multimedia 2011*, pp. 1125–1128 (2011).
[16] Dhome, M., Richetin, M., Lapreste, J. and Rives, G.: Determination of the attitude of 3-D objects from a single perspective view, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 12, pp. 1265–1278 (1989).
[17] Lee, S. C., Jung, S. K. and Nevatia, R.: Integrating ground and aerial views for urban site modeling, *Proc. Int. Conf. on Pattern Recognition*, Vol. 4, pp. 107–112 (2002).
[18] Ramalingam, S., Bouaziz, S. and Sturm, P.: Pose estimation using both points and lines for geo-localization, *Proc. IEEE Int. Conf. on Robotics and Automation*, pp. 4716–4723 (2011).
[19] Zhang, Z., Anandan, P. and Shum, H. Y.: What can be determined from a full and a weak perspective image?, *Proc. Int. Conf. on Computer Vision*, Vol. 1, pp. 680–687 (1999).
[20] Criminisi, A., Reid, I. and Zisserman, A.: Single view metrology, *Int. J. of Computer Vision*, Vol. 40, No. 2, pp. 123–148 (2000).
[21] Tamura, H. and Kato, H.: Proposal of international voluntary activities on establishing benchmark test schemes for AR/MR geometric registration and tracking methods, *Proc. Int. Symp. on Mixed and Augmented Reality*, pp. 233–236 (2009).
[22] Aiger, D., Cohen-Or, D. and Mitra, N. J.: Repetition maximization based texture rectification, *EUROGRAPHICS 2012*, Vol. 31, No. 2, pp. 439–448 (2012).
[23] Zhang, Z., Ganesh, A., Liang, X. and Ma, Y.: Tilt: Transform invariant low-rank textures, *Int. J. of Computer Vision*, Vol. 99, No. 1, pp. 1–24 (2012).
[24] Matas, J., Chum, O., Urban, M. and Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions, *Image and Vision Computing*, Vol. 22, No. 10, pp. 761–767 (2004).
[25] Morel, J. and Yu, G.: ASIFT: A new framework for fully affine invariant image comparison, *SIAM J. Imaging Sciences*, Vol. 2, No. 2, pp. 438–469 (2009).

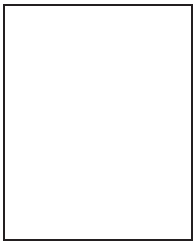
Taiki Sekii received his B.E. degree in computer science and systems engineering from Kobe University in 2009. He received his M.E. degree in information science from Nara Institute of Science and Technology in 2012. He has been working at Panasonic System Networks R&D Lab. Co.,Ltd. since 2012.



Tomokazu Sato received his B.E. degree in computer and system science from Osaka Prefecture University in 1999. He received his M.E. and Ph.D. degrees in information science from Nara Institute of Science and Technology in 2001 and 2003, respectively. He became an assistant professor at Nara Institute of Science and Technology in 2003. He was a visiting researcher in Czech Technical University in Prague in 2010-2011 and has been an associate professor at Nara Institute of Science and Technology since 2011.



Hideyuki Kume received his M.E. degree in information science from Nara Institute of Science and Technology in 2010. He is currently pursuing his Ph.D. degree in Nara Institute of Science and Technology. From 2011 to 2012, he was a Visiting Student at Carnegie Mellon University.



Naokazu Yokoya received his B.E., M.E., and Ph.D. degrees in information and computer science from Osaka University in 1974, 1976, and 1979, respectively. He joined the Electrotechnical Laboratory (ETL) in 1979. He was a visiting professor at McGill University in 1986-1987 and has been a professor at Nara Institute of Science and Technology since 1992.