LiDAR and Camera Calibration using Motion Estimated by Sensor Fusion Odometry

Ryoichi Ishikawa[1], Takeshi Oishi[1] and Katsushi Ikeuchi[2]

***Abstract*— This paper proposes a targetless and automatic camera-LiDAR calibration method. Our approach extends the hand-eye calibration framework to 2D-3D calibration. The scaled camera motions are accurately calculated using a sensor-fusion odometry method. We also clarify the suitable motions for our calibration method.**

Whereas other calibrations require the LiDAR reflectance data and an initial extrinsic parameter, the proposed method requires only the three-dimensional point cloud and the camera image. The effectiveness of the method is demonstrated in experiments using several sensor configurations in indoor and outdoor scenes. Our method achieved higher accuracy than comparable state-of-the-art methods.

I. Introduction

Sensor fusion of the camera and LiDAR has been widely studied in the robotics and computer vision fields. The multimodal systems enable the better performances in the robustness and accuracy in many vision problems than single-modal systems. One of the successful studies of sensor fusion is the three-dimensional scanning systems [1], [2], [3], [4]. 2D-3D sensor fusion provides reliable 3D point clouds with color information. The system also achieves accurate motion estimation while reconstructing 3D points.

The extrinsic calibration of multimodal sensors is one of the critical issues in 2D-3D sensor fusion systems. The error in the extrinsic parameters causes the distortions in between 2D and 3D images. The extrinsic parameters are usually estimated by using target cues or by manually associating 2D points on the image with 3D points on the point cloud.

Unfortunately, accurate calibration by manually establishing the correspondences is laborious because it requires multiple matches. Automated methods such as those developed in [5], [6], [7] use targets that are detectable on both 2D images and 3D point clouds. However, as prepared targets must be detectable by both the camera and LiDAR, these methods are impractical and undesirable for on-site calibration. Some recent studies have proposed targetless automatic 2D-3D calibration [8], [9], [10]. These methods use the mutual information for evaluating the distance between 2D and 3D images. As each sensor collects multimodal information, the calibration result depends on the modality among the sensors. The motion-based approaches use the relative motions that are separately estimated for single modalities as with the hand-eye calibration [11]. The motion-based 2D-3D calibration obtains the camera motion from the 2D images alone; the translation of the camera motion is without scaling. Accordingly, the precision of the extrinsic parameter largely depends on the motion error in the hand-eye calibration. Although the scale and extrinsic parameter can be simultaneously calculated from multiple movements, the scaleless camera motion deteriorates the accuracy of the hand-eye calibration.

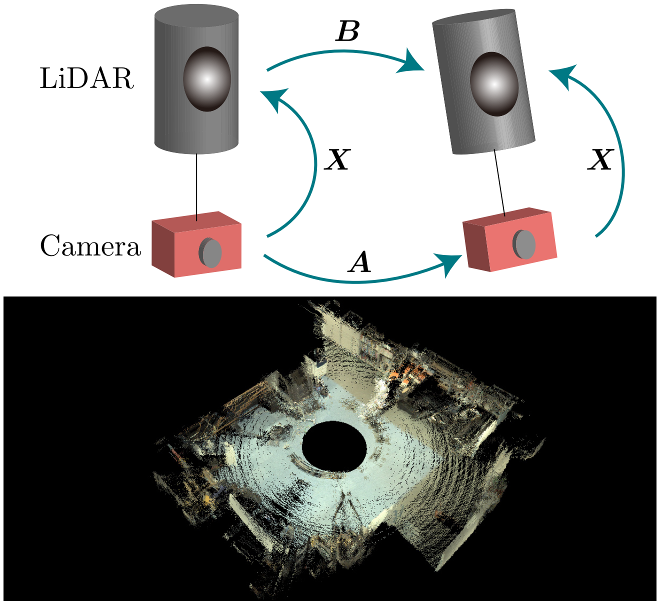

Fig. 1. Top: Proposed motion-based 2D-3D calibration. The LiDAR motion is estimated by ICP algorithm. The camera motion is initially estimated by feature-point matching, then estimated to scale by a sensor-fusion system. Bottom: Colored scan captured by HDL-64E. The texture is taken by Ladybug 3 and corrected by our calibration result.

The present paper proposes an automatic targetless calibration method between a fixed camera and LiDAR. The proposed method employs a motion-based approach (see Fig. 1). The method estimates the relative motions of the sensors separately for several movements and estimates the calibration parameters. The extrinsic parameter is numerically solved from the sensor motions which are calculated in the same modality. To solve the scale issue and achieve accurate camera translation, our method estimates the camera motion using sensor-fusion odometry [1], [2], [3], [4]. Though the sensor fusion odometry gives accurate camera motions with scale, it requires known calibration parameters.

[1] Ryoichi Ishikawa and Takeshi Oishi are with Institute of Industrial Science, The University of Tokyo, Japan {ishikawa, oishi}@cvl.iis.u-tokyo.ac.jp

[2] Katsushi Ikeuchi is with Microsoft, USA, katsuike@microsoft.com

Therefore, it first obtain an initial extrinsic parameter from the LiDAR motions and the ”scaleless” camera motions. Next, the camera motions are recalculated using the scale determined by the initial extrinsic parameter and the point cloud from the LiDAR. The extrinsic parameter then recalculated using the updated motions. The camera motions and the extrinsic parameter are repeatedly calculated until the estimation converges.

The contributions of this paper are summarized below.

- To our knowledge, the estimation of camera motion estimation in a sensor fusion system has not been previously applied in 2D-3D calibration.
- We determine the sensor motion that optimizes the effectiveness of the calibration method.
- The only inputs are the RGB image from a camera and the three-dimensional point cloud from a LiDAR. No additional data such as the LiDAR reflectance or initial value of the extrinsic parameter are required. The method estimates the extrinsic parameter more accurately than other methods from a small number of motions.

Our proposed method requires that the measurement ranges of the camera and the LiDAR overlap. To align the scans for the LiDAR motion estimation, the LiDAR’s measurement range must also be two-dimensional for securing overlap.

II. Related Work

Our work is related to targetless automatic 2D-3D calibration and hand-eye calibration.

A. Targetless multimodal calibration

Targetless automatic 2D-3D calibration methods generally use the mutual information in the image and the point cloud. For example, portions that appear as discontinuous 3D shapes will probably appear as edges on an RGB image. Therefore some calibration methods align this three-dimensional discontinuous portion with the 2D edges [9], [10]. Viola *et al.*[12] proposed a multimodal alignment method based on mutual information (MI), which has been developed mainly for medical imaging purpose. More recently, MI-based 2D-3D calibration methods that evaluate the commonality between LiDAR and camera data have been proposed. Among the indicators for MI evaluations are the reflectance - (gray-scale intensity) [8], surface normal - (gray-scale intensity) [13], and multiple evaluation indicators, including the discontinuities in LiDAR data and the edge strengths in images [14]. A gradient-based metric was proposed by Taylor and Nieto [15].

The above methods align 3D point clouds to 2D images by projecting the points in the 3D space onto a 2D surface. Alternatively, texturing methods construct a stereo image from camera images taken at multiple sites. The 3d structure is then reconstructed from the images and aligned with a 3D point cloud. The authors of [16] developed a method that computes the extrinsic parameter between the LiDAR and the camera for texturing a dense 3D scan. The method uses the anisotropic error distribution but the accuracy of the alignment is deteriorated because of the sparse stereo reconstruction.

B. Hand-eye Calibration

In hand-eye calibration, the position and orientation of the sensor are changed and the extrinsic parameter is calibrated using the motions observed by each sensor. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be the position and orientation changes observed by two fixed sensors, and let $\boldsymbol{X}$ be the unknown relative position and orientation between the sensors. Then $\boldsymbol{A}$ and $\boldsymbol{B}$ are related as $\boldsymbol{AX} = \boldsymbol{XB}$ (upper panel of Fig. 1). Solving this expression for $\boldsymbol{X}$ provides the extrinsic parameter between the two sensors [17], [18]. Furthermore, as the sensor is influenced by noise, the calibration is accompanied by estimation of the sensor bias using a Kalman filter [19], [20].

Heng *et al.*[21] proposed a method that calibrates the transitions between four cameras mounted on a car using visual odometry. Taylor and Nieto [11] obtained the motion of a sensor in a 2D-3D calibration and estimated the extrinsic parameter combining the motion and the multimodal alignment. Although their method achieves highly accurate calibration, it estimates a scaleless translation from the camera images, which reduces the accuracy of the camera motions. This deteriorate the accuracy of the extrinsic translation parameter particularly for a small number of motions.

III. Methodology

Our method is overviewed in Fig. 2. The proposed method is divided into two main steps. The initialization phase estimates the extrinsic parameter from the LiDAR motions $\{\boldsymbol{R}_{lid}, \boldsymbol{t}_{lid}\}$ using iterative-closest-point (ICP) alignment and from the camera motions $\{\boldsymbol{R}_{cam}, \bar{\boldsymbol{t}}_{cam}\}$ using feature-point matching. The iterative stage alternates between updating the extrinsic parameter $\hat{\boldsymbol{R}}, \hat{\boldsymbol{t}}$ and the scaled camera motions $\{\hat{\boldsymbol{R}}_{cam}, \hat{\boldsymbol{t}}_{cam}\}$ through sensor-fusion odometry until the convergence condition is reached.

A. Initial calibration parameter estimation

1) LiDAR motion estimation: The LiDAR motion is estimated by a high-speed registration method in the ICP algorithm, which searches for corresponding points in the gaze direction [22]. Meshes on the point clouds are created in advance by projecting a sequence of points onto a two-dimensional plane and applying Voronoi splitting. When aligning the scans, the threshold distance between corresponding points is initially set to a large value, and the outlier correspondences are eliminated while gradually decreasing the threshold. Assuming that two point clouds are scanned at sufficiently close positions for ICP alignment, the initial position of the ICP alignment is the center of the two LiDAR coordinates.

2) Camera-motion estimation: The initial motion of the camera is estimated by standard feature-point matching. We first extract the feature points from two images using the AKAZE algorithm [23], calculate the descriptors, and make the matchings. From the matchings, we calculate the

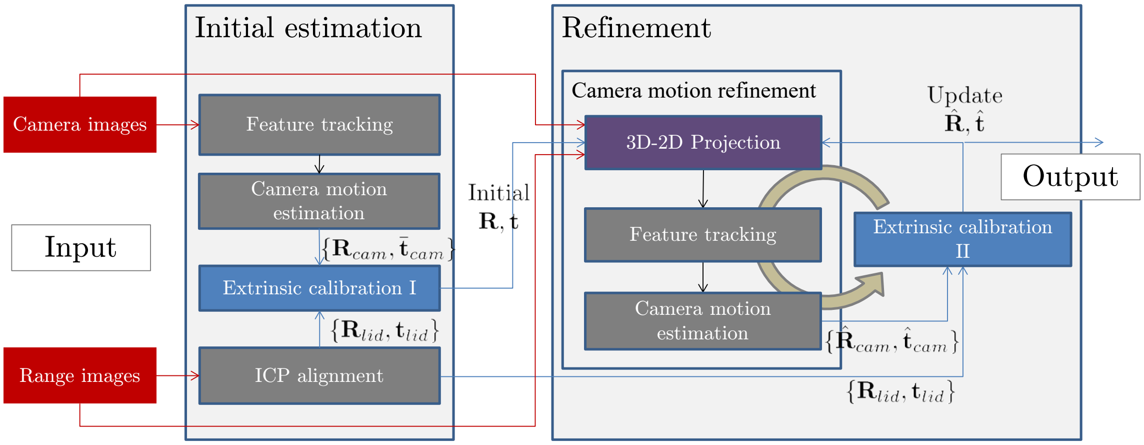
Initial estimation
Refinement
Camera images
Feature tracking
Camera motion estimation
$\{\mathbf{R}_{cam}, \bar{\mathbf{t}}_{cam}\}$
Extrinsic calibration I
$\{\mathbf{R}_{lid}, \mathbf{t}_{lid}\}$
Range images
ICP alignment
Input
Initial $\mathbf{R}, \mathbf{t}$
Camera motion refinement
3D-2D Projection
Feature tracking
Camera motion estimation
Update $\hat{\mathbf{R}}, \hat{\mathbf{t}}$
Output
Extrinsic calibration II
$\{\hat{\mathbf{R}}_{cam}, \hat{\mathbf{t}}_{cam}\}$
$\{\mathbf{R}_{lid}, \mathbf{t}_{lid}\}$

Fig. 2. Overview

initial relative position and orientation between the camera frames using a 5-point algorithm [24] and the random sample consensus algorithm [25]. Let $\boldsymbol{u}_j$ and $\boldsymbol{v}^c_j$ be the j th matched pair of feature points in the first and second image of the i th motion (expressed as a unit vector from the camera center). $\boldsymbol{v}^c_j$ is expressed in the coordinate frame of the second camera. After obtaining the initial relative position and orientation of the camera motion $\boldsymbol{R}^i_{cam}, \bar{\boldsymbol{t}}^i_{cam}$ by linear computation, the camera motions are optimized by minimizing the projection error using an angle error metric with an epipolar plane [26]:

$$\boldsymbol{R}^i_{cam}, \bar{\boldsymbol{t}}^i_{cam} = \underset{\boldsymbol{R}_{cam}, \bar{\boldsymbol{t}}_{cam}}{\arg\min} \sum_j \frac{\boldsymbol{u}_j^\top [\bar{\boldsymbol{t}}_{cam}]_\times \boldsymbol{R}_{cam} \boldsymbol{v}^c_j}{|[\bar{\boldsymbol{t}}_{cam}]_\times \boldsymbol{R}_{cam} \boldsymbol{v}^c_j|} \qquad (1)$$

where $[\cdot]_\times$ is a skew-symmetric matrix expressing the crossproduct operation.

3) Extrinsic calibration I: To obtain the relative position and orientation between the two sensors from the initially estimated motions, we extend the normal hand-eye calibration to include the scale estimation of the camera motion. The series of camera motion $\boldsymbol{R}^i_{cam}, \bar{\boldsymbol{t}}^i_{cam}$ and LiDAR motion $\boldsymbol{R}^i_{lid}, \boldsymbol{t}^i_{lid}$ satisfy the following relationships [17]:

$$\boldsymbol{R}^i_{cam}\boldsymbol{R} = \boldsymbol{R}\boldsymbol{R}^i_{lid} \qquad (2)$$

$$\boldsymbol{R}^i_{cam}\boldsymbol{t} + s^i \bar{\boldsymbol{t}}^i_{cam} = \boldsymbol{R}\boldsymbol{t}^i_{lid} + \boldsymbol{t}, \qquad (3)$$

where s^i is the scale factor. The extrinsic calibration parameter $\boldsymbol{R}$ is initially estimated from the following equation derived from Eq. 2:

$$\boldsymbol{k}^i_{cam} = \boldsymbol{R}\boldsymbol{k}^i_{lid} \qquad (4)$$

where $\boldsymbol{k}^i_{cam}$ and $\boldsymbol{k}^i_{lid}$ are rotational axes of rotation matrices $\boldsymbol{R}^i_{cam}$ and $\boldsymbol{R}^i_{lid}$, respectively, and $\boldsymbol{t}$ is solved through Eq. 3. $\boldsymbol{R}$ is linearly solved from the series of $\boldsymbol{k}^i_{cam}, \boldsymbol{k}^i_{lid}$ by singular-value decomposition. However, the rotation solution requires at least two position and orientation transitions, and the series of transitions must include rotations in different directions. $\mathbf{R}$ is then nonlinearly optimized by minimizing the following cost function:

$$\boldsymbol{R} = \underset{\boldsymbol{R}}{\arg\min} \sum_i \left|\mathbf{R}^i_{cam}\mathbf{R} - \mathbf{R}\mathbf{R}^i_{lid}\right|. \qquad (5)$$

After optimizing $\boldsymbol{R}$, $\boldsymbol{t}$ and s^i are obtained by constructing simultaneous equations and solving them by the linear least-squares method.

B. Iteration of camera-motion estimation and sensor calibration

The sensor-fusion system enables highly accurate estimates of the scaled camera motion [1], [3] even when the base line is short. In contrast, estimating the relative translation between two camera images is inaccurate without scaling. Having estimated the extrinsic parameter, we can compute the scaled camera motion $\hat{\boldsymbol{R}}_{cam}, \hat{\boldsymbol{t}}_{cam}$, and update $\hat{\boldsymbol{R}}, \hat{\boldsymbol{t}}$ by the extrinsic calibration II. The extrinsic parameter is then optimized by repeatedly alternating the camera motion and extrinsic parameter estimations until convergence.

1) Camera-motion estimation with range data: This step first constructs the 2D-3D correspondence for computing $\hat{\boldsymbol{R}}_{cam}, \hat{\boldsymbol{t}}_{cam}$. Figure 3 shows the construction of the 2D-3D correspondence. The inputs are a point cloud in LiDAR coordinates, two camera images taken from the positions of cameras 1, and 2, and the extrinsic parameter. First, the point cloud is projected onto the image from camera 1. A certain point in the point cloud with local LiDAR coordinates $\boldsymbol{p}^l$ is transformed into camera coordinates and projected onto a two-dimensional plane by a projection function $Proj(\boldsymbol{p}) = \frac{\boldsymbol{p}}{|\boldsymbol{p}|}$. The pixel onto which $\boldsymbol{p}^l$ is projected is expressed as

$$\boldsymbol{u} = Proj(\boldsymbol{R}\boldsymbol{p}^l + \boldsymbol{t}), \qquad (6)$$

where $\boldsymbol{u}$ is the vector heading from the center of camera 1 to the corresponding pixel (See Fig. 3 (a)). The pixel onto which $\boldsymbol{p}$ is projected is then tracked from camera image 1 to image 2 using a KLT tracker [27]. Let $\boldsymbol{v}^c$ be the vector heading from the center of camera 2 to the tracked pixel, where c indicates that the vector is expressed in the local

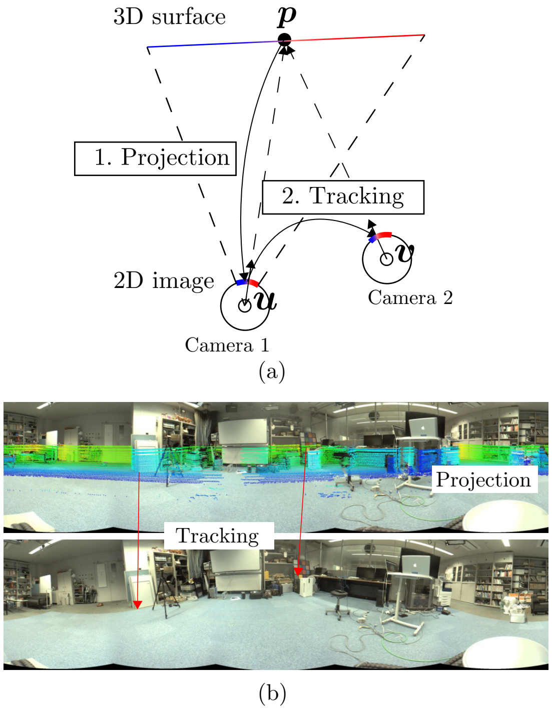

Fig. 3. (a) Construction of the 2D-3D correspondences. The range data are projected onto first image. The pixels with depth are tracked to second image. (b) Example of the construction process of the 2D-3D correspondences.

coordinates of camera 2. The point $\boldsymbol{p}$ and vector $\boldsymbol{v}^c$ of the 2D-3D correspondences are then constructed.

After constructing the 2D-3D correspondences, the relative position and orientation of camera 1 with respect to camera 2 can be optimized by minimizing the projection error. Let $(\boldsymbol{v}^c_j, \boldsymbol{p}^l_j)$ be the j th 2D-3D correspondence in the i th motion. The position and orientation transition $\hat{\boldsymbol{R}}^i_{cam}, \hat{\boldsymbol{t}}^i_{cam}$ between the cameras is optimized by minimizing the following angle-metric cost function

$$\hat{\boldsymbol{R}}^i_{cam}, \hat{\boldsymbol{t}}^i_{cam} = \underset{\boldsymbol{R}_{cam},\boldsymbol{t}_{cam}}{\arg\min} \sum_j \left| \boldsymbol{v}^c_j \times Proj(\boldsymbol{R}^\top_{cam}((\boldsymbol{R}\boldsymbol{p}^l_j + \boldsymbol{t}) - \boldsymbol{t}_{cam}))\right|. \tag{7}$$

The first iteration implements a generalized perspective 3-point algorithm on the initial values of $\hat{\boldsymbol{R}}_{cam}$ and $\hat{\boldsymbol{t}}_{cam}$ [28]. The estimation result of the previous iteration is used in subsequent iterations.

2) Extrinsic calibration II: After updating the camera motions $\hat{\boldsymbol{R}}_{cam}$, $\hat{\boldsymbol{t}}_{cam}$, the method updates the extrinsic parameter $\hat{\boldsymbol{R}}$, $\hat{\boldsymbol{t}}$. In each iteration, $\hat{\boldsymbol{R}}_{cam}$ and $\hat{\boldsymbol{t}}_{cam}$ are solved both linearly and nonlinearly. The camera and LiDAR motions are related as follows [17]:

$$\hat{\boldsymbol{R}}^i_{cam}\hat{\boldsymbol{t}} + \hat{\boldsymbol{t}}^i_{cam} = \hat{\boldsymbol{R}}\boldsymbol{t}^i_{lid} + \hat{\boldsymbol{t}}, \tag{8}$$

In the nonlinear optimization step, $\hat{\boldsymbol{R}}$ is optimized by Eq. 5 and $\boldsymbol{t}$ is optimized by the following cost function based on Eq. 8:

$$\hat{\boldsymbol{t}} = \underset{t}{\arg\min} \sum_i \left|(\boldsymbol{R}^i_{cam}\boldsymbol{t} + \boldsymbol{t}^i_{cam}) - \hat{\boldsymbol{R}}\boldsymbol{t}^i_{lid} + \boldsymbol{t})\right|. \tag{9}$$

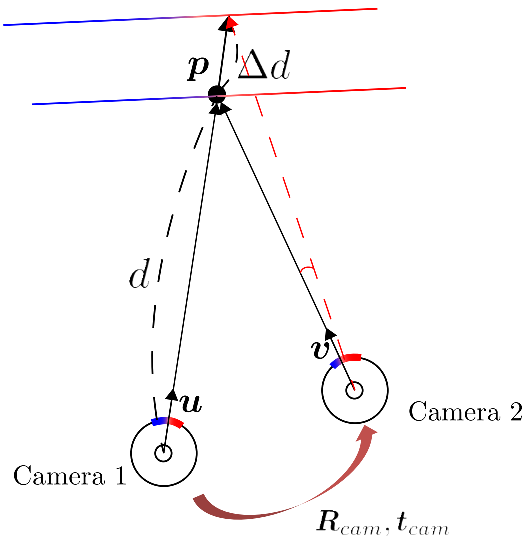

Fig. 4. Case of an error in the localized range data in camera 1 coordinates

IV. Optimal motion for 2D-3D calibration

When determining the most suitable sensor motion for the calibration, we must consider the influence of the extrinsic error on the camera-motion estimation. During the repeated alternating estimation of the extrinsic parameter between the sensors and the camera motion, extrinsic parameter error in the camera motion is inevitable. As the motion and extrinsic parameter estimates also depend on the measured environment and the number of motions, precise convergence conditions are difficult to obtain. However, the motions will likely converge the estimation are determinable.

A. Camera motion estimation

Let us consider the conditions of successful camera-motion estimation when the extrinsic parameter is erroneous. We associate the depth information to the camera images. $\boldsymbol{v}^c$ and $\boldsymbol{p}^l$ is transformed to $\boldsymbol{v}$ and $\boldsymbol{p}$ in global coordinates, respectively. $\boldsymbol{p}$, $\boldsymbol{u}$, and $\boldsymbol{v}$ are aligned in the same coordinate system. Define d as the distance from camera 1 to $\boldsymbol{p}$. During the projection step, the LiDAR point cloud is transformed by the extrinsic parameter, and becomes localized in the coordinate system of camera 1. However, as the extrinsic parameter contains error, the localization is incorrect and the depth associated with $\boldsymbol{u}$ includes an error Δd.

This depth error shifts the estimated position of camera 2 from the ground truth when minimizing the projection error. At the correct position of camera 2, the projection error is given by

$$e = \left|\boldsymbol{v} \times \frac{(\Delta d + d)\boldsymbol{u} - \boldsymbol{t}_{cam}}{|(\Delta d + d)\boldsymbol{u} - \boldsymbol{t}_{cam}|}\right|. \tag{10}$$

From Fig. 4, $\boldsymbol{v}$ is given by

$$\boldsymbol{v} = \frac{d\boldsymbol{u} - \boldsymbol{t}_{cam}}{|d\boldsymbol{u} - \boldsymbol{t}_{cam}|}. \tag{11}$$

Substituting Eq. 11 into Eq. 10, we obtain

$$e \quad = \quad \left|\frac{\Delta d(\boldsymbol{t}_{cam} \times \boldsymbol{u})}{|d\boldsymbol{u} - \boldsymbol{t}_{cam}||(\Delta d + d)\boldsymbol{u} - \boldsymbol{t}_{cam}|}\right|. \tag{12}$$

Now, to estimate the extrinsic parameter accurately, we should estimate the ideal camera motion during the actual movement of the camera. In other words, the estimated orientation and position of camera 2 should approach $\boldsymbol{R}_{cam}$ and $\boldsymbol{t}_{cam}$, respectively. This requires The projection error becomes vanishingly small ($e \to 0$) when estimated camera 2 is localized on ground truth. From Eq. 12, followings is said:

- According to Eq. 12, reducing $\boldsymbol{t}_{cam}$ reduces e even when the extrinsic parameter contains error. That is, the smaller the moving distance of the camera, the more accurate is the estimation becomes. In other words, when $\boldsymbol{t}_{cam}$ is small, the estimated position of camera 2 will more likely lie at the periphery of the ideal site. In subsequent extrinsic parameter estimations, the existence probability of the calibrated extrinsic parameter will also sharply distribute around the true value.
- Decreasing Δd also reduces the projection error. Δd can deviate under certain points. For example, when there are depth discontinuity parts nearby or when the incident angle from the camera is shallow. To avoid these problems, the calibration environment should be surrounded by smooth terrain as possible.

B. Extrinsic parameter estimation

We next consider how the error in the estimated motions influences the extrinsic parameter calibration. Ignoring the rotation error for simplicity, we assume errors in the translations of the camera motion and the extrinsic parameter in Eq. 8. Let $\boldsymbol{e}_{cam}$ and $\boldsymbol{e}$ be the error in $\boldsymbol{t}_{cam}$ and $\boldsymbol{t}$, respectively. Then we have

$$\boldsymbol{R}_{cam}(\boldsymbol{t} + \boldsymbol{e}) + \boldsymbol{t}_{cam} + \boldsymbol{e}_{cam} = \boldsymbol{R}\boldsymbol{t}_{cam} + \boldsymbol{t} + \boldsymbol{e}. \tag{13}$$

The difference between Eqs. 13 and 8 is

$$\boldsymbol{e}_{cam} = (\boldsymbol{I} - \boldsymbol{R}_{cam})\boldsymbol{e}. \tag{14}$$

We assume that Eq. 14 describes a single unit. When the rotation $\boldsymbol{R}_{cam}$ is small, the translation errors satisfy $|\boldsymbol{e}_{cam}| < |\boldsymbol{e}|$. This indicates that the error of the camera motion is amplified and propagates to the extrinsic parameter.

If the error propagated when estimating the extrinsic parameter from the camera motions does not exceed the error reduction when estimating the camera motion from the distance image, the proposed method will improve the accuracy of the relative position and orientation. Therefore, to reduce the propagated amount of error in the relative position and orientation estimation, one can rotate the camera motion through a larger angle. Sampling a plurality of motions (as many as possible) will also ensure a robust extrinsic parameter estimation. When rotation the camera motion, one must realize that if the image appearance changes significantly, the accuracy of the motion estimation can degrade. One must therefore account for this possibility. Although the proposed method is applicable to perspective cameras, omnidirectional cameras are advantageous because they can secure a common field of view even when rotated significantly.

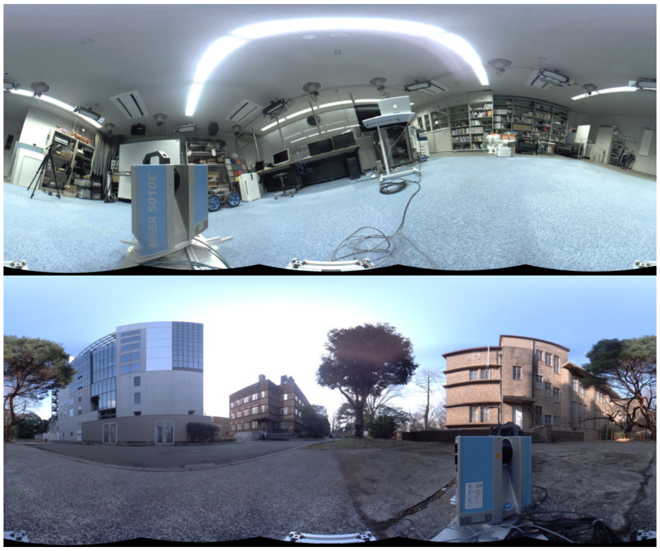

Fig. 5. Indoor and outdoor calibration scene taken by Ladybug 3

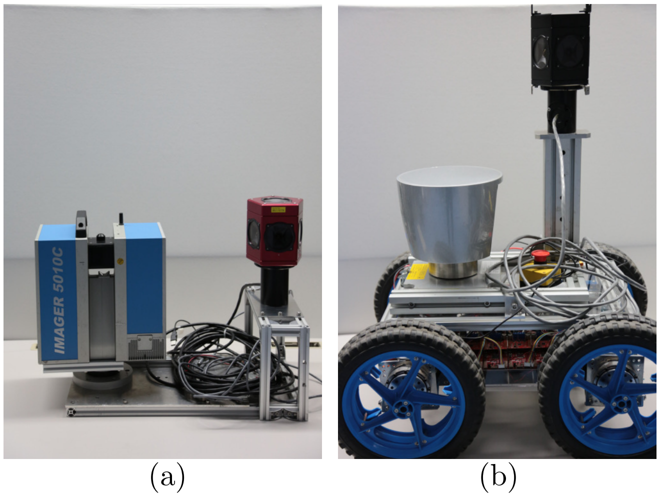

Fig. 6. Sensor configurations. (a)Imager 5010C and Ladybug 3, (b)HDL-64E mounted with Ladybug 3

V. Experimental results

The experimental calibrations were conducted in indoor and outdoor environments (see Fig. 5). Images and point clouds were collected by a panoramic LiDAR, a multibeam LiDAR, and an omnidirectional camera. As the comparison method, we selected image-based calibration using scaleless camera motions, as applied in [11]. This calibration, hereafter referred to as "*Scaleless*", is the initialization output in Fig. 2.

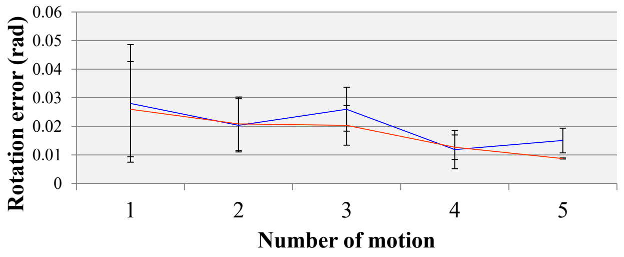

Fig. 7. Rotation error versus the number of motions. Blue line: *Scaleless*, Red line: Our method.

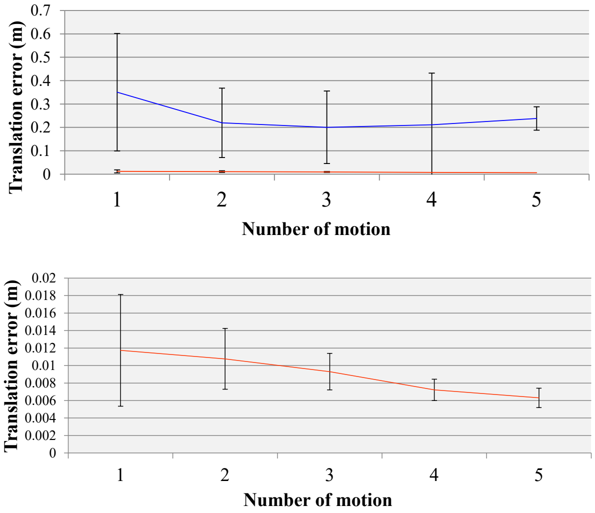

Fig. 8. Translation error versus number of motions. Blue line: *Scaleless*, Red line: Our method. Bottom panel shows the result of our method only.

As the other comparison methods, we chose calibration by *Manual* correspondence acquisition and calibration with alignment using *MI* [8].

A. Evaluation with colored range data

We first show the results of the evaluation datasets. The data were measured by two range sensors, Focus S 150 by FARO Inc.[1] and Imager 5010C by Zollar+Flöhlich Inc.[2]. High-accuracy panoramic laser range scanners can be used for mobile scanning in profiler mode [3], [4]. Three-dimensional panoramic point clouds were scanned with both range sensors. The data measured by Focus S 150 were converted to a colored point cloud using a photo texture function. The evaluation inputs were a pseudo-panorama-rendered image obtained from the colored point cloud scanned by Focus S 150, and a point cloud scanned by Imager 5010C. The ground truth was computed by registering the two point clouds scanned by the two sensors.

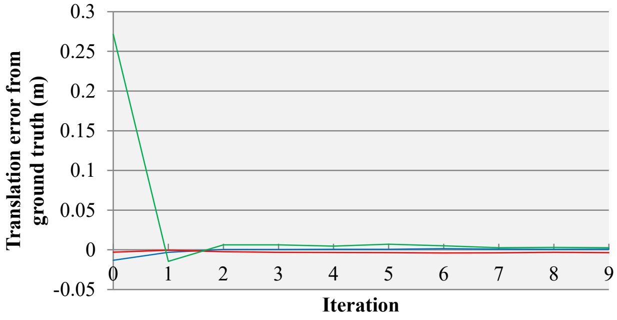

Fig. 9. Example of translation parameter x (blue line), y (red line), and z (green line) error versus iteration times using ten motions.

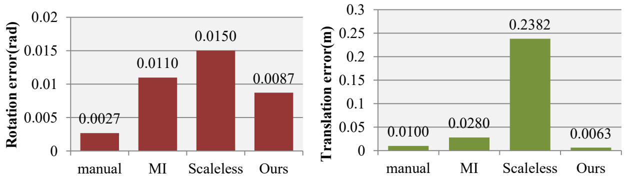

Fig. 10. Error in the calibration results of each method (relative to the ground truth) in the indoor scene

Motion in the indoor scene were captured by rotating the sensors five times each in the vertical and horizontal directions. The rotational and translational calibration with changing the number of motion in the indoor scene is shown in Figs. 7 and 8, respectively. The horizontal axes represent half the numbers of horizontal and vertical motions in the calibration. For example, one motion denotes that the calibration was performed with two motions (one horizontal and one vertical motion). Each calibration was performed 10 times for each number of motion(1-5), with random sampling of the motions. The result of Figs. 7 and 8 are the averages and standard deviations of the rotational and translational errors in the extrinsic parameter, respectively, in both *Scaleless* (blue lines) and the proposed method (red lines). The proposed method did not significantly improve the rotational error, but dramatically improved translational error. Moreover, the translational error in the proposed method gradually decreased with increasing number of motions. Figure 9 shows the example of translation parameter convergence.

Figure. 10 compares the calibration results of the proposed, *Scaleless*, *Manual*, and *MI* methods. Registration by maximizing MI completes only one scan. This method shifts the initial point from the ground truth by a fixed distance (0.1 m) in a random direction (the rotational parameter is not shifted). The *Manual* calibration acquires 30 correspondences in as many directions as possible. As shown in Fig. 10, the rotational error was below 1 ° in all method. However, our method achieved considerably lower translational error than the other methods.

Figure 11 plots the evaluation results in the outdoor environment. The motions were captured by rotating the sensors by three times in the vertical direction and three times

[1] https://www.faro.com

[2] http://www.zf-laser.com

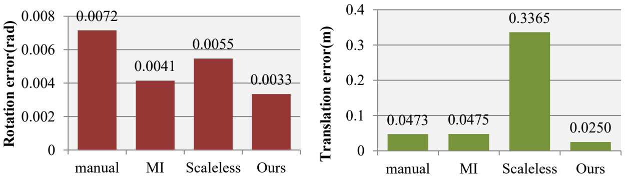

Fig. 11. Error in the calibration results of each method (relative to the ground truth) in the outdoor scene

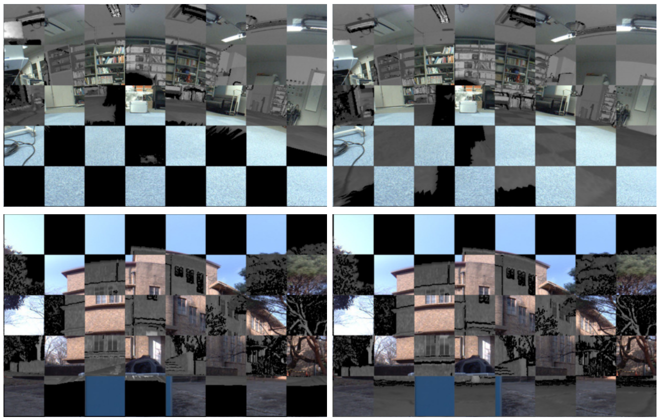
Scaleless
Ours

Fig. 12. Checkerboard arrangement of the panorame images taken by Ladybug 3 and the panorama-rendered reflectance image scan scanned by Imager 5010C. When the extrinsic parameter is correct, the two images are consistent.

in the horizontal direction. Although our method minimized the rotation error, the errors in all methods were less then $0.5\,^{\circ}$, with no significant differences among the methods. On the other hand, our method most accurately estimated the translation. For the same number of motions, the translational accuracy was below $1\,\mathrm{cm}$ in the indoor environment, versus $2.5\,\mathrm{cm}$ in the outdoor environment. Therefore the indoor scene is preferred for calibration purposes.

B. Ladybug 3 and Imager 5010C

We now show the calibration results of the omnidirectional camera Ladybug 3 by FLIR Inc.[3] and Imager 5010C. The sensor configuration is shown in Fig. 6 (a). The evaluation was performed by overlaying the image of Ladybug 3 and the reflectance image obtained by the panorama rendering from the center of the estimated camera position in the point cloud (See Fig. 12). The Images from the two sensors were alternated in a checkerboard pattern. We thus confirmed the consistency of the two images. As shown in Fig. 12, the pre-optimized *Scaleless* lacked consistency between RGB and reflectance images, but the proposed method achieved consistency between the two images.

C. Ladybug 3 and Velodyne HDL-64E

Finally, we show the extrinsic calibration results of Ladybug 3 and the multibeam LiDAR HDL-64E by Velodyne Inc. [4] in the indoor scene. The sensor configuration is shown in Fig. 6 (b). While measuring the data, the rover loaded with the sensors was operated to generate rotational motion in the vertical direction and the horizontal directions. The vertical motions were generated by raising and lowering the front wheel through steps of approximately $4\,\mathrm{cm}$. The data were measured in stop-and-scan mode; that is, the rover was stopped while the data were scanned. To obtain the range data and scanned camera images at the same position, we visually checked the times at which the LiDAR and the camera were stationary.

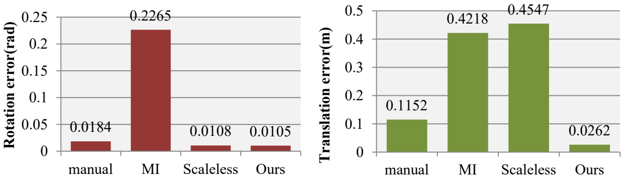

Fig. 13. Error in the calibration results of each method (relative to the ground truth) using the HDL-64E and Ladybug 3 sensors. In *MI*, *Scaleless*, and proposed method, the calibration was performed with 16 scans.

The ground truth of the extrinsic parameter between HDL-64E and Ladybug 3 was indirectly obtained by computing the relative position and orientations of the point cloud measured by Imager 5010C in the same environment. The positions and orientations of HDL-64E and Imager 5010C were obtained by aligning the range data scanned by each sensor. The extrinsic parameter between Ladybug 3 and Imager 5010C was obtained by manually specifying the correspondence points between the panorama image and the three-dimensional reflectance image. The extrinsic parameter was then computed using the correspondences. When calibrating *Scaleless* and the proposed method, eight horizontal rotation motions and eight vertical rotation motions were randomly sampled at each time, and the average error of 10 calibrations was recorded. The *Manual* calibration was carried out using a single scan, taking the corresponding points between the panoramic image of Ladybug 3 and the 3D reflectance image of HDL-64E. In the *MI* calibration, the MI was calculated from 16 scan sets of an image-point cloud scanned at each position.

As shown in Fig. 13, the *Manual* calibration was inaccurate because the correspondence construction failed under the narrow scan range and the sparse point cloud of HDL-64E. Meanwhile, the *MI* optimization failed on the HDL-64E dataset because of the low resolution and unclear reflectance information. On the other hand, the accuracies of the motion-based methods (*Scaleless* and the proposed method) were below $1\,^{\circ}$ in the rotational direction. However, the translation in *Scaleless* significantly differed from the ground truth, whereas the proposed method achieved highly accurate translation results.

The proposed method can also process the motions acquired by an operating rover. To obtain the extrinsic param-

[3] https://www.ptgrey.com/

[4] http://velodynelidar.com/

eters of a six degree-of-freedom by hand-eye calibration, the sensor must be rotated in two or more directions. However, in vertical rotational motion, part of the platform on which the sensors are mounted must be raised. Although this operation is more difficult than horizontal rotation, this experiment demonstrates that the proposed method works well when the vertical rotational motions are implemented by reasonable mobile-platform operations, such as raising and lowering by a small step.

VI. Conclusion

We presented a targetless automatic 2D-3D calibration for camera-motion estimation. The calibration is based on hand-eye calibration and uses sensor fusion odometry. The proposed method can be fully utilized with fewer translations and larger rotations of the camera than other approaches. The calibration measurements are best acquired in environments surrounded by flat terrains.

Hand-eye calibration requires multi-directional rotations of the sensor. However, in many situations, vertical rotations are more difficult than horizontal rotations. Although the proposed method must also satisfy this condition, sufficient vertical rotation was achieved by reasonable movements of the mobile platform. Therefore, this method is highly practical. Although the proposed method is for off-line calibration, it can be extended to on-line calibration during scanning, provided that the motions are appropriately selected and computational time is reduced.

Acknowledgment

This work was partially supported by the social corporate program (Base Technologies for Future Robots) sponsored by NIDEC corporation. This work was also supported by JSPS KAKENHI Grant Number JP16747698, JP17923471, and JSPS Research Fellow Grant No.16J09277.

References

[1] Y. Bok, Y. Jeong, D. Choi, and I. Kweon, "Capturing Village-level Heritages with a Hand-held Camera-Laser Fusion Sensor," *International Journal of Computer Vision*, vol. 94, pp. 36–53, 2011.

[2] J. Zhang, M. Kaess, and S. Singh, "A real-time method for depth enhanced visual odometry," *Autonomous Robots*, vol. 41, no. 1, pp. 31–43, 2017.

[3] B. Zheng, X. Huang, R. Ishikawa, T. Oishi, and K. Ikeuchi, " A New Flying Range Sensor: Aerial Scan in Omini-directions ," in *International Conference on 3D Vision*, 2015.

[4] R. Ishikawa, M. Roxas, Y. Sato, T. Oishi, T. Masuda, and K. Ikeuchi, "A 3d reconstruction with high density and accuracy using laser profiler and camera fusion system on a rover," in *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 2016, pp. 620–628.

[5] Q. Zhang and R. Pless, "Extrinsic calibration of a camera and laser range finder (improves camera calibration)," in *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, vol. 3. IEEE, 2004, pp. 2301–2306.

[6] V. Fremont, P. Bonnifait, *et al.*, "Extrinsic calibration between a multi-layer lidar and a camera," in *Multisensor Fusion and Integration for Intelligent Systems, 2008. MFI 2008. IEEE International Conference on*. IEEE, 2008, pp. 214–219.

[7] Y. Bok, D.-G. Choi, and I. S. Kweon, "Extrinsic calibration of a camera and a 2d laser without overlap," *Robotics and Autonomous Systems*, vol. 78, pp. 17–28, 2016.

[8] G. Pandey, J. R. McBride, S. Savarese, and R. M. Eustice, "Automatic extrinsic calibration of vision and lidar by maximizing mutual information," *Journal of Field Robotics*, vol. 32, no. 5, pp. 696–722, 2015.

[9] J. Levinson and S. Thrun, "Automatic online calibration of cameras and lasers." in *Robotics: Science and Systems*, vol. 2, 2013.

[10] T. Cui, S. Ji, J. Shan, J. Gong, and K. Liu, "Line-based registration of panoramic images and lidar point clouds for mobile mapping," *Sensors*, vol. 17, no. 1, p. 70, 2016.

[11] Z. Taylor and J. Nieto, "Motion-based calibration of multimodal sensor extrinsics and timing offset estimation," *IEEE Transactions on Robotics*, vol. 32, no. 5, pp. 1215–1229, 2016.

[12] P. Viola and W. Wells III, "Alignment by Maximization of Mutual Information," *International Journal of Computer Vision*, vol. 24, no. 2, pp. 137–154, 1997.

[13] Z. Taylor and J. Nieto, "A mutual information approach to automatic calibration of camera and lidar in natural environments," in *Australian Conference on Robotics and Automation*, 2012, pp. 3–5.

[14] K. Irie, M. Sugiyama, and M. Tomono, "Target-less camera-lidar extrinsic calibration using a bagged dependence estimator," in *Automation Science and Engineering (CASE), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1340–1347.

[15] Z. Taylor, J. Nieto, and D. Johnson, "Multi-Modal Sensor Calibration Using a Gradient Orientation Measure," *Journal of Field Robotics*, vol. 32, no. 5, pp. 675–695, 2015.

[16] A. Banno and K. Ikeuchi, "Omnidirectional texturing based on robust 3d registration through euclidean reconstruction from two spherical images," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 491–499, 2010.

[17] Y. C. Shiu and S. Ahmad, "Calibration of wrist-mounted robotic sensors by solving homogeneous transform equations of the form AX= XB," *ieee Transactions on Robotics and Automation*, vol. 5, no. 1, pp. 16–29, 1989.

[18] I. Fassi and G. Legnani, "Hand to sensor calibration: A geometrical interpretation of the matrix equation AX= XB," *Journal of Field Robotics*, vol. 22, no. 9, pp. 497–506, 2005.

[19] J. D. Hol, T. B. Schön, and F. Gustafsson, "Modeling and calibration of inertial and vision sensors," *The international journal of robotics research*, vol. 29, no. 2-3, pp. 231–244, 2010.

[20] J. Kelly and G. S. Sukhatme, "Visual-inertial sensor fusion: Localization, mapping and sensor-to-sensor self-calibration," *The International Journal of Robotics Research*, vol. 30, no. 1, pp. 56–79, 2011.

[21] L. Heng, B. Li, and M. Pollefeys, "Camodocal: Automatic intrinsic and extrinsic calibration of a rig with multiple generic cameras and odometry," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 1793–1800.

[22] T. Oishi, A. Nakazawa, R. Kurazume, and K. Ikeuchi, "Fast Simultaneous Alignment of Multiple Range Images using Index Images," *Proc. of 5th International Conference on 3-D Digital Imaging and Modeling (3DIM2005)*, pp. 476–483, 2005.

[23] P. F. Alcantarilla and T. Solutions, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," *IEEE Trans. Patt. Anal. Mach. Intell*, vol. 34, no. 7, pp. 1281–1298, 2011.

[24] D. Nister, O. Naroditsky, and J. Bergen, "Visual odometry for ground vehicle applications," *Journal of Field Robotics*, vol. 23, no. 1, pp. 3–20, 2006.

[25] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[26] A. Pagani and D. Stricker, "Structure from motion using full spherical panoramic cameras," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 375–382.

[27] B. D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," in *International Joint Conference on Artificial Intelligence*, 1981, pp. 674–679.

[28] L. Kneip, D. Scaramuzza, and R. Siegwart, "A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2969–2976.