IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 January 2023, accepted 27 January 2023, date of publication 13 February 2023, date of current version 1 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3244212

METHODS

Accuracy Evaluation and Prediction of Single-Image Camera Calibration

SUSUMU KIKKAWA[1,2], FUMIO OKURA[1], (Member, IEEE), DAIGO MURAMATSU[3], (Member, IEEE), YASUSHI YAGI[4], (Senior Member, IEEE), AND HIDEO SAITO[5], (Senior Member, IEEE)

[1]Graduate School of Information Science and Technology, Osaka University, Suita, Osaka 567-0871, Japan
[2]Forensic Science Laboratory, Osaka Prefectural Police, Chuo, Osaka 540-8540, Japan
[3]Faculty of Science and Technology, Seikei University, Musashino, Tokyo 180-8633, Japan
[4]SANKEN (The Institute of Scientific and Industrial Research), Osaka University, Osaka 567-0047, Japan
[5]Department of Information and Computer Science, Keio University, Kouhoku, Yokohama 223-8522, Japan

Corresponding author: Susumu Kikkawa (kikkawa@am.sanken.osaka-u.ac.jp)

This work was supported in part by JST Fusion Oriented REsearch for disruptive Science and Technology (FOREST) Grant Number JPMJFR206F and JSPS Grants-in-Aid for Scientific Research (KAKENHI) Grant Number JP21H03466.

ABSTRACT This paper proposes an application to statistically predict the accuracy of single-image geometric camera calibration that uses given 2D-3D correspondences. Deriving both camera intrinsics and extrinsics from correspondences between a single image and a 3D shape, is important for the scene analysis when the optical system of the camera is lost, such as in the analyses of traffic accidents. It is unclear whether the single-image calibration will be successful in practice, particularly when the number of 2D-3D correspondences is small, even if we could assign accurate correspondences by manual labor. To this end, we perform a systematic evaluation of the camera parameter accuracy using synthetic environments. Based on the statistics observed during the experiments, our application predicts the calibration accuracy from simple variables (*e.g.*, the area that correspondences could be given). Since the prediction process does not rely on 3D shapes, it provides an estimate of the success of the calibration before time-consuming processes, *i.e.*, 3D scanning and 2D-3D correspondence mapping.

INDEX TERMS Camera calibration, traffic accident reconstruction, computer vision.

I. INTRODUCTION

Camera calibration of intrinsic and extrinsic parameters is a traditional yet essential problem in computer vision. Given only a single image and a three-dimensional (3D) model of the environment, the problem of finding the camera parameters becomes practically challenging, even though it is of practical importance.

A major application of single-image camera calibration with known 3D geometry is to estimate the camera parameters for images capturing traffic accidents, which can be important evidence. During traffic accident reconstruction (TAR) [1], investigators often estimate the position and movement of cars and pedestrians from the given *evidence* images [2], [3], which are often captured by dashboard cameras or smartphones. Since the camera parameters of off-the-shelf cameras are generally unknown, camera calibration from the single evidence image is fundamental for the accurate reconstruction of the target scene [4]. Static scene geometry may be obtained after the incident with a 3D scanner, while it is difficult to reproduce the same optical system as when the image was taken because the camera can be damaged or misaligned.

Geometric calibration methods [5], [6] have been widely used to estimate camera parameters. In particular, the methods without relying on planar markers (*e.g.*, Tsai's method [5]) can, in principle, be used to estimate both intrinsic and extrinsic parameters using the correspondences on an arbitrary 3D model and a two-dimensional (2D) image. Methods tailored for single-image calibration are also studied [7]. These geometric methods are known to be valid if the good (in terms of both numbers and quality) correspondences are obtained [8], [9].

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Liu.

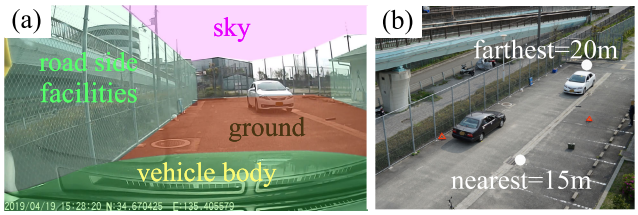


FIGURE 1. Challenges in single-image camera calibration in traffic accident reconstruction (TAR).

It is, however, uncertain that the calibration with a single image that captures an urban landscape will be successful, which is often the case of the target scenes of TAR as shown in Fig. 1. Even if accurate correspondences are given by manual labor, the number of correspondences will be smaller and tends to be biased in the image and 3D space, compared to using a well-designed rig. The image captured with a dashboard camera (see Fig. 1(a)) contains large areas that we cannot yield the correspondences (*e.g.*, the sky at infinity, or a vehicle body moving along with the camera). Besides, most of the possible corresponding points are often distributed on three planar surfaces, *i.e.*, the ground and roadside facilities such as buildings. They are poorly textured, thus limiting the distribution of the possible corresponding points on the image. In the case of overlooking a scene (see Fig. 1(b)), which often occur in security camera images, correspondences can be found in various part on the image; however, since the variation of distance of the scene is small, it may affect the calibration accuracy. These issues can lead to inaccurate or unstable solutions when estimating the position or speed of cars or pedestrians at the time of the accident.

Our goal is to develop an application that statistically predicts the accuracy of single-image camera calibration via comprehensive and systematic experiments. In this study, we use a traditional geometric calibration method. Since the instability of the solution is inevitable in single-image cases with relatively few correspondences, we employ a practical technique to obtain a reasonable solution by sampling the initial intrinsics. We evaluate the calibration accuracy with changing various factors (*e.g.*, the number of correspondence and distribution in the image) by generating 2D-3D correspondences in synthetic environments. Based on the experimental results, we develop an application that predicts the calibration accuracy from an image and simple additional information.

Our application is intended to provide users a guide to acquiring enough information for scene analysis by estimating the success of the camera calibration for a given set of variables. Given an image, it helps to make decisions on whether or not to engage in time-consuming processes in 3D shape acquisition (*e.g.*, via 3D scanners) and 2D-3D correspondence mapping (*e.g.*, via manual labor). Also, it can be used to estimate the number of points or coverage of 3D shapes to meet the required accuracy. To assess the reliability of TAR and its admissibility as evidence at the court, predicting the accuracy of the camera parameter estimation, as well as its stability and confidence intervals, are quite important. Through experiments in a real-world environment that mimics traffic accidents, we show that the predicted accuracy is well in line with the practical scenarios. Our implementation is available at https://github.com/Kikkawa-OPP/CalibPrediction.[1]

A. CONTRIBUTION

We provide a practical analysis of single-image calibration accuracy and its confidence interval depending on various factors, which emphasizes the application to predict the calibration accuracy in practical environments. The prediction process does not require 3D shapes and 2D-3D correspondences; it thus provides the estimation of the success of the calibration before time-consuming processes (*i.e.*, 3D scanning and correspondence mapping).

II. RELATED WORKS

Our goal is to provide a quantitative measure and a prediction tool of the camera calibration accuracy through systematic evaluation. Our work is thus closely related to camera calibration and its evaluation.

A. GEOMETRIC CALIBRATION

Geometric calibration of camera intrinsic and extrinsic parameters is a fundamental technique in computer vision [10], [11]. Intrinsic parameters are often estimated using the correspondences on known 3D geometry [5] or planes [6]. Several different models for camera intrinsics have been proposed, such as using sixth-order radial distortion [12] or tangental distortion [13]. These methods, in principle, optimize both intrinsics and extrinsics, thus can also be used for extrinsic parameter estimation against a known geometry.

The perspective-n-point (PnP) problem [14], [15], [16] is to estimate the extrinsic parameters with given intrinsics and 2D-3D correspondences. The PnP problem can be extended to estimate (a part of) intrinsics like focal length [17], [18]. Recent studies use deep learning to estimate extrinsic parameters for the alignment between a camera and a depth sensor [19]. Though slightly different from the calibration using 2D-3D correspondences, recent structure-from-motion (SfM) methods targeting unordered image collection often compute camera intrinsics, as well as extrinsics, using numbers of image correspondences [20], [21].

B. SINGLE-IMAGE FULL CALIBRATION

Even the 2D-3D correspondences on a single image are given, traditional methods (*e.g.* [5]) can still achieve a full calibration (*i.e.*, estimation of both intrinsics and extrinsics). In practice, this task is often done using calibration rigs [22], which enables detecting plenty of correspondences. There are also methods specialized for single-image calibration using orthogonal planes [23] or lines [24]. Similarly, circles [25], vanishing points [7], [26], and low-rank textures [27] are

[1]The code will be uploaded upon publication.

known to be useful assumptions for single-image calibration. In the augmented reality application, fiducial markers are designed for estimating intrinsic parameters [28]. To avoid the lack of generalization, we will investigate a general method for geometric calibration, which does not rely on planar scenes or any other special assumptions.

Recently, single-image calibration using deep learning is also studied. Some recent methods do not require 3D shapes but relying on numbers of 2D training images [29]. They, however, only calculate a rough extrinsics (*i.e.*, place recognition) or partial rotation (*i.e.*, only roll and pitch) [29], [30], which is difficult for the use in TAR applications.

C. EVALUATION OF CALIBRATION ACCURACY

The accuracy evaluation of camera calibration has a close relation with ours. Early attempts include the comparison of distortion models [31] and different calibration methods [32], [33]. The influences of measurement noise were also investigated as an important factor for calibration accuracy [34]. Several studies perform task-oriented evaluation, which assesses the influence of calibration error to the accuracy of stereo vision [35], [36] or 3D reconstruction [37], [38], [39]. Also, a recent paper [40] seeks the camera calibration options suitable for autonomous driving applications.

Similar to our work, Sun and Cooperstock [9] provide a systematic evaluation of traditional calibration methods. The experiment was carried out using synthetic 3D models to assess the influence of noises and the number of correspondences. However, the previous study focused on the use of well-designed rigs (*i.e.*, 3D patterns, or 2D checkerboard) captured with multiple images. For single-image calibration with urban scenes, the problem becomes notably challenging in terms of both the number of correspondences and the distribution of the corresponding points.

III. SINGLE-IMAGE CAMERA CALIBRATION WITH INITIAL PARAMETER SEARCH

Although proposing a calibration method is not our main contribution, we here introduce a practical technique for single-image camera calibration, which is used in our experiment.

Our method estimates both the intrinsic and extrinsic parameters from given 2D-3D correspondences defined between a single image and a 3D point cloud such as acquired by a 3D scanner. When performing gradient-based nonlinear minimization like the Levenberg-Marquardt (LM) method [41], local minima far from the actual solution are likely to be derived if the number of corresponding points is small or if there is a bias in the distribution in the image. It is possible to obtain the initial values of intrinsic parameters by solving a linear system using singular value decomposition (SVD) and other methods [5], [6]. However, it is easily assumed that both the linear solution by minimizing the algebraic distance and the local solution of the re-projection error by the nonlinear least-squares method will be unstable, especially when only a few correspondences are given.

In this study, we restrict the solution space by sampling the initial values of the parameters that largely affect the re-projection error. We use a grid search of focal length $\mathbf{f} = (f_x, f_y)$ and the second-order radial distortion k_1, and search the best intrinsic parameters that minimize the re-projection error using the LM method. During the grid search, we first optimize the remaining parameters while fixing $\{\mathbf{f}, k_1\}$ at the grid point, then optimize the all parameters using the given solution as the initial guess.

A. IMPLEMENTATION DETAILS

Since the calibration functions implemented in OpenCV are commonly used nowadays, we employ the camera model based on the definitions in OpenCV, which is slightly different from Tsai's model [5] but includes the sixth-order radial distortion [12] and tangental distortion [13]. We thus used the intrinsic parameters consist of focal length $\mathbf{f} = (f_x, f_y)$, principal point $\mathbf{c} = (c_x, c_y)$, and the distortion coefficients including three radial $\mathbf{k} = (k_1, k_2, k_3)$ and two tangental distortion terms $\mathbf{p} = (p_1, p_2)$. Similar to traditional methods, this study uses the mean square of the re-projection error as the objective function and alternatingly optimizes the extrinsic and intrinsic parameters by the LM method [41].

For grid search, we sample the initial focal length $\mathbf{f} = (f_x, f_y)$ converted from the vertical field of view (FoV) α.

$$f_x = f_y = \frac{\frac{\mathcal{I}_h}{2}}{\tan\frac{\alpha}{2}}, \tag{1}$$

where $\mathcal{I}_h$ denotes the height of the image. The FoV α is searched in the range $[10°, 170°]$ with an interval of $10°$. For the distortion coefficients, we initially set $\mathbf{k} = (k_1, \mathbf{0})$, $\mathbf{p} = \mathbf{0}$ and sample k_1 in the range $[-10, 10]$ with the interval of 0.5. The initial values of the principal point $\mathbf{c}$ are equivalent to the image center. Although a rough FoV may be obtained from the specification information of the camera used, the focal length converted from the FoV using Eq. (1) becomes notably different from the actual value when the radial distortion is large, thus it is useful to search for the focal length using the grid search.

Our implementation is based on `calibrateCamera` function in OpenCV. With our single-threaded Python implementation, the whole process took 1.9 [sec] on a CPU (Intel i7-8700K, 3.70 GHz) when 50 correspondences were given.

IV. SYSTEMATIC EVALUATION

The heart of this study is to analyze the error factors for single-image geometric camera calibration to develop an error prediction application. Previous studies reported that the noise and the number of correspondences affects the calibration accuracy via experiments on synthetic 2D-3D correspondences simulating 3D rigs [9]. We also investigate the factors that are likely affecting when using correspondences on a single urban image, such as the scene geometry and the correspondence distribution on the image.

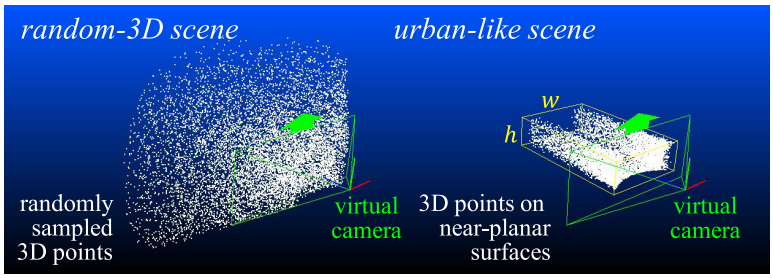


FIGURE 2. Synthetic environments (densely sampled for visualization).

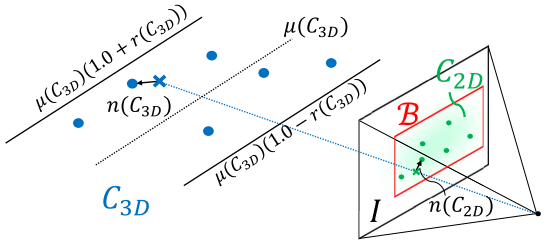


FIGURE 3. Correspondence generation.

TABLE 1. The list of variables used in the systematic experiment.

Variables	Description	Min	Max	Interval
$\frac{\#(\mathcal{B})}{\#(\mathcal{I})}$	Bounding box size	0.1	1.0	0.1
$\rho(\mathcal{C}_{2D})$	Correspondence density (par 100 × 100 pixels)	0.1	1.0	0.1
$\mu(\mathbf{d}(\mathcal{C}_{3D}))$	Representative depth [m]	10.0	100.0	10.0
$r(\mathbf{d}(\mathcal{C}_{3D}))$	Depth range	0.1	1.0	0.1
$n(\mathcal{C}_{2D})$	Noise on 2D points [px]	0.0	5.0	0.5
$n(\mathcal{C}_{3D})$	Noise on 3D points [cm]	0.0	5.0	0.5
w	Road width [m]	5.0	50.0	5.0
h	Building height [m]	2.0	20.0	2.0

A. SYNTHETIC ENVIRONMENTS

To assess the influences of the variables (*e.g.*, number of correspondences) under both ideal and practical scenarios, we use two types of synthetic 3D scenes: *random-3D* and *urban-like* scenes as shown in Figure 2. We generate these types of scenes to automatically acquire 2D-3D correspondences by randomly selecting 3D points from the 3D scenes and projecting them to the virtual camera. In order to obtain reasonable results, the pose of the virtual camera was randomly set each time (translations ranged from 0m to 100m)

1) RANDOM-3D SCENE

To analyze in a similar setting with a previous study [9], we prepare *random-3D* scenes. For this type of scenes, we randomly select the 3D object points in the view frustum of the virtual camera, to fulfill the given variables listed in sec. IV-B. The *random-3D* scenario simulates an *ideal* case for 2D-3D correspondences generation, intending to evaluate the overall trend of the influence of the variables on the calibration accuracy, which does not rely on scene geometry.

2) URBAN-LIKE SCENE

We also create *urban-like* scenes that simulate practical scenarios in TAR, which usually deal the roadside images captured by, *e.g.*, dashboard cameras. As discussed in Sec. I, we assume the scene geometry of the roadside is mostly composed of three orthogonal planes (*i.e.*, roads and buildings). Since the road and building surfaces are not perfectly planar, we add the random noises on the point locations with the standard deviation of 5 [cm] for the road, and 100 [cm] for the building. In this analysis, we fix the height of the virtual camera at 1.5 [m]. This represents the approximate height of the rear-view mirror, which is the location of the dashboard camera in most cars. We evaluate the calibration accuracy by changing the road width and the height of the buildings.

For both scenes, the intrinsic parameters of virtual cameras simulate an actual dashboard camera with a resolution of 1920 × 1080. We use a wide FoV camera, in which $\mathbf{f} = (1000, 1010)$ and $\mathbf{k} = (-0.3, 0.1, 0.0)$. We add slight tangential distortions $\mathbf{p} = (0.02, 0.01)$ and shifted principal point $\mathbf{c} = (1020, 560)$, which often occur by windshields. To avoid the effect of the initial guess of parameter estimation, we randomly translated and rotated the entire scene (*i.e.*, both 3D points and the camera) before each computation.

B. VARIABLES

For the synthetic experiments, we consider the scene shape and the quality of 2D-3D correspondence could influence the calibration accuracy. We, therefore, systematically evaluate the calibration accuracy by changing these variables. Specifically, based on the existing study for 3D rigs [9], we assume the calibration accuracy relies on the following factors: correspondence distribution, scene geometry, and noises.

Table 1 summarizes the range and intervals of the variables we assess during the experiment, which are used to construct the statistics of calibration accuracy. These variables are related to the distribution and the quality of correspondences, as well as the scene characteristics. Given the set of variables, we randomly sample the 3D points from the point cloud that are projected in a given bounding box, as depicted in Fig. 3.

We here denote the set of correspondences on 2D image points and 3D object points as $\mathcal{C}_{2D}$ and $\mathcal{C}_{3D}$, respectively. We use the functions to represent the range r, density ρ, average μ, and noise level n of the given set of correspondences. Also, # counts the members of a given set. The variables are now defined as follows.

1) SIZE OF BOUNDING BOX CONTAINING IMAGE POINTS $\frac{\#(\mathcal{B})}{\#(\mathcal{I})}$

This variable affects the distribution of 2D image points. Let $\mathcal{I}$ and $\mathcal{B}$ as the sets of the pixels representing the image and a bounding box that contains the corresponding points $\mathcal{C}_{2D}$ (*i.e.*, we only sampled the correspondences in the given bounding box). The size of the bounding box $\frac{\#(\mathcal{B})}{\#(\mathcal{I})} \in [0, 1]$ is denoted as the ratio of the area (*i.e.*, the number of pixels) of the bounding box $\#(\mathcal{B})$ in the whole image $\#(\mathcal{I})$. During the experiment, we randomly generated bounding boxes which match the designated area ratio $\frac{\#(\mathcal{B})}{\#(\mathcal{I})}$.

2) CORRESPONDENCE DENSITY $\rho(\mathcal{C}_{2D})$

We control the number of the corresponding points in the given bounding box. To simplicity, we denote the correspondence density $\rho(\mathcal{C}_{2D})$ as the number of points par $100 \times 100 = 10,000$ pixels (when using cameras with 1920×1080 resolutions).

$$\rho(\mathcal{C}_{2D}) = \#(\mathcal{C}_{2D}) \frac{100 \times 100}{\#(\mathcal{B})}. \tag{2}$$

3) REPRESENTATIVE DEPTH $\mu(\mathbf{d}(\mathcal{C}_{3D}))$ [m] AND RANGE $r(\mathbf{d}(\mathcal{C}_{3D}))$

Letting the depth values of the object points $\mathcal{C}_{3D}$ as $\mathbf{d}(\mathcal{C}_{3D})$, we randomly sample the object points in the depth range defined as

$$[\mu(\mathbf{d}(\mathcal{C}_{3D}))(1.0 - r(\mathbf{d}(\mathcal{C}_{3D}))),\ \mu(\mathbf{d}(\mathcal{C}_{3D}))(1.0 + r(\mathbf{d}(\mathcal{C}_{3D})))], \tag{3}$$

where $\mathbf{d}(\mathcal{C}_{3D})$ denotes the depth values of the set of 3D points. Thus, $\mu(\mathbf{d}(\mathcal{C}_{3D}))$ and $r(\mathbf{d}(\mathcal{C}_{3D}))$ denote the representative value and the range of the depth as illustrated in Fig. 3.

4) NOISE ON CORRESPONDENCES $n(\mathcal{C}_{2D})$ [px], $n(\mathcal{C}_{3D})$ [cm]

To the given 2D and 3D corresponding points, we respectively add the Gaussian noises based on the standard deviation defined as $n(\mathcal{C}_{2D})$ and $n(\mathcal{C}_{3D})$. While the noises on 2D image points $n(\mathcal{C}_{2D})$ simulate the errors of feature point detection or manual correspondence assignment, $n(\mathcal{C}_{3D})$ simulates the 3D measurement error during laser scan or the fusion of multiple scans. The direction of the noise vectors is randomly selected.

5) ROAD WIDTH w [m] AND BUILDING HEIGHT h [m] (FOR *URBAN-LIKE* SCENES)

During the experiment using *urban-like* scenes, we also control the scene characteristics. Since the scenes simulate the road-side scenario, we change the road width w and the height of the road-side buildings h.

C. EVALUATION METRICS

We evaluate the accuracy of intrinsics and extrinsics using the well-known measures: The re-projection error, the camera position error, and the orientation error.

1) RE-PROJECTION ERROR (GIVEN CORRESPONDENCE) $Re_{\mathcal{C}_{2D}}$ [px]

We compute the root mean square (RMS) of the re-projection error $Re_{\mathcal{C}_{2D}}$ on the given 2D-3D correspondences.

2) RE-PROJECTION ERROR (ENTIRE IMAGE) $Re_{\mathcal{I}}$ [px]

We evaluate the RMS of the re-projection error on equally-distributed pixels (*i.e.*, grid points at 10 [px] intervals) in the image, $Re_{\mathcal{I}}$, not only on the given correspondences. This is computed by the back projection of the pixels to the representative depth of the scene, $\mu(\mathbf{d}(\mathcal{C}_{3D}))$.

3) RE-PROJECTION ERROR (BOUNDING BOX) $Re_{\mathcal{B}}$ [px]

In practical TAR scenarios, *e.g.*, to estimate vehicle position, it is often sufficient to be accurately calibrated in the image region around the vehicle. In this case, accurate re-projection errors may not necessarily be required for the entire image. We, therefore, evaluate the RMS of the re-projection error at grid points *inside* the bounding box, $Re_{\mathcal{B}}$, which contains the corresponding points.

4) POSITIONAL ERROR E_{pos} [cm]

To evaluate the accuracy of the extrinsics, we calculate the positional error of the estimated camera E_{pos}. This measure is useful for predicting the accuracy of the self-localization of vehicles.

5) ORIENTATION ERROR E_{ori} [deg]

Similar to the positional error, we also evaluate the orientation of the camera E_{ori}. This is computed as the angle between the optical axes of the estimated and the ground-truth camera.

D. RESULTS

Figure 4 shows the errors on the correspondence projection $Re_{\mathcal{C}_{2D}}, Re_{\mathcal{I}}, Re_{\mathcal{B}}$ and extrinsics E_{pos}, E_{ori} while changing variables for both scenes. The results shown were generated by changing each of variables independently, while fixing the other parameters as a set of the default values $\frac{\#(\mathcal{B})}{\#(\mathcal{I})} = 1.0$, $\rho(\mathcal{C}_{2D}) = 0.5$, $\mu(\mathbf{d}(\mathcal{C}_{3D})) = 20$, $r(\mathbf{d}(\mathcal{C}_{3D})) = 0.6$, $n(\mathcal{C}_{2D}) = 0.7$, $n(\mathcal{C}_{3D}) = 0.5$, $w = 10$, $h = 5$. For a given set of variables, we repeated the estimation $1,000$ times using random correspondences, and the figure shows the mean and confidence intervals calculated from these samples. Overall, the results of two (*random-3D* and *urban-like*) scenes share a similar tendency, while the error scale is different due to the different correspondence distribution on 3D space and 2D planes. We then introduce detailed discussions related to the correspondence distribution, depth variation, noises, and scene geometry.

1) CORRESPONDENCE DISTRIBUTION ON 2D IMAGES

If the bounding box size is relatively small (*e.g.*, $\frac{\#(\mathcal{B})}{\#(\mathcal{I})} < 0.6$), the estimate outside the bounding box is quite unstable. Meanwhile, even in more challenging cases, the estimate inside the bounding box and the camera localization accuracy

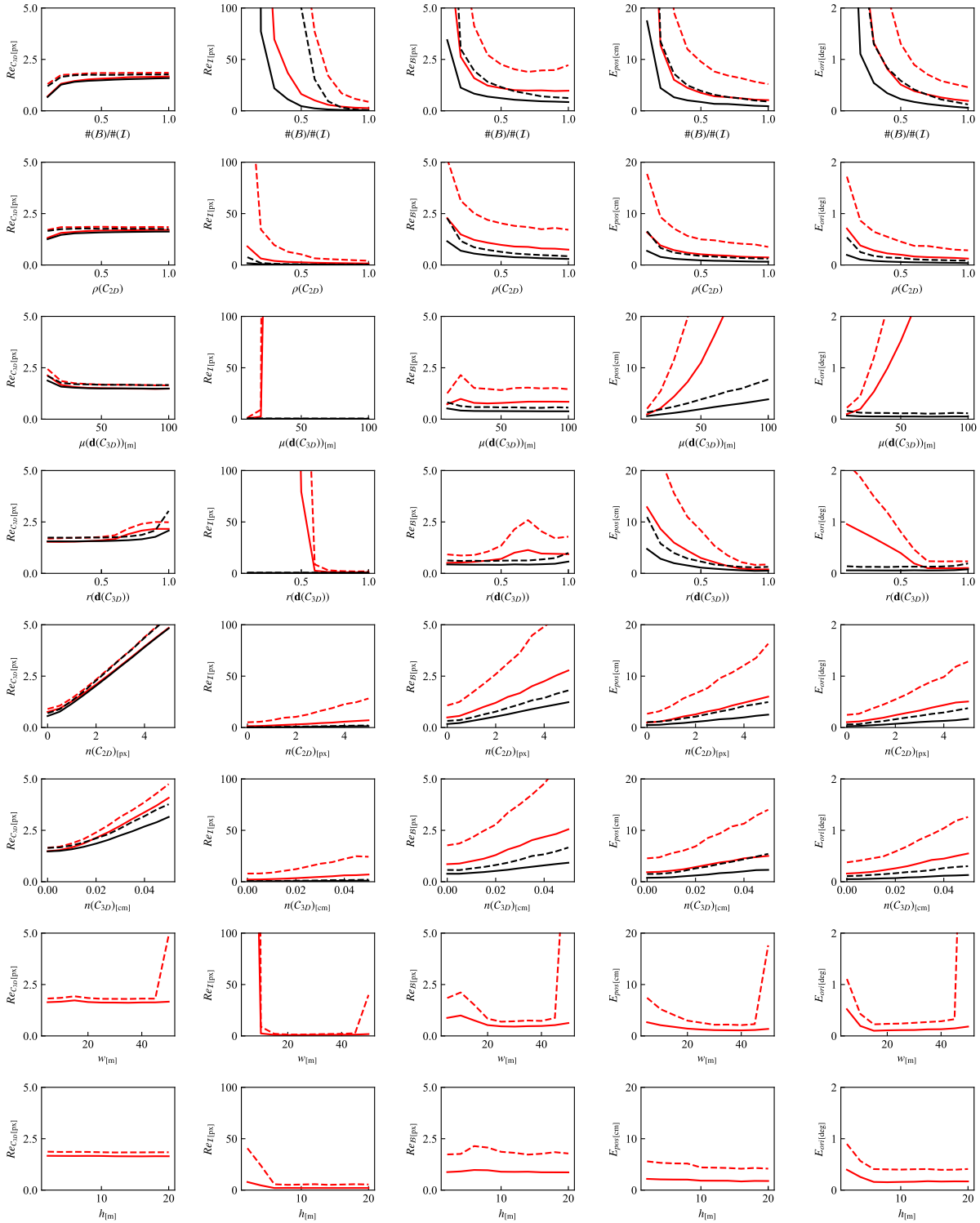


FIGURE 4. Calibration errors in synthetic environments. Black lines: Results using *random-3D* scenes. Red lines: Results using *urban-like* scenes. Each plot visualizes the 50 percentile (median; a solid line) and 95 percentile (confidence interval; a dashed line) values calculated from 1,000 samples. The results shown were generated by changing each of variables independently, while fixing the others as a set of the default values $\frac{\#(\mathcal{B})}{\#(\mathcal{I})} = 1.0$, $\rho(\mathcal{C}_{2D}) = 0.5$, $\mu(\mathbf{d}(\mathcal{C}_{3D})) = 20$, $r(\mathbf{d}(\mathcal{C}_{3D})) = 0.6$, $n(\mathcal{C}_{2D}) = 1.5$, $n(\mathcal{C}_{3D}) = 1.0$, $w = 10$, $h = 5$.

can be improved by giving larger numbers of correspondences $\rho(\mathcal{C}_{2D})$.

For further investigation, Fig. 5 visualizes the spatial distribution of the re-projection errors on the image plane by changing both $\frac{\#(\mathcal{B})}{\#(\mathcal{I})}$ and $\rho(\mathcal{C}_{2D})$. If the bounding box has an enough size ($\frac{\#(\mathcal{B})}{\#(\mathcal{I})} \geq 0.3$ for *random-3D* scenes) and contains the enough number of correspondences (*e.g.*, $\rho(\mathcal{C}_{2D}) \geq 3.0$), the size of bounding boxes $\frac{\#(\mathcal{B})}{\#(\mathcal{I})}$ did not notably influence

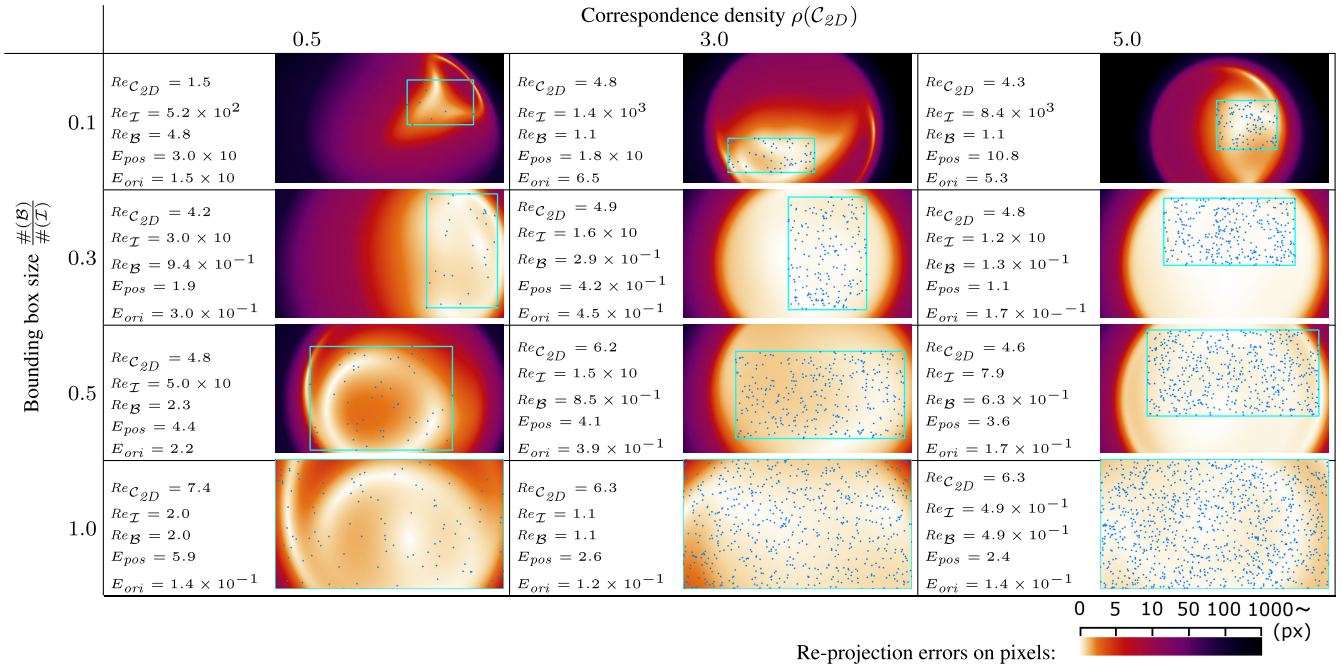


FIGURE 5. **Visualization of the re-projection errors for different correspondence distribution (*random-3D* scene).**

the re-projection error *inside* the box $Re_{\mathcal{B}}$. In such cases, the camera localization accuracy E_{pos}, E_{ori} are also acceptable (*i.e.*, $E_{pos} < 5$ [cm]) intending that the correspondences only given around the regions of interest can be used for TAR scenarios only using a local part of the image. Restricting the bounding boxes of correspondences is theoretically the same as to use narrow-FoV cameras; in our case, $\frac{\#(\mathcal{B})}{\#(\mathcal{I})} = 0.3$ is converted to the vertical FoV of approximately 30°, if they share the same principal point and aspect ratio.

2) DEPTH VARIATIONS

The depth variation affect the accuracy of extrinsics. Larger depth $\mu(\mathbf{d}(\mathcal{C}_{3D}))$ with a smaller range $r(\mathbf{d}(\mathcal{C}_{3D}))$ leaded the larger camera positional errors E_{pos}. Only in *urban-like* scenes, larger scene depths also led the inaccurate guess for camera orientation E_{ori} and re-projection errors $Re_{\mathcal{I}}$. The cause is related to the scene characteristics. Especially when the road width is narrow, it is difficult to obtain the correspondences far from the camera on the large areas in the image plane. This is a similar effect when decreasing the size of the bounding box $\frac{\#(\mathcal{B})}{\#(\mathcal{I})}$.

3) NOISES

Noises $n(\mathcal{C}_{2D}), n(\mathcal{C}_{3D})$ also influence the overall accuracy, as reported in [9]. The relationship between the noise levels and the errors was almost linear.

4) SCENE GEOMETRY

The scene geometry slightly influenced the stability of the calibration. Regarding the confidence interval, the estimation in narrow (*e.g.*, < 5 [m]) or wide (*e.g.*, > 45 [m]) roads sometimes, slightly, drop the accuracy related to the re-projection errors and positional errors. In such cases, the scene can be approximated as a plane.

V. PREDICTION OF CALIBRATION ACCURACY

Based on the systematic experiments, we can easily develop an application that predicts the calibration accuracy. For TAR-related analyses such as the localization of vehicles, it is necessary to measure the scene geometry (*e.g.*, using laser scanners), which may lead to traffic restrictions. For practical use cases, therefore, our application leverages users' prior knowledge of the target scene, *without* acquiring 3D point clouds and 2D-3D correspondences. It can be used for primary screening of images as possible evidence of the incident (*e.g.*, traffic accident) to avoid or minimize the burden of investigators and society. Also, demonstrating accuracy with confidence interval is important to ensure the reliability of evidence at court.

In this section, we describe the detail of the application software as well as experiments using real-world scenes using our software.

A. APPLICATION DETAILS

As shown in Fig. 6, we suppose to use our application in pre-survey of geometric camera calibration. The application estimates the mean and confidence interval of the calibration errors from the scene type (*random-3D* or *urban-like*) and variables used in our experiment (see Tab. 1). This is simply doable via the systematic evaluation using the synthetic correspondences that fulfill the given variables as we showed in the previous sections. While it predicts the success of calibration for given variables, it can also be used to estimate the lower or upper bounds of variables that meet a required accuracy

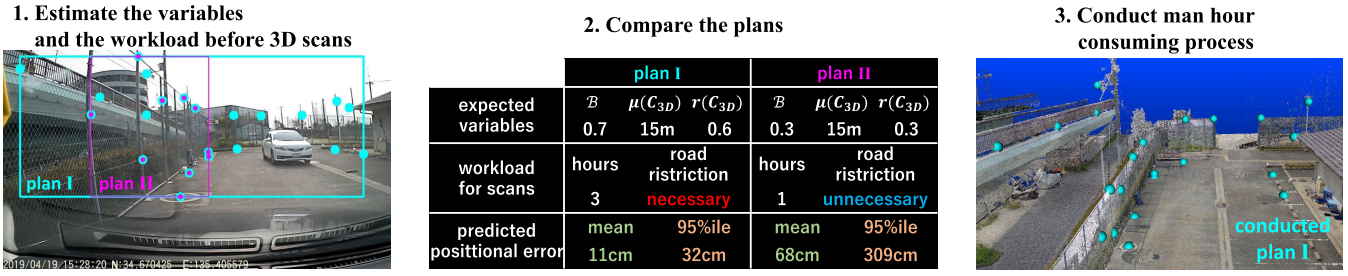
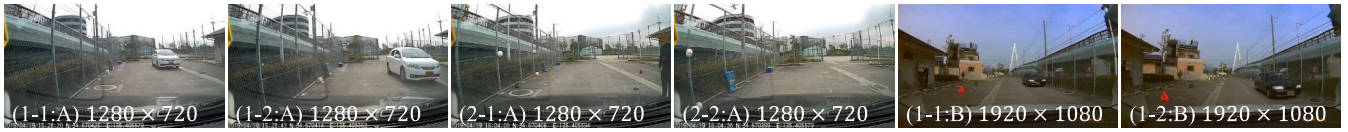
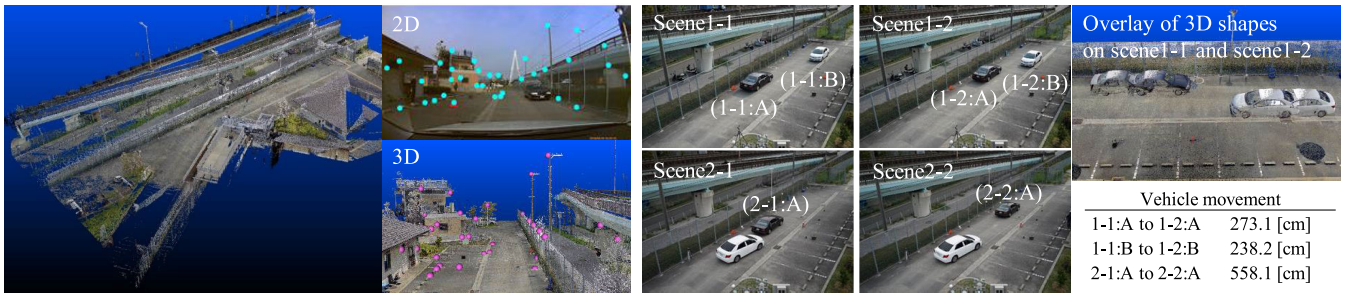


FIGURE 6. A use case of our application. It can be used to predict the calibration accuracy from multiple measurement plans and examine whether TAR can be conducted with reasonable accuracy.



(a) Input images for the real-world experiment.



(b) Scene points and example correspondences.

(c) The ground truth of vehicle movement.

FIGURE 7. Dataset for the real-world experiment. As input data for our accuracy prediction software, we acquire (a) images from dashboard cameras, (b) manual correspondences between 3D models and images. We also measure (c) the ground-truth camera movement acquired by a 3D scanner for the evaluation purpose.

TABLE 2. Results of the real-world experiment compared to the prediction by our application. Each row shows a set of input, prediction, and actual results for an image. For all images, $\{n(\mathcal{C}_{2D}), n(\mathcal{C}_{3D}), w, h\}$ were set to $\{1.0, 1.5, 10, 3\}$.

Img	Variables				Prediction by *urban-like* simulation *without* 3D acquisition										Real-world results		Percentile	
					Mean					Upper confidence intervals								
	$\rho(\mathcal{C}_{2D})$ $[\#\mathcal{C}_{2D}]$	$\frac{\#(\mathcal{B})}{\#(\mathcal{I})}$	$\mu(\mathbf{d}(\mathcal{C}_{3D}))$	$r(\mathbf{d}(\mathcal{C}_{3D}))$	$Re_{\mathcal{C}_{2D}}$	$Re_{\mathcal{I}}$	$Re_{\mathcal{B}}$	E_{pos}	E_{ori}	$Re_{\mathcal{C}_{2D}}$	$Re_{\mathcal{I}}$	$Re_{\mathcal{B}}$	E_{pos}	E_{ori}	$Re_{\mathcal{C}_{2D}}$	E_{pos}	$Re_{\mathcal{C}_{2D}}$	E_{pos}
1-1:A	0.24 [22]	0.57	15	0.8	1.1	7.2	4.2	9.5	1.1	1.5	25.1	13.6	25.8	2.7	1.1	1.2	55.3	0.3
1-2:A	0.18 [17]	0.47	15	0.8	1.0	28.5	5.9	14.5	1.9	1.3	274.3	30.1	45.6	4.9	1.2		80.2	0.2
1-1:B	0.16 [34]	0.55	30	0.8	1.6	5.6	3.5	4.2	0.5	2.1	14.6	8.9	11.1	1.2	1.4	2.3	21.2	22.9
1-2:B	0.15 [31]	0.61	30	0.8	1.6	5.9	3.1	3.3	0.5	2.0	18.7	8.1	9.5	1.2	1.6		56.3	29.9
2-1:A	0.24 [22]	0.61	12	0.6	1.1	5.5	3.5	7.8	1.0	1.5	15.1	9.8	22.2	2.3	1.2	5.6	70.4	31.7
2-2:A	0.26 [24]	0.60	12	0.6	1.2	6.9	3.5	8.4	0.8	1.5	26.7	10.5	22.1	2.0	1.2		58.8	27.0

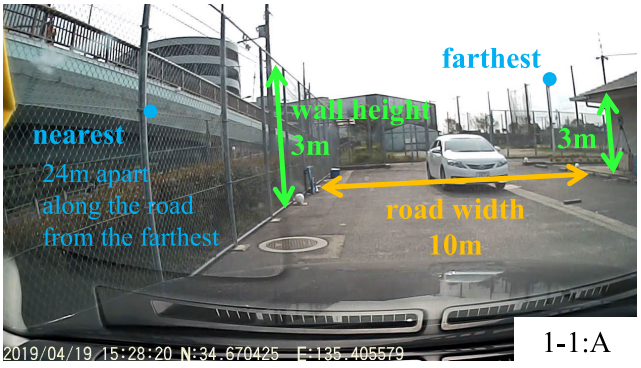
in principle. The application, for example, can be used to predict the minimum number of correspondence or coverage of 3D scans to meet the given accuracy requirement, which contribute to minimizing manual labor and traffic restrictions. While we can use the pre-computed error statistics yielded during the previous experiments, re-computing the statistics for new camera settings (*e.g.*, for different resolution or largely different FoV) is possible by a reasonable time (approximately 10 minutes for 1, 000 trials when parallelized on a CPU, Intel i7-8700K, 3.70 GHz, 6 cores, 12 threads).

B. REAL-WORLD EXPERIMENTS

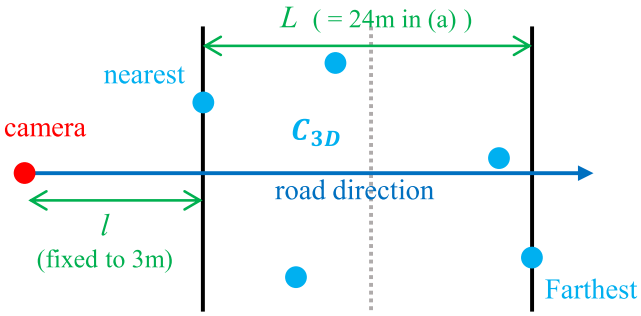
1) DATASET

To validate our application, we experimented using a real-world scene shown in Fig. 7. We captured six images from two dashboard cameras (denoted as A and B in Fig. 7), which have similar FoV used in the previous experiments, by changing the position of the cars equipping the cameras. The resolutions of four images were 1280 × 720 while the others were 1920 × 1080 pixels. The captured dataset mimics traffic accidents, where Scene 1 (1-1 and 1-2) and Scene 2 (2-1 and 2-2) simulate the frontal and rear-end collisions as shown in Figure 7(c). Image IDs (*e.g.*, 1-1:A) correspond to the scene name and camera ID.

To acquire the ground truth of the prediction, a 3D point cloud of the scene was obtained by a laser scanner (Z+F IMAGER 5010C, Zoller+Fröhlich, Wangen im Allgäu, Germany). Since there is no access to the accurate ground truth for camera pose, we indirectly evaluated the camera localization accuracy E_{pos} via the amount of the car movement, which was measured as the difference of car positions between two



(a) the prior knowledge on 1-1:A



(b) measurements to determine expective depths

FIGURE 8. Determining the variables for the real-world scenario.

frames (*e.g.*, 1-1 and 1-2) captured by the laser scanner. Figure 7(c) shows the measured values of vehicle movements. We also evaluated the re-projection error on manually assigned 2D-3D correspondences $Re_{\mathcal{C}_{2D}}$, where the number of the corresponding points varies from 17 to 35 depending on the scenes.

2) VARIABLES

Since the real-world scene was mostly composed of near-planar surfaces, we use the *urban-like* dataset for the prediction. To determine the bounding box $\mathcal{B}$, we roughly set an area containing discriminative points in the given image. Instead of generating a large number of bounding boxes, we use the designated bounding box $\mathcal{B}$ for the accuracy prediction. To the correspondence density $\rho(\mathcal{C}_{2D})$, we count the discriminative points on given images. To the noise levels, $n(\mathcal{C}_{2D})$ and $n(\mathcal{C}_{3D})$, we assumed that the error in correspondence assignment by the human annotator was a few pixels/centimeters.

We approximate the variables regarding the scene geometry, according to the prior knowledge that is easy to measure as shown as Fig. 8. We set w and h as actual road width and wall height. To the representative depths, $\mu(\mathbf{d}(\mathcal{C}_{3D}))$ and $r(\mathbf{d}(\mathcal{C}_{3D}))$, we determine them from two measures, the distance to the nearest point l and the distance between the nearest and farthest point L as

$$\mu(\mathbf{d}(\mathcal{C}_{3D})) \sim l + \frac{1}{2}L, \quad r(\mathbf{d}(\mathcal{C}_{3D})) \sim \frac{L}{2l+L}.$$

We fix $l = 3$ [m] in our experiment since it is usually realistic to yield discriminative points around the front end of the vehicle. The farthest point depends on the target scene, we thus select a reasonable point (*i.e.*, semantically discriminative and easy to measure its 3D location by the laser scanner) from a given image. We empirically confirmed that the approximation was reasonable since the actual representative depth calculated from the well-estimated optical center differs from these approximations by only a few meters at most.

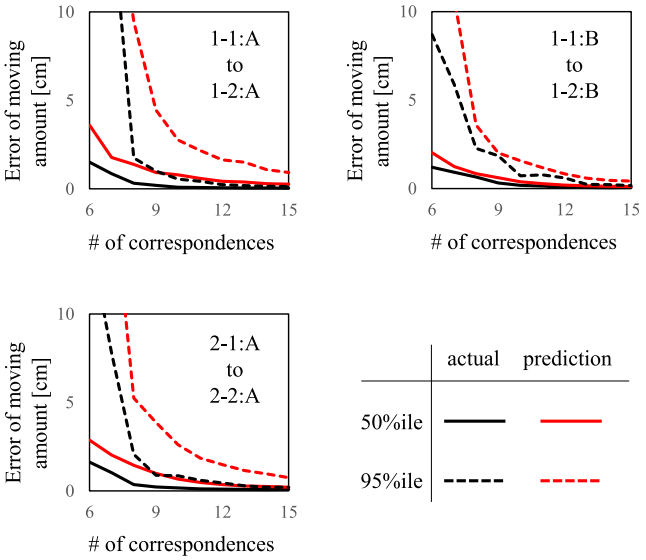


FIGURE 9. Comparison of error distributions between the (subsampled) real scenes and predicted ones.

3) RESULT

Table 2 summarizes the prediction by our application as well as the ground-truth calibration errors. The last set of columns indicates the percentile values of the actual errors across 1, 000 synthetic samples generated by the application (a smaller percentile means a smaller error). In most cases, the real-world results were in the confidence interval predicted by the application.

4) COMPARISON OF ERROR DISTRIBUTION

To confirm if the confidence intervals yielded by our application fit the real-world scenario, we compute the distribution of the positional accuracy of actual 2D-3D correspondences using subsampled correspondence sets.

We evaluate the amount of car movement in the real-world images, which originally have more than 15 correspondences, along with the ground-truth car movement observed by the 3D scanner. We randomly subsample the corresponding points from 6 to 15 points (1000 trials for each) and estimate the car movement based on the calibrated camera parameters using the subsampled correspondences. Since each trial yields a single estimation, we can get the error distribution over the subsampled correspondence sets. We use the *urban-like* scene for the error prediction. Figure 9 shows the comparisons between the actual (*i.e.*, subsampled)

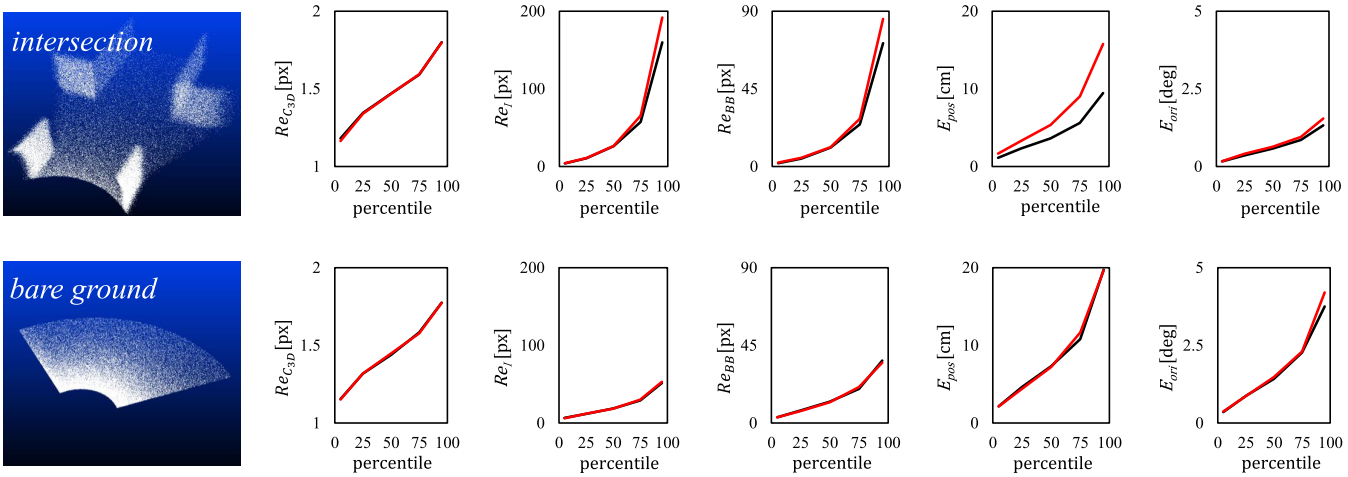


FIGURE 10. Comparisons of calibration accuracy for different geometry complexity. Percentiles of each evaluation metric from 1000 trials are visualized. Black lines are results from *intersection* and *bare ground* scenes. Red lines are predicted values using the *urban-like* geometry.

and predicted error distribution. Our application reasonably reproduces the real-world error distribution.

VI. DISCUSSION

We have introduced an evaluation of single-image camera calibration and a new application that predicts the calibration accuracy for scene analysis with TAR. The accuracy of re-projection errors and camera localization in a practical real-world scene was reasonably estimated via the error simulation using synthetic environments. Since our application does not rely on the 3D shape of target scenes, it can be used for primary screening of images as possible evidence of the incident to minimize the burden of investigators (*i.e.*, 3D acquisition, and correspondence assignments) and society (*e.g.*, road traffic restrictions).

A. GENERALIZATION ABILITY

While we use shape templates mimicking practical scenarios, *i.e.*, *urban-like* scenes, it is worth discussing the generalization ability of the error prediction using our model. To assess the generalization ability of our predefined shape templates, we conduct an experiment on different types of synthetic scene geometry that mimic road intersections and bare grounds. For the prediction of calibration accuracies, we use *urban-like* scene approximations. Specifically, we set the width and height parameters that are the same as the main road ($w = 10$ [m], $h = 5$ [m] at the *intersection*. For the *bare ground* scene, we set large road width ($w = 1000$ [m]) and zero height.

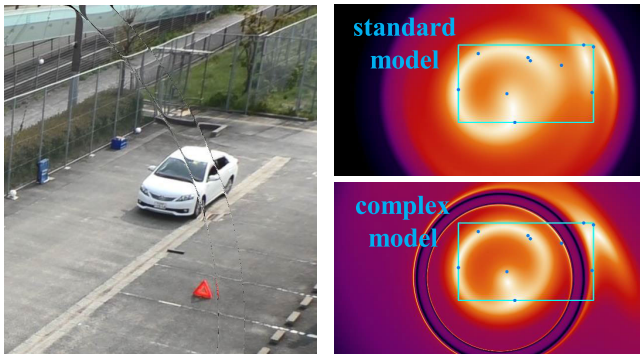
Figure 10 compares the calibration errors. The calibration accuracy for the *intersection* scene is slightly better compared to the *urban-like* scene due to the higher degree of freedom for the correspondence selection. Since the errors are still inside the confidence interval by the predictions using *urban-like* geometry, we consider our design choice of using the geometry mimicking road reasonable. Meanwhile, a practical future direction for better prediction is to increase the variation of shape templates for our predictor in addition to *random-3D* and *urban-like* scenes. For the *bare-ground* scene, the *urban-like* template accurately represents the observed scene geometry.

B. DISTORTION MODEL SELECTION

A number of distortion models have been proposed. Given a enough number of correspondences, it is known that using camera models with larger number of parameters can achieve the accurate camera calibration [40]. However, through the experiments, we found the complex camera models are often not suitable for our conditions where the correspondences are only sparsely obtained. Figure 11 shows the comparison between the models with different number of radial distortions, where we use three coefficients for the standard model and six for the complex model. We observe fault-like artifacts when increasing the number of distortion parameters. The visualization of re-projection errors (Fig. 11 (b)) highlights the artifacts appearing along circumferences.

C. SINGLE-IMAGE RECONSTRUCTION FOR TAR

The estimated camera parameters are essential information for TAR to estimate the position, shape, and behavior of target objects (*e.g.*, car). As shown in sec. V, the speed of cars can simply be computed from the positional difference between consecutive images. Also, it can be used for estimating how many points or the coverage of 3D scans on the target scene are needed to meet the accuracy requirement of the scene analysis. Although most 3D reconstruction methods for TAR [2], [3], [4] have used multi-view images, single-image 3D reconstruction is another promising application that broadens the availability of criminal investigation because the single-image metrology is fundamentally well studied [42]. We are keen to develop and deploy a whole



(a) Undistort with complex model

(b) Re-projection error (same correspondence)

FIGURE 11. Comparisons of different distortion model implementations for single-image calibration. The complex model using larger number of distortion coefficients lead fault-like artifacts.

framework for TAR for the actual investigation scenarios, including the error prediction process proposed in this paper.

ACKNOWLEDGMENT

This work was supported in part by JST Fusion Oriented REsearch for disruptive Science and Technology (FOREST) Grant Number JPMJFR206F and JSPS Grants-in-Aid for Scientific Research (KAKENHI) Grant Number JP21H03466.

REFERENCES

[1] L. B. Fricke, "Minimum training criteria for police traffic accident reconstructioninsts," Nat. Highway Traffic Saf. Admin. (NHTSA), Washington, DC, USA, Final Rep., 1987. [Online]. Available: https://rosap.ntl.bts.gov/view/dot/29602
[2] C. Fraser, "Accident reconstruction via digital close-range photogrammetry," in *Proc. Amer. Soc. Photogramm. Remote Sens. Annu. Conf. (ASPRS)*, 2006, pp. 1–7.
[3] X. Liao, Z. Zhu, Y. Yan, and T. Lv, "Traffic accident reconstruction technology research and simulation realization," in *Proc. IEEE Symp. Electr. Electron. Eng. (EEESYM)*, Jun. 2012, pp. 152–155.
[4] M. R. Osman and K. N. Tahar, "3D accident reconstruction using low-cost imaging technique," *Adv. Eng. Softw.*, vol. 100, pp. 231–237, Oct. 2016.
[5] R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *IEEE J. Robot. Autom.*, vol. RA-3, no. 4, pp. 323–344, Aug. 1987.
[6] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.
[7] J. Deutscher, M. Isard, and J. MacCormick, "Automatic camera calibration from a single Manhattan image," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2002, pp. 175–188.
[8] L. Song, W. Wu, J. Guo, and X. Li, "Survey on camera calibration technique," in *Proc. 5th Int. Conf. Intell. Hum. Mach. Syst. Cybern.*, vol. 2, Aug. 2013, pp. 389–392.
[9] W. Sun and J. R. Cooperstock, "Requirements for camera calibration: Must accuracy come with a high price?" in *Proc. 7th IEEE Workshops Appl. Comput. Vis. (WACV/MOTION)*, vol. 1, Jan. 2005, pp. 356–361.
[10] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[11] R. Szeliski, *Computer Vision: Algorithms and Applications*. Cham, Switzerland: Springer, 2010.
[12] J.-M. Lavest, M. Viala, and M. Dhome, "Do we really need an accurate calibration pattern to achieve a reliable camera calibration?" in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 1998, pp. 158–174.
[13] J. Heikkila, "Geometric camera calibration using circular control points," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1066–1077, Oct. 2000.
[14] R. Horaud, B. Conio, O. Leboulleux, and B. Lacolle, "An analytic solution for the perspective 4-point problem," *Comput. Vis., Graph., Image Process.*, vol. 47, no. 1, pp. 33–44, Jul. 1989.
[15] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EP*n*P: An accurate O(*n*) solution to the P*n*P problem," *Int. J. Comput. Vis.*, vol. 81, no. 2, p. 155, Feb. 2009.
[16] S. Li, C. Xu, and M. Xie, "A robust O(n) solution to the perspective-n-point problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1444–1450, Jul. 2012.
[17] A. Penate-Sanchez, J. Andrade-Cetto, and F. Moreno-Noguer, "Exhaustive linearization for robust camera pose and focal length estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2387–2400, Oct. 2013.
[18] T. Taketomi, K. Okada, G. Yamamoto, J. Miyazaki, and H. Kato, "Camera pose estimation under dynamic intrinsic parameter change for augmented reality," *Comput. Graph.*, vol. 44, pp. 11–19, Nov. 2014.
[19] G. Iyer, R. K. Ram, J. K. Murthy, and K. M. Krishna, "CalibNet: Geometrically supervised extrinsic calibration using 3D spatial transformer networks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1110–1117.
[20] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building Rome in a day," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 72–79.
[21] J. L. Schonberger and J.-M. Frahm, "Structure-from-Motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4104–4113.
[22] A. Geiger, F. Moosmann, O. Car, and B. Schuster, "Automatic camera and range sensor calibration using a single shot," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 3936–3943.
[23] P. F. Sturm and S. J. Maybank, "On plane-based camera calibration: A general algorithm, singularities, applications," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 1999, pp. 432–437.
[24] I. Miyagawa, H. Arai, and H. Koike, "Simple camera calibration from a single image using five points on two orthogonal 1-D objects," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1528–1538, Jun. 2010.
[25] Y. Chen, H. Ip, Z. Huang, and G. Wang, "Full camera calibration from a single view of planar scene," in *Proc. Int. Symp. Vis. Comput. (ISVC)*. Cham, Switzerland: Springer, 2008, pp. 815–824.
[26] B. Caprile and V. Torre, "Using vanishing points for camera calibration," *Int. J. Comput. Vis.*, vol. 4, no. 2, pp. 127–139, Mar. 1990.
[27] Z. Zhang, Y. Matsushita, and Y. Ma, "Camera calibration with lens distortion from low-rank textures," in *Proc. CVPR*, Jun. 2011, pp. 2321–2328.
[28] I. Schillebeeckx and R. Pless, "Single image camera calibration with lenticular arrays for augmented reality," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3290–3298.
[29] M. Lopez, R. Mari, P. Gargallo, Y. Kuang, J. Gonzalez-Jimenez, and G. Haro, "Deep single image camera calibration with radial distortion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11817–11825.
[30] Y. Hold-Geoffroy, K. Sunkavalli, J. Eisenmann, M. Fisher, E. Gambaretto, S. Hadap, and J.-F. Lalonde, "A perceptual measure for deep single image camera calibration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2354–2363.
[31] J. Weng, P. Cohen, and M. Herniou, "Camera calibration with distortion models and accuracy evaluation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 10, pp. 965–980, Oct. 1992.
[32] J. Salvi, X. Armangué, and J. Batlle, "A comparative review of camera calibrating methods with accuracy evaluation," *Pattern Recognit.*, vol. 35, no. 7, pp. 1617–1635, Jul. 2002.
[33] W. Li, T. Gee, H. Friedrich, and P. Delmas, "A practical comparison between Zhang's and Tsai's calibration approaches," in *Proc. 29th Int. Conf. Image Vis. Comput.*, New Zealand, Nov. 2014, pp. 166–171.
[34] S. K. Kopparapu and P. Corke, "The effect of measurement noise on intrinsic camera calibration parameters," in *Proc. IEEE Int. Conf. Robot. Autom.*, vol. 2, May 1999, pp. 1281–1286.
[35] K. Schreve, "How accurate can a stereovision measurement be?" in *Proc. 15th Int. Workshop Res. Educ. Mechatronics (REM)*, Sep. 2014, pp. 1–7.
[36] A. Bódis-Szomorú, T. Dabóczi, and Z. Fazekas, "Calibration and sensitivity analysis of a stereo vision-based driver assistance system," in *Stereo Vision*, A. Bhatti, Ed. London, U.K.: IntechOpen, 2008, ch. 1, pp. 1–26.
[37] C. Strecha, W. Von Hansen, L. Van Gool, P. Fua, and U. Thoennessen, "On benchmarking camera calibration and multi-view stereo for high resolution imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[38] A.-S. Poulin-Girard, S. Thibault, and D. Laurendeau, "Influence of camera calibration conditions on the accuracy of 3D reconstruction," *Opt. Exp.*, vol. 24, no. 3, pp. 2678–2686, 2016.

[39] E. Dima, M. Sjostrom, and R. Olsson, "Assessment of multi-camera calibration algorithms for two-dimensional camera arrays relative to ground truth position and direction," in *Proc. Int. Conf. 3D Vis. (3DV)*. Jul. 2016, pp. 1–4.
[40] P. F. Martins, H. Costelha, L. C. Bento, and C. Neves, "Monocular camera calibration for autonomous driving—A comparative study," in *Proc. IEEE Int. Conf. Auto. Robot Syst. Competitions (ICARSC)*, Apr. 2020, pp. 306–311.
[41] J. J. Moré, "The Levenberg–Marquardt algorithm: Implementation and theory," in *Numerical Analysis*. Cham, Switzerland: Springer, 1978, pp. 105–116.
[42] A. Criminisi, I. Reid, and A. Zisserman, "Single view metrology," *Int. J. Comput. Vis. (IJCV)*, vol. 40, no. 2, pp. 123–148, 2000.

SUSUMU KIKKAWA received the M.S. degree in mathematics from the Faculty of Mathematics, Kyushu University, in 2008. He is currently pursuing the Doctoral degree with the Graduate School of Information Science and Technology, Osaka University.

He is also a Researcher with the Forensic Science Laboratory, Criminal Department, Osaka Prefectural Police Headquarter. His research interests include forensic engineering and computer vision.

FUMIO OKURA (Member, IEEE) received the M.S. and Ph.D. degrees in engineering from the Nara Institute of Science and Technology, in 2011 and 2014, respectively.

He was an Assistant Professor with SANKEN (The Institute of Scientific and Industrial Research), Osaka University, in 2020. He is currently an Associate Professor with the Graduate School of Information Science and Technology, Osaka University. His research interest includes the boundary domain between computer vision and computer graphics.

Dr. Okura is a member of IEICE, IPSJ, and VRSJ.

DAIGO MURAMATSU (Member, IEEE) received the B.S., M.E., and Ph.D. degrees in engineering from Waseda University, Tokyo, Japan, in 1997, 1999, and 2006, respectively.

From 2015 to 2020, he was an Associate Professor with the Institute of Scientific and Industrial Research, Osaka University. He is currently a Professor with the Department of Computer and Information Science, Faculty of Science and Technology, Seikei University. His research interests include pattern recognition and biometrics, including gait recognition.

Dr. Muramatsu is a member of IPSJ and IEICE.

YASUSHI YAGI (Senior Member, IEEE) received the Ph.D. degree from Osaka University, in 1991.

In 1985, he joined the Product Development Laboratory, Mitsubishi Electric Corporation, where he was involved in robotics and inspections. He became a Research Associate, in 1990, a Lecturer, in 1993, an Associate Professor, in 1996, and a Professor, in 2003, with Osaka University, where he was the Director of SANKEN (The Institute of Scientific and Industrial Research), from 2012 to 2015. He was the Executive Vice President of Osaka University, from 2015 to 2019. His research interests include computer vision, pattern recognition, biometrics, human sensing, medical engineering, and robotics.

Dr. Yagi is a fellow of IPSJ and a member of IEICE and RSJ. He is also a member of the Editorial Board of the *International Journal of Computer Vision*. He is the Vice President of the Asian Federation of Computer Vision Societies. He was awarded the ACM VRST2003 Honorable Mention Award, the IEEE ROBIO2006 Finalist of the T. J. Tan Best Paper in Robotics, the IEEE ICRA2008 Finalist for the Best Vision Paper, the PSIVT2010 Best Paper Award, the MIRU2008 Nagao Award, the IEEE ICCP2013 Honorable Mention Award, the MVA2013 Best Poster Award, the IWBF2014 IAPR Best Paper Award, and the *IPSJ Transactions on Computer Vision and Applications* Outstanding Paper Award (2011 and 2013). He has served as the Chair for international conferences, including ROBIO2006 (PC), ACCV (2007PC and 2009GC), PSVIT2009 (FC), and ACPR (2011PC, 2013GC, 2021GC, and 2023GC). He has also served as an Editor for the IEEE ICRA Conference Editorial Board (2008 and 2011). He was the Editor-in-Chief of the *IPSJ Transactions on Computer Vision and Applications*.

HIDEO SAITO (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Keio University, Japan, in 1992.

Since 1992, he has been with the Faculty of Science and Technology, Keio University. From 1997 to 1999, he joined the Virtualized Reality Project with the Robotics Institute, Carnegie Mellon University, as a Visiting Researcher. Since 2006, he has been a Full Professor with the Department of Information and Computer Science, Keio University. His research interests include computer vision and pattern recognition, and their applications to augmented reality, virtual reality, and human–robotic interaction.

Dr. Saito's recent activities in academic conferences, include being the Program Chair of ACCV 2014 and ISMAR 2016, the General Chair of ISMAR 2015, and the Scientific Program Chair of EuroVR2020.