

# Calibration Method for Sparse Multi-view Cameras by Bridging with a Mobile Camera

Hidehiko Shishido<sup>1</sup> and Itaru Kitahara<sup>1</sup>

<sup>1</sup>University of Tsukuba

e-mail: {shishido,kitahara}@ccs.tsukuba.ac.jp

**Abstract**— Camera calibration that estimates the projective relationship between 3D and 2D image spaces is one of the most crucial processes for such 3D image processing as 3D reconstruction and 3D tracking. A strong calibration method, which needs to place landmarks with known 3D positions, is a common technique. However, as the target space becomes large, landmark placement becomes more complicated. Although a weak-calibration method does not need known landmarks to estimate a projective transformation matrix from the correspondence information among multi-view images, the estimation precision depends on the accuracy of the correspondence. When multiple cameras are arranged sparsely, detecting sufficient corresponding points is difficult. In this research, we propose a calibration method that bridges sparse multiple cameras with mobile camera images. The mobile camera captures video images while moving among sparse multi-view cameras. The captured video resembles dense multi-view images and includes sparse multi-view images so that weak-calibration is effective. We confirmed the appropriate spacing between the images through comparative experiments of camera calibration accuracy by changing the number of bridging images and applied our proposed method to multiple capturing experiments in a large-scale space and verified its robustness.

**Keywords**—camera calibration, multi-view camera, bridging image, pose estimation, badminton image

## I. INTRODUCTION

Such 3D image processing approaches as 3D tracking and 3D reconstruction are active research topics in computer vision. 3D positional estimation in large-scale spaces is being scrutinized for various scenes [1]. For such processes, the projective relationship must be obtained between the 3D space and the 2D image space, given by the camera parameters of the capturing camera. In general camera calibration processes, we need to place landmarks with known 3D positions in the space and estimate the projective transformation matrix from the correspondence relationship between the 3D points and their observed positions in a 2D image plane. This is called strong calibration [2]. However, when it is applied to a large-scale space, the time and effort of landmark installation becomes a problem. On the other hand, the weak-calibration method (or self calibration) [3] does not require landmarks to be placed. Relative position and orientation information can be estimated among multiple cameras as well as the intrinsic parameters of the capturing cameras from the correspondence information among multiple viewpoint images. However, when the cameras are arranged sparsely, sufficient correspondence points cannot be obtained, and the estimation precision of the projection relationship falls. In such large-scale spaces as a gymnasium or a stadium where we must apply 3D image processing, densely arranging cameras is usually difficult.

As an application target of 3D image processing, we focus on badminton games in a relatively large-scale space where installing 3D tracking equipment to capture players and shuttlecocks is difficult. As shown in Fig. 1, when we capture badminton match images, the entire court region is usually captured from a distance. In many other sports, the entire court image is often captured and utilized for strategic analysis purposes. In sports analysis, the 3D position of athletes and balls is crucial as basic data for improving performance and will eventually be applied for the automatic referring and management of games in the near future. Research on the 3D position estimation of a

subject using game images is active. Since the target space is too large to install enough cameras to guarantee weak-calibration's accuracy, strong calibration is normally used. However, since estimating such 2D-3D projective relationships as placing landmarks is time-consuming, acquiring the observed position of the landmarks in captured images is often onerous. Providing 3D position data to athletes and coaches is also time-consuming. Moreover, in official international competition, placing landmarks in the target space is difficult.

In this research, we combine the advantages of the strong and weak-calibration methods by bridging them with mobile camera images. A mobile camera captures video while moving among sparse multi-view cameras (Fig. 1). The captured video resembles dense multi-view images, including sparse multi-view images so that weak-calibration effectively works.

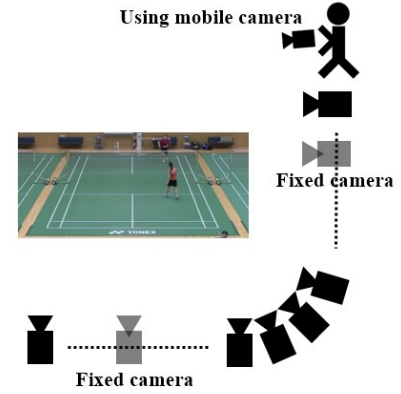


Fig. 1 Dense images captured in our proposed method. Mobile camera captures target scene while moving among sparsely fixed multi-view cameras.

## II. RELATED WORKS

Strong calibration using known landmarks (e.g., checkerboards) is one common approach [4, 5]. Davide et al. proposed a flexible new technique for single viewpoint omnidirectional camera calibration using checkerboards that calibrated a panoramic camera with a vertical field of view over 200°. Chen et al. proposed a refractive calibration method for an underwater stereo camera system where both cameras are looking through multiple parallel flat refractive interfaces. In research on improving calibration accuracy, the estimation error was minimized by calculating the epipolar geometry from dynamic silhouettes [6] and color codes [7]. The motion barcode of Gil et al. is a binary temporal sequence for lines that indicate the existence of at least one foreground pixel on that line. The search for corresponding epipolar lines was limited to lines with similar barcodes. Schillebeeckx et al. introduced a calibration object based on a flat lenticular array that creates a color-coded light-field whose observed color changes depending on the angle from which it is viewed. Other studies have also addressed environments where it is difficult to calibrate cameras, such as underwater or medical endoscopes.

Nishimura et al. [8] proposed a camera calibration algorithm for camera systems involving distortions by unknown refraction and reflection processes. Melo et al. [9] proposed a complete software-based system that calibrates and corrects radial distortion in clinical endoscopy in real time.

All of the above approaches and methods target relatively small spaces. On the other hand, in the large-scale space targeted in this research, since we must place landmarks to cover the entire space, much labor is required. To solve such problems, Workman et al. proposed [10] a camera calibration method that used the geometry of a rainbow and described the minimal sets of constraints that are sufficient for estimating camera calibration and presented both semi-automatic and fully automatic methods to calibrate a camera. However, rainbows are relatively rare, and applying them in a large indoor space is difficult. Calibration methods have been extensively studied that utilize the corresponding information between multi-viewpoint images without the installation of landmarks [11, 12, 13]. By analyzing the motion field of radially distorted images, Wu et al. [12] solved for critical surface pairs that can lead to the same motion field under different radial distortions and possibly different camera motions.

Cohen et al. described [14] an example of robust calibration by adding corresponding points and proposed a combinatorial approach for solving this variant by automatically stitching multiple sides of a building together. However, when obtaining sufficient corresponding points is difficult and when the cameras are installed sparsely, the estimation accuracy easily decreases. In weak-calibration, the relative position, the orientation information, and the camera's internal parameters are estimated from the correspondence information of multi-viewpoint images. Therefore, uncertainty exists about the scale in the relative positional relationship among cameras. As a countermeasure, we convert the coordinate system obtained by weak-calibration to a world coordinate system using information defined by each individual sport (e.g., court size).

### III. MULTI-VIEW CAMERA CALIBRATION METHOD

#### A. Acquisition of projective transformation matrix using weak-calibration

As shown in Fig 1, multi-view images are photographed by sparsely installed fixed cameras. At the same time, we captured video while moving a mobile camera among fixed cameras facing the same direction as the fixed cameras. This means that a mobile camera visually bridges sparsely arranged multi-viewpoint cameras. As a result, we acquired a group of dense multi-view images that include fixed cameras. By applying weak-calibration to the image groups, we can estimate the projective transformation matrix for all multi-view images including sparse fixed cameras without setting landmarks, since detecting sufficient corresponding points is necessary to improve the estimation accuracy. We assume that enough image features can be observed to obtain corresponding points in the captured images of the target space, where at least the size of one object is known in the captured space to estimate the scale parameters.

#### B. Calculation of 3D coordinates

A 3D coordinate of an arbitrary point in the weak-calibration coordinate system is defined as  $M_{sfm} = [X_s, Y_s, Z_s, 1]^T$ . When  $m = [u, v, 1]^T$  is observed in the camera coordinate system, the projection relation between the weak-calibration coordinate system and the camera coordinate system is expressed by Eq. (1). Projective transformation matrix  $P$  of the camera in the weak-calibration

coordinate system, acquired by the method in Section III-A, is used:

$$\lambda m \approx PM_{sfm}. \quad (1)$$

The projection relation is similarly estimated in multiple viewpoint images. The 3D coordinates are calculated from the observation coordinates on the image by the stereo method using the estimated projective transformation matrix.

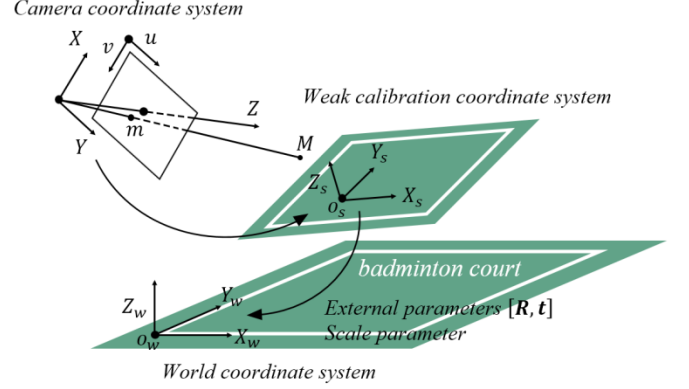


Fig. 2 Geometric relationship of camera coordinate system, weak-calibration coordinate system, and world coordinate system

#### C. Transformation from weak-calibration coordinate system to world coordinate system

In a weak-calibration coordinate system, the coordinate system is based on the distribution of the corresponding points to be observed. Therefore, the origin and the direction of each axis are different for each capturing process. For long-term 3D image processing, a unified 3D coordinate system is needed. For that purpose, we set the world coordinate system of the capturing space and transformed it from a weak-calibration coordinate system into a world coordinate system (Fig. 2).

An arbitrary point of the world coordinate system is defined as  $M_{world} = [X_w, Y_w, Z_w]^T$ . The world coordinate system's transformation from a weak-calibration coordinate system is expressed by a transformation matrix using rotation matrix  $R$  and translation vector  $t$  (Eq. (2)):

$$M_{world} = RM_{sfm} + t. \quad (2)$$

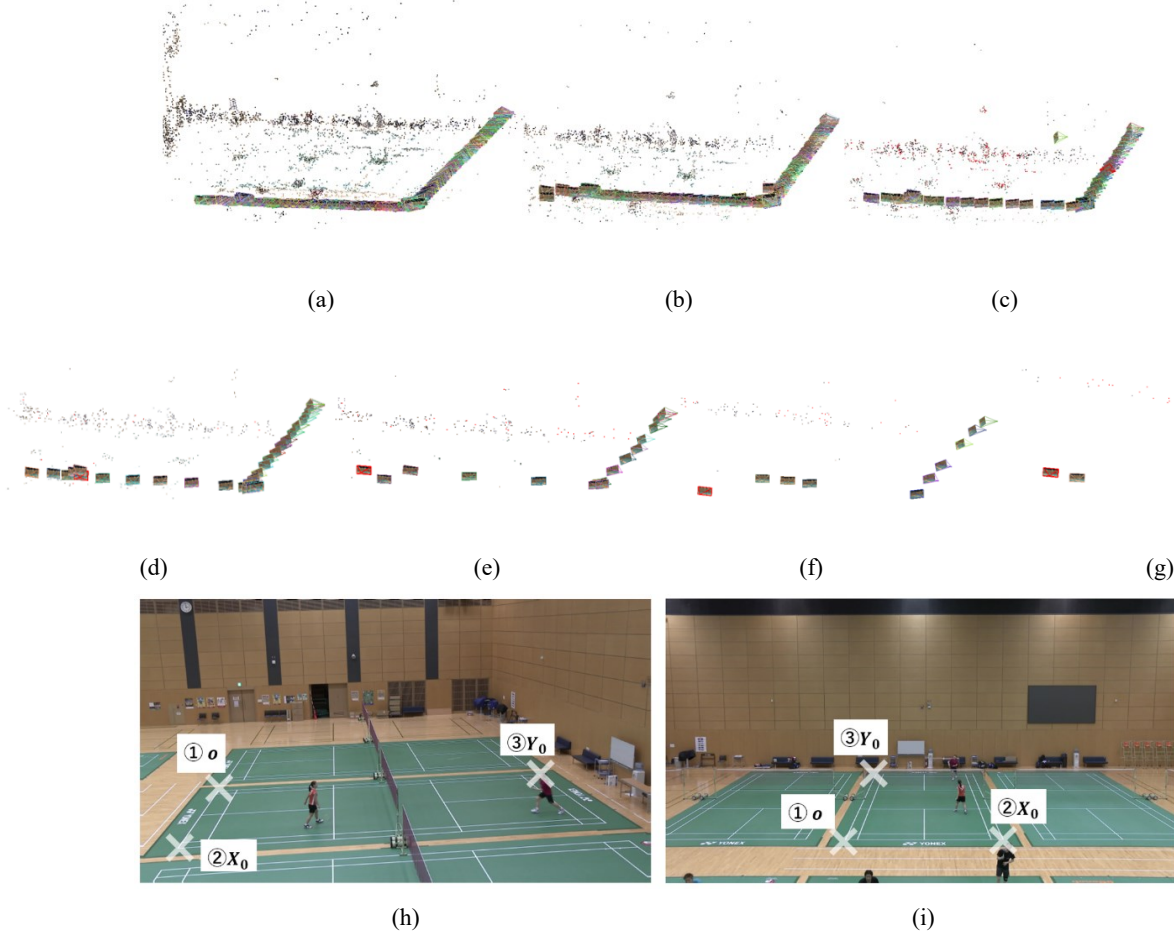
Here, 3D transformation matrix  $D$  is

$$D = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix}. \quad (3)$$

Equation 4 is expressed using transformation matrix  $D$ :

$$\tilde{M}_{world} = D\tilde{M}_{sfm}. \quad (4)$$

In this paper, we define the origin of the world coordinate system (Figs. 3(h) and (i)) to satisfy the following two conditions. First, two straight lines (edges) intersect vertically from the capturing scene of the multi-view video. Second, there is an object whose size is known.



Figs. 3 (a-g): Results of estimating camera parameters using mobile camera images shown in Fig 1. Number of bridging images: (a: 300, b: 150, c: 75, d: 40, e: 20, f: 10, and g: 5). (h and i): World coordinate system's capturing-image environment.

Vector  $\mathbf{t}$  is the parallel translation amount from point  $S_o$  to origin  $o_{sfm}$  in the weak-calibration coordinate system that corresponds to the origin of the world coordinate system. In addition, the scale is obtained from the ratio with the size in the weak-calibration coordinate system using an object whose size is known in the world coordinate system. An orthonormal vector of the weak-calibration coordinate system represented by Eq. 5 is calculated by points  $S_x, S_y$ , and  $S_z$  in the weak-calibration coordinate system that corresponds to the points on the X-, Y-, and Z-axes of the world coordinate system. Rotation matrix  $\mathbf{R}$  is obtained from the components of each vector  $e_i$ :

$$e_i = \frac{S_i - S_o}{|S_i - S_o|} \quad (i = x, y, z). \quad (5)$$

By the transformation from a weak-calibration to a world coordinate system, we calculated the subject's 3D position in the world coordinate system.

#### IV. ACCURACY EVALUATION EXPERIMENT OF MULTI-VIEW CAMERA CALIBRATION METHOD

Next we captured badminton scenes in a gymnasium. We installed two cameras so that their optical axis is orthogonal to the X- and Y-axes of the world coordinate system (Fig. 1). Figs. 3(h) and (i) show images taken by each camera. The origin is set at the corner of the court, and the X- and Y-axes are set along the court line, defined by standard badminton regulations. The distance of ① ② is 6.1 m, and

the distance of ① ③ is 13.4 m (Figs. 3(h) and (i)). The scale parameters are estimated based on these distances.

Multi-videos were captured using digital video cameras (Sony FDR AX-1,) with 3,840 x 2,160 pixel resolution at 30 frames/second. We also captured video of the same space by moving (bridging) between two fixed cameras using a camera with identical specifications. A bridging image is obtained by dividing the video into frames. In this experiment, we moved the mobile camera (Fig. 1) along the gymnasium's layout.

To evaluate the relationship between the accuracy of the camera calibration and each frame's intervals (i.e., bridging gap), we adjusted the gap to 1.0°, 1.5°, 2.5°, 6°, 12°, 21° and 26°. As shown in Figs. 3(a)-(g), weak-calibration processing is applied to the captured bridging image. Using the estimated camera parameters, ① origin:  $o_{world}$ , ②  $X_o$ , and ③  $Y_o$  of the world coordinate system are calculated (Figs. 3(h) and (i)). We verified the estimation accuracy of the 3D position.

As shown in Fig. 4, to evaluate the accuracy of the camera calibration, strong calibration processing was done with the known badminton court coordinates. The court's line is defined based on badminton regulations and shown in Fig. 4. The court coordinates of the world coordinate system and the pole tip position (Nos. 1-18 in Fig. 4) are calculated based on the origin's position defined in Fig. 3. Similarly, we acquired the specified position coordinates of each image. Strong calibration was done using this information. Thus, with

the calculated camera parameters, ①the origin of the world coordinate system:  $o_{world}$ , ②  $X_o$ , and ③  $Y_o$ , (Figs. 3(h) and (i)) were calculated and the estimation accuracy of the 3D position was verified.

### A. Estimation error vs. number of bridging images

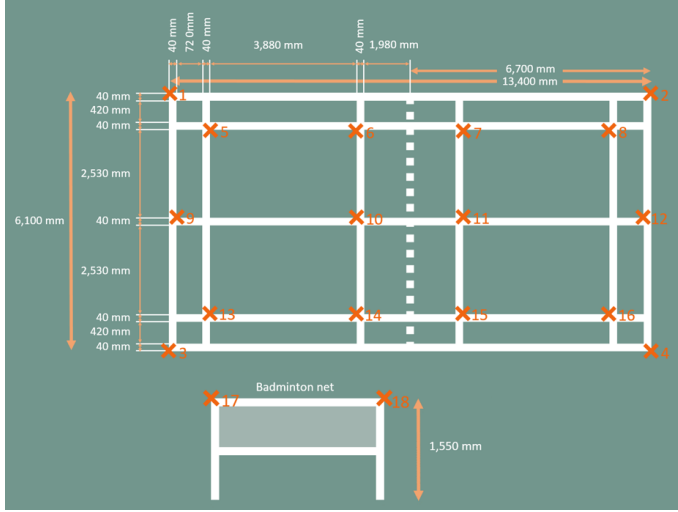


Fig. 4 Court coordinates of world coordinate system and pole tip position (Nos. 1-18)

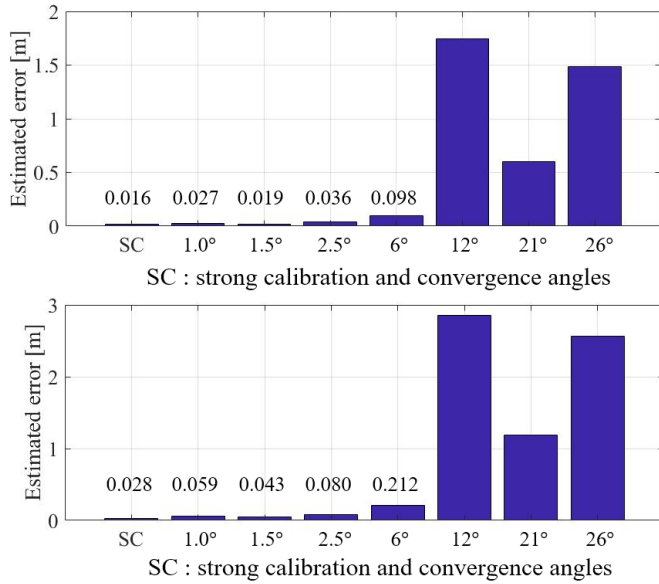


Fig. 5 Calculation error of Euclidean distance (upper:  $X_o$  and ground truth, lower:  $Y_o$  and ground truth)

We compared the values defined in the world coordinate system (① origin:  $o_{world}$ , ②  $X_o$ , and ③  $Y_o$ ) with those defined in badminton regulations (①② 6.1 m, ① ② ③ using 13.4 m). The calculation error of the Euclidean distance ( $X_o$  and ground truth /  $Y_o$  and ground truth) is shown in Fig. 5. The average estimated error due to strong calibration using the badminton court coordinates was 2.2 cm. The minimum average error value was 4.3 cm when using the 1.0° convergence angle between each bridging image. The maximum average error value was 229.5 cm when using the 12° convergence angle between each bridging image. As a result, 2.1 cm is the difference between the average estimation error by the proposed

method using the 1.0° convergence angle between each bridging image and the average estimation error by strong calibration. The proposed method's precision was almost the same as the strong calibration.

In estimation processing using a small number of bridging images, the error in the world coordinate system is large. When the 1.0°, 1.5° and 2.5° intervals of each frame are used, we can estimate the 3D position within 10-cm estimation error. When the number of intervals of each frame exceeds 12°, the error sharply increases.

Figure 5 shows a large difference in the average error between 40 bridging images and 20 bridging images. When we used 40 bridging images, the convergence angle for each camera pair is approximately 6°. When we used 20 bridging images, it is about 12°. As a result, the complement images should be divided so that the convergence angle is less than 6° to increase our proposed method's effectiveness.

In this experimental environment, the distance between the cameras in Fig. 1 is approximately 40 m (20 m long, 20 m wide). In this case, if a bridging image is captured by walking at 1 m / sec and divided by one frame per second, the convergence angle is about 6°.

### B. 3D pose estimation using proposed method

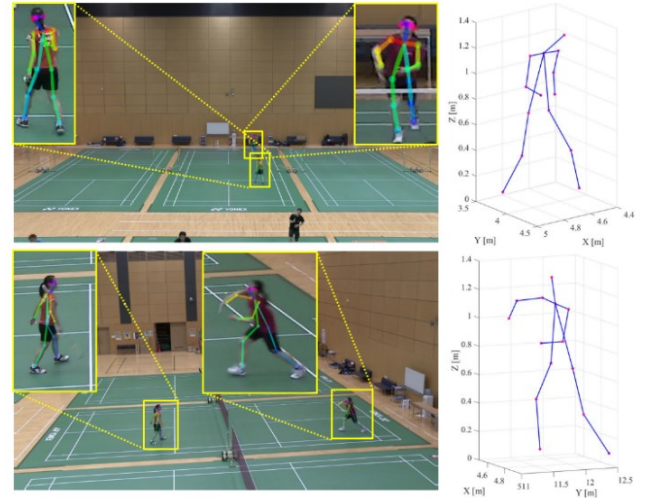


Fig. 6 Detected 2D pose (right) and estimated 3D pose (left)

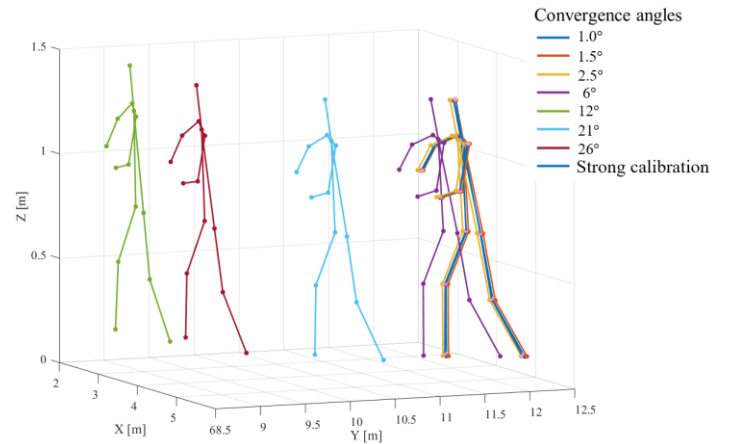


Fig. 7 3D pose estimated by our proposed method with different convergence angles between each bridging image



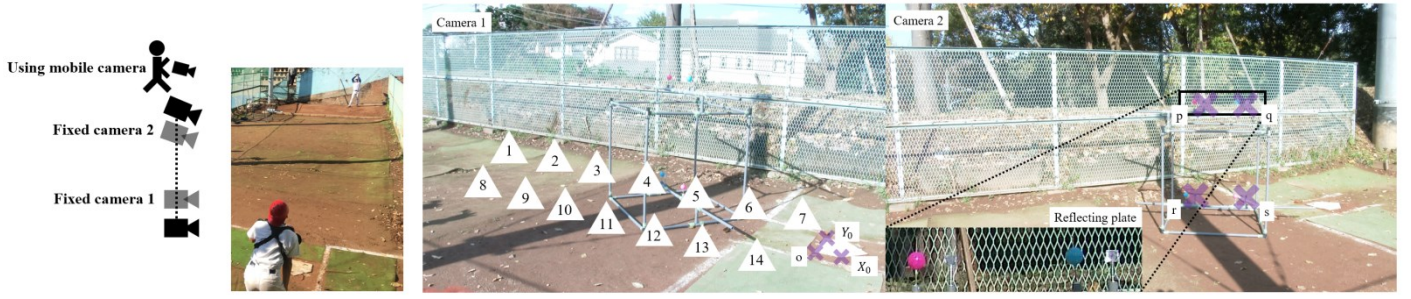


Fig. 8 Left: capturing environment and camera layout. Center: landmark installation location. World coordinate system of capturing environment (home plate). Right: landmarks with four reflectors (positions p, q, r, s) mounted on a 1-m square cube.

Since the 3D pose estimation of badminton players is a promising application of our method, we examined how a subject's pose estimation accuracy changes with the proposed method. To estimate a subject's pose in the captured image, we applied the pose estimation method of the Convolutional Neural Network (CNN) [15]. Fig. 6 shows the result of applying Convolutional Pose Machines [15] to the captured multi-view images. The estimation result of the 3D pose position from the pose information detected at two viewpoints is shown in Fig. 6 on the right. In this case, the projective transformation matrix for stereo processing was estimated by the proposed method. We confirmed that the 3D pose position is well estimated.

The result of the 3D pose position estimation with strong calibration is shown in Fig. 7. The result using a convergence angle between each bridging image is shown in Fig. 7, too. We compared the 3D pose position results estimated by the  $1.0^\circ$  convergence angle between each bridging image and strong calibration (comparison of positions under the neck) and confirmed that the average error value was 2.2 cm. When the convergence angle between each bridging image exceeded  $12^\circ$ , the error drastically increased. Thus, the proposed method's accuracy is almost identical as strong calibration. Based on this result, since the 3D position used during strong calibration includes the height of the net, we obtained estimation accuracy that was equivalent to the proposed method by including height information that matched the height of the athletes. Therefore, when estimating a position that exceeds the net's height, such as the position of the badminton shuttlecock, the estimation error of the strong calibration used in this experiment will probably increase. In the proposed method, we can estimate positions with high accuracy even in a space that is higher than the net by calibrating the whole imaging space.

### C. Application for baseball pitching

We expanded the applications of our proposed method by also experimenting with a baseball scene.

As in the above accuracy evaluation experiment, we also used badminton court coordinates for strong calibration. However, in a baseball context, extracting similar information from the capturing environment is difficult. Therefore, as shown on the right in Fig. 8, we used a landmark with four reflectors mounted to a 1-m square cube. The four positions (p, q, r, s) (Fig. 8, right) are measured by laser-surveying instrument (Total Station). We also measured the distance of home plate's apex. As shown in the center of Fig. 8, its apex is the origin, and the X and Y axes are defined. Furthermore, landmarks were placed in the whole imaging space. As shown in Fig. 8, center, a 1-m square cube with four reflectors was moved 14 places. We measured the position of each reflector with a surveying instrument for strong calibration.

The arrangement of multiple cameras is shown in Fig. 8 on the left. Set up the installed camera in the same way as in the IV section. To evaluate their calibration accuracy, we adjusted the division intervals of the frames from the moving captured images. We set  $0.8^\circ$ ,  $1.5^\circ$ ,  $3.5^\circ$ , and  $4.5^\circ$  convergence angles between each bridging image. Fig. 9 shows an example in which the weak-calibration process is applied to the captured bridging image (70 bridging images, convergence angle,  $0.8^\circ$ ). The scale parameter can be estimated based on the regulation size of home plate. With the estimated camera parameters, we verified the estimation accuracy of the 3D position of the 56 positions of the reflection plate (four reflection plates of the landmark  $\times$  14 landmark positions) (Fig. 8, center). Verification of this estimation accuracy covers both weak-calibration by the proposed method and strong calibration.

Figure 10 shows the estimation values of 56 locations in the world coordinate system (strong and weak-calibrations by the proposed method) and the calculation error of the Euclidean distance of the 3D position of the 56 places measured by the surveying instrument. The average estimation error by strong calibration was 1.5 cm, and the average estimation error by the proposed method using an  $0.8^\circ$  convergence angle between each bridging image was 4.9 cm,  $1.5^\circ$ : 5.1 cm,  $3.5^\circ$ : 5.3 cm, and  $4.5^\circ$ : 6.4 cm. In the proposed method, the estimation error in the world coordinate system increased as the convergence angle between the bridging images increased. However, even if the convergence angle increased from  $0.8^\circ$  to  $4.5^\circ$ , the estimation error remained at about 1.5 cm because the distance between the two cameras is small (about 5 m). Therefore, for capturing environments where the distance between sparsely installed cameras like badminton court is large, the proposed method's effectiveness is confirmed.

Based on the above experiment results, the effectiveness of our proposed method is shown in Table 1. No labor-intensive or low accuracy methods were compared. However, labor is intensive and accuracy is low when no landmarks were installed in the whole capturing space during strong calibration. In this experiment, when both labor and estimation accuracy are low, a strong calibration method just uses the badminton court coordinates. The 3D position for strong calibration includes the court's net height information. We obtained the same estimation accuracy as the proposed method that includes the same height information as the athletes. When estimating a position that is higher than the net, the estimation error of the strong calibration is large. When both the labor and estimation accuracy are high, strong calibration is applied to the baseball situation in this experiment. We calibrated 56 measuring points in the world coordinate system with a surveying instrument and also manually obtained 56 positions on two observation images to calculate the camera parameters. This step was very time-consuming. On the other hand, the estimation accuracy averaged approximately 1.5 cm, confirming highly accurate calibration. Finally, we confirmed that our proposed method provides high estimation accuracy with less labor.

There is no need for strong calibration labor, as was applied to baseball. Since we only calculated the camera parameters by capturing the bridging image, the labor is small. When applied to this experiment's badminton game, the proposed method's estimation accuracy nearly matched strong calibration. Therefore, when the distance between the sparsely installed cameras like badminton court is large, the proposed method works with high accuracy and less labor.

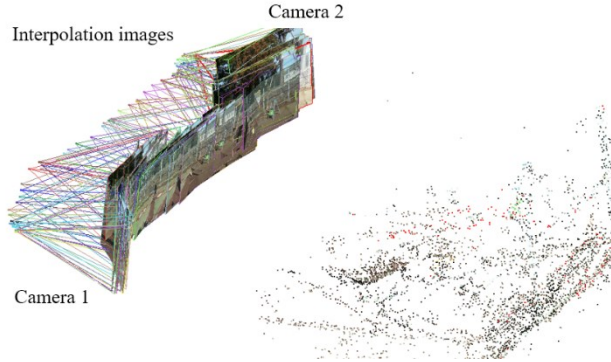


Fig. 9 Example where weak-calibration process is applied to captured bridging image (70 bridging images: convergence angle,  $0.8^\circ$ )

Table 1 Accuracy and labor of each estimation method

	Low accuracy	High accuracy
Low labor	Strong calibration (using only badminton court coordinates)	Proposed method
High labor	×	Strong calibration

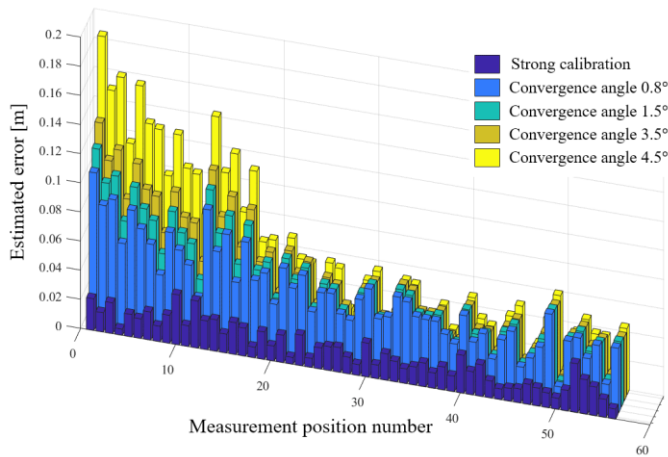


Fig. 10 Estimation values of 56 locations in world coordinate system (both strong and weak-calibration by proposed method) and calculation error of Euclidean distance of 3D position of 56 places measured by surveying instrument

## V. CONCLUSION

We introduced a method that achieves a calibration method with sparse multiple cameras by bridging them with mobile camera images. Mobile cameras captured video images while moving among sparse multi-view cameras. Such captured video resembles dense multi-view images, including sparse multi-view images so that weak-calibration

was effective. We conducted experiments on camera calibration accuracy by changing the convergence angles between each bridging image and verified our proposed method's effectiveness. When the distance between the sparsely installed cameras was increased, the proposed method worked with high accuracy and less labor.

## ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Numbers 17K13180 and JST CREST Grant Number JPMJCR16E3 including AIP challenge program, Japan.

## REFERENCES

- [1] Yuanlu Xu, Xiaobai Liu, Yang Liu and Song-Chun Zhu, "Multi-View People Tracking via Hierarchical Trajectory Composition," The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4256-4265, 2016.
- [2] Hidehiko Shishido, Yoshinari Kameda, Yuichi Ohta, and Itaru Kitahara, "Visual Tracking Method of a Quick and Anomalously Moving Badminton Shuttlecock," ITE Transactions on Media Technology and Applications (MTA), Vol. 5, No. 3, pp. 110-120, 2017.
- [3] K. Kanatani, N. Ohta and Y. Shimizu, "3-D reconstruction from uncalibrated-camera optical flow and its reliability evaluation," Systems and Computers in Japan, Vol. 33, No. 9, pp. 1-10, 2002.
- [4] D. Scaramuzza, A. Martinelli and R. Siegwart, "A Flexible Technique for Accurate Omnidirectional Camera Calibration and Structure from Motion," Fourth IEEE International Conference on Computer Vision Systems (ICVS'06), pp. 45-45, 2006.
- [5] X. Chen and Y. H. Yang, "Two-View Camera Housing Parameters Calibration for Multi-layer Flat Refractive Interface," The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 524-531, 2014.
- [6] Gil Ben-Artzi, Yoni Kasten, Shmuel Peleg and Michael Werman, "Camera Calibration From Dynamic Silhouettes Using Motion Barcodes," The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4095-4103, 2016.
- [7] I. Schillebeeckx and R. Pless, "Single Image Camera Calibration with Lenticular Arrays for Augmented Reality," The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3290-3298, 2016.
- [8] Mai Nishimura, Shohei Nobuhara, Takashi Matsuyama, Shinya Shimizu and Kensaku Fujii, "A Linear Generalized Camera Calibration From Three Intersecting Reference Planes," The IEEE International Conference on Computer Vision (ICCV), pp. 2354-2362, 2015.
- [9] R. Melo, J. P. Barreto and G. Falcao, "A New Solution for Camera Calibration and Real-Time Image Distortion Correction in Medical Endoscopy-Initial Technical Evaluation," IEEE Transactions on Biomedical Engineering, vol. 59, no. 3, pp. 634-644, 2012.
- [10] Scott Workman, Radu Paul Mihail and Nathan Jacobs, "A Pot of Gold: Rainbows as a Calibration Cue," European Conference on Computer Vision (ECCV), pp. 820-835, 2014.
- [11] C. Wu, "Towards Linear-Time Incremental Structure from Motion," 2013 International Conference on 3D Vision - 3DV 2013, pp. 127-134, 2013.
- [12] C. Wu, "Critical Configurations for Radial Distortion Self-Calibration," The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 25-32, 2014.
- [13] Kyle Wilson and Noah Snavely, "Network Principles for SfM: Disambiguating Repeated Structures with Local Context," The IEEE International Conference on Computer Vision (ICCV), pp. 513-520, 2013.
- [14] Andrea Cohen, Torsten Sattler and Marc Pollefeys, "Merging the Unmatchable: Stitching Visually Disconnected SfM Models," The IEEE International Conference on Computer Vision (ICCV), pp. 2129-2137, 2015.
- [15] Shih-En Wei, Varun Ramakrishna, Takeo Kanade and Yaser Sheikh, "Convolutional Pose Machines," The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724-4732, 2016.