View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Tsukuba Repository

Calibration of Multiple Sparsely Distributed Cameras Using a Mobile Camera

著者（英）	Hidehiko SHISHIDO, Itaru Kitahara
journal or publication title	Proceedings of the Institution of Mechanical Engineers. Part P, Journal of sports engineering and technology
volume	234
number	1
page range	37-48
year	2019-09
権利	(C) IMechE 2019
URL	http://hdl.handle.net/2241/00161339

doi: 10.1177/1754337119874276

Calibration of Multiple Sparsely-Distributed Cameras Using a Mobile Camera

Hidehiko Shishido, and Itaru Kitahara

University of Tsukuba, 1-1-1 Tennodai Tsukuba,Ibaraki, Japan
{shishido,kitahara}@ccs.tsukuba.ac.jp

Abstract

In sports science research, there are many topics that utilize the body motion of athletes extracted by motion capture system, since motion information is valuable data for improving an athlete's skills. However, one of the unsolved challenges in motion capture is extraction of athletes' motion information during the actual game or match, as placing markers on athletes is a challenge during game play. In this research, the authors propose a method for acquisition of motion information without attaching a marker, utilizing computer vision technology. In the proposed method, the 3D world joint position of the athlete's body can be acquired using just two cameras without any visual markers. Furthermore, the athlete's 3D joint position during game play can also be obtained without complicated preparations. Camera calibration that estimates the projective relationship between 3D world and 2D image spaces is one of the principal processes for the respective 3D image processing, such as 3D reconstruction and 3D tracking. A strong calibration method, which needs to set up landmarks with known 3D positions, is a common technique. However, as the target space expands, landmark placement becomes increasingly complicated. Although a weak calibration method does not need known landmarks, the estimation precision depends on the accuracy of the correspondence between image captures. When multiple cameras are arranged sparsely, sufficient detection of corresponding points is difficult. In this research, the authors propose a calibration method that bridges multiple sparsely-distributed cameras using mobile camera images. Appropriate spacing was confirmed between the images through comparative experiments evaluating camera calibration accuracy by changing the number of bridging images. Further, the proposed method was applied to multiple capturing experiments in a large-scale space to verify its robustness. As a relevant

example, the proposed method was applied to the 3D skeleton estimation of badminton players. Subsequently, a quantitative evaluation was conducted on camera calibration for the 3D skeleton. The reprojection error of each part of the skeletons and standard deviations were approximately 2.72 mm and 0.81 mm, respectively, confirming that the proposed method was highly accurate when applied to camera calibration. Consequently, a quantitative evaluation was conducted on the proposed calibration method and a calibration method using the coordinates of eight manual points. In conclusion, the proposed method stabilizes calibration accuracy in the vertical direction of the world coordinate system.

Keywords: Markerless pose estimation, camera calibration, multiple cameras, bridging image, badminton image

1. Introduction

In sports science research, there are many topics that utilize the body motion of athletes extracted by motion capture system, since motion information is valuable data for improving an athlete's skills. [17]. However, in order to acquire motion information, markers must be attached to the body of the athlete. These markers are prone to issues (size, mass, etc.) that hinder the athlete's mobility. In particular, sports research often investigates high-level skills, which are not possible to measure accurately when wearing markers. Sports locations are large spaces, such as fields, links, and arenas. Motion capture surrounds and digitizes the space using multiple cameras. Creating a capturable space is necessary preliminary work, and one of the more complicated tasks. Additionally, one of the unsolved challenges for motion capture is extracting the athlete's motion information during the actual game or match due to the difficulty of putting markers on athletes during game play.

In this research, the authors propose a method for acquisition of motion capture information without attaching a marker, utilizing computer vision technology. In the proposed method, it is possible to acquire the 3D world joint position of the athlete's body using just two cameras without markers. Furthermore, the athlete's 3D joint

position during game play is also possible to obtain without complicated preparations. In this way, the athlete's 3D joint position acquired by the proposed method can be expected as suitable for use with athletes. Although, the proposed method is not suitable for all types of sports analysis, it is possible to apply the method to various sports held in an environment which is similar to the experiment of the proposed method. If the measurement accuracy of the proposed method is high, analysis of body motion in general sports can be used to obtain joint angles. However, instantaneous motion analysis cannot be applied because the error is too large for measurements such as reaction velocity. If the measurement accuracy of the proposed method is low, it is not valid data for analyzing body movement. However, the data can be used as a positioning sensor by calculating the center of gravity of the estimated joint position, which can be applied to strategy and performance analysis.

3D image processing approaches, such 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 world and the 2D image space, established by the camera parameters of the capturing camera. In camera calibration processes, landmarks need to be set up with known 3D positions in the space and the projective transformation matrix from the correspondence relationship between the 3D points and their observed positions in a 2D image plane needs to be estimated. This is called strong calibration [2]. However, calibration of a large-scale space can be problematic due to the increased time and effort required for landmark installation. Alternatively, the weak calibration method (or self-calibration) [3] does not require landmark placement. 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 is reduced. In large-scale spaces, such as a gymnasium or stadium where 3D image processing is often implemented, dense camera arrangement becomes increasingly difficult.

In sports analysis, the 3D position of athletes, implements, and projectiles is basic data that is crucial for improving performance, and will eventually be applied to the

automation of game officiating and management in the near future. The 3D position estimation of a subject using game images is currently being researched. As a target application for 3D image processing, the authors focused on badminton games in a relatively large-scale space, where the installation of 3D tracking equipment for capturing players and the shuttlecock is difficult. Since the target space is too large to install enough cameras to guarantee the accuracy of weak calibration, strong calibration is typically employed. However, setting up many landmarks for an accurate strong calibration is time-consuming in such a space. Moreover, in official international competitions, setting up landmarks in the target space is even more difficult because permission must be obtained from the association to perform measurements with landmarks before the official match. Setting up the landmarks in such a large space could take half to several days, which can be expensive.

The authors have proposed a method to combine the advantages of strong and weak calibration by bridging multiple sparsely-distributed cameras with a mobile camera [4]. As illustrated in Fig. 1, the target space was captured while moving among sparse multi-view cameras using a mobile camera, so that densely captured multi-viewpoint images (interpolation images) could be acquired virtually. By utilizing the captured images, a simple and accurate calibration method is realized.

In the previous paper by Shishido and Kitahara [4], however, the experiment demonstrated is insufficient in portraying the effectiveness of the proposed method due to its limited application range. In the current paper, the authors also conduct an evaluation experiment on the calibration accuracy using manually and automatically acquired interpolation images to show the effectiveness of the proposed method. Furthermore, the proposed method was applied to the 3D skeleton estimation of a badminton player to confirm the possibility of application. Subsequently, a quantitative evaluation was conducted on camera calibration for the 3D skeleton. Moreover, a quantitative evaluation was conducted on the proposed calibration method and a calibration method using the coordinates of eight manual points.

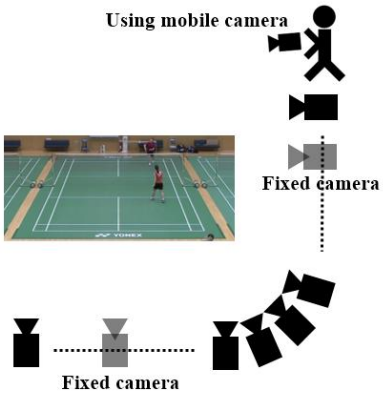


Figure 1: Dense images captured in the proposed method. Mobile camera captures the target scene while moving among sparsely-fixed multi-view cameras.

2. Related Works

Strong calibration using known landmarks (e.g., checkerboards) is one common approach for camera calibration. Scaramuzza et al. [5] proposed a flexible technique for single viewpoint omnidirectional camera calibration using checkerboards that calibrated a panoramic camera with a vertical field of view over 200 degrees. Chen et al. [6] 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 and color codes. The motion barcode of Ben-Artzi et al. [7] 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. [8] introduced a calibration object based on a flat lenticular array that creates a color-coded light field in which observed colour 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 medical endoscopes. Nishimura et al. [9] proposed a camera calibration algorithm for camera systems involving distortions by unknown refraction and reflection processes. Melo et al. [10] 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, due to the large-scale space targeted in this research, significant labour is required because landmarks needed to be set up to cover the entire space. To solve such problems, Workman et al. [11] proposed a camera calibration method that used the geometry of a rainbow to describe the minimum set of constraints that is sufficient for estimating camera calibration, and presented both semi and fully automatic methods for camera calibration. However, rainbows are relatively rare, and applying them in a large indoor space is difficult. Calibration methods that utilize the corresponding information between multi-viewpoint images without the installation of landmarks have been studied extensively [12,13,14]. By analysing the motion field of radially distorted images, Wu et al. [13] found critical surface pairs that can render the same motion field under different radial distortions and possibly different camera motions.

Cohen et al. [15] described 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 the cameras are installed sparsely, the estimation accuracy readily decreases. In weak calibration, the relative position, orientation information, and intrinsic camera parameters are estimated from the correspondence information of the multi-viewpoint images. Therefore, the scale parameter between the captured 3D space (the world coordinate system) and the reconstructed 3D space obtained by weak calibration (the camera coordinate system) is difficult to estimate. As a countermeasure, the camera coordinate system was converted to the world coordinate system using 3D information defined by each individual sport (e.g., court size).

Chen et al. [18] proposed a camera calibration method for soccer video. First, two manually corresponding 3D - 2D points are acquired, and the focal length between them is estimated. Next, the initial pan/tilt angle is estimated using one point. Finally, both points are used to minimize reprojection error. The PTZ (Pan Tilt Zoom) parameters were then optimized, and subsequently integrated and applied to the calibration of sports cameras. Obtaining corresponding points of the soccer field lines requires complicated manual work. To solve this problem, automatic calibration corresponding to 3D - 2D points was realized by detecting the soccer field lines from images [19]. In

addition, rendering football fields and athletes from YouTube video frames with a 3D viewer/AR device was accomplished by combining the automatic calibration method using soccer field lines and the depth estimation method of a player using a trained deep network [20]. These calibration methods can be conducted by corresponding the soccer field 3D - 2D points. The calibration accuracy decreases with increasing distance from the soccer field ground, but the influence on the calibration accuracy in soccer is negligible because of the many uses of information near the ground, such as player foot positions. On the other hand, the calibration accuracy in badminton is subject to greater influence because of the many uses of information far from the ground, such as player arm movement.

In this research, the authors propose a solution for accurate calibration in the vertical direction on the court, which has not been achieved in previous research.

In general, utilization of the court lines is a valuable mechanism in camera calibration of the sports scene. However, calibration accuracy proves unstable when utilizing the court lines far from the ground (high height). Therefore, a calibration is proposed that stabilizes accuracy for positions high off the ground utilizing weak calibration.

3. Multiple Camera Calibration Method

3.1 Acquisition of projective transformation matrix using weak calibration

As depicted in Fig. 1, multi-view images are captured by sparsely-installed fixed cameras. At the same time, a video sequence was captured using a mobile camera moving among and facing in the same direction as the fixed cameras. This means that a mobile camera visually bridges sparsely-arranged multi-viewpoint cameras. As a result, dense multi-view images were acquired, including images captured by the fixed cameras. By applying weak calibration to the images, the projective transformation matrix can be estimated for all multi-view images, including sparse fixed cameras, without setting landmarks, which is salient because sufficient detection of

corresponding points is necessary for improving the estimation accuracy. The authors assume that the observed image features are sufficient for obtaining corresponding points in the captured images of the target space, where the size of at least one object is known in order to estimate the scale parameters.

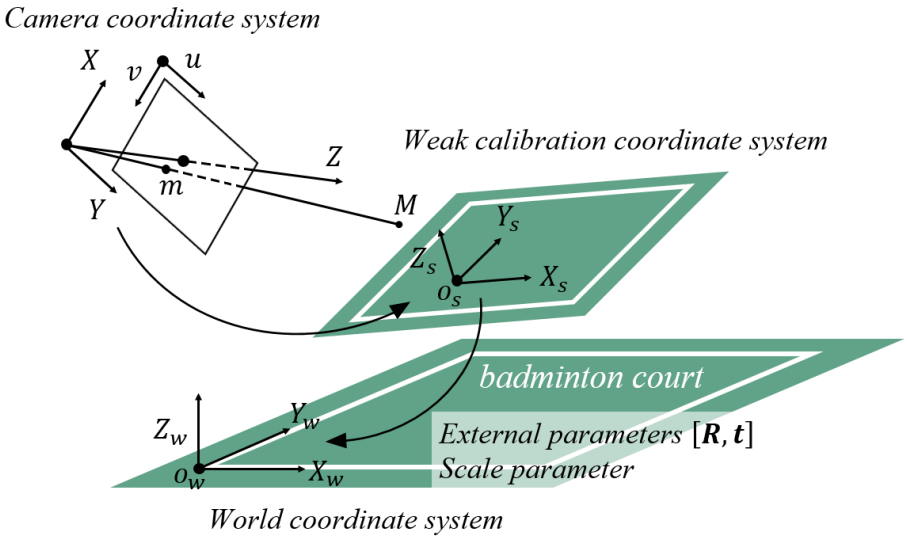


Figure 2: Geometric relationship among camera coordinate system, weak-calibration coordinate system, and world coordinate system.

3.2 Calculation of 3D coordinates

In this research, a weak calibration method was adopted that uses correspondence between multiple viewpoint images without setting landmarks [12,13]. This method is called Structure-from-Motion and is hereinafter referred to as sfm. A 3D coordinate of an arbitrary point in the weak-calibration sfm 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 projective relationship between the weak-calibration and camera coordinate system is expressed by Eq. (1). The projective transformation matrix (***P***) of the camera in the weak-calibration coordinate system, acquired by the method in Section 3.1, is employed:

$$\lambda m \simeq \boldsymbol{P} M_{sfm}. \quad (1)$$

The projective relationship is similarly estimated in multiple viewpoint images. The 3D coordinates are calculated from the observed image coordinates by the stereo vision method using the estimated projective transformation matrix. The stereo vision method is an approach for acquiring depth coordinates which are 3D coordinates from multiple images having parallax. This technique applies the principle of triangulation.

3.3 Transformation from weak-calibration coordinate system to world coordinate system

A weak-calibration coordinate system is defined by the distribution of the observed corresponding points. Therefore, the origin and direction of each axis are different for each calibration process. As shown in Fig. 2, the capturing space was originally assigned the weak-coordinate system, but it was transformed into a world coordinate system to unify measurements across different capturing data.

An arbitrary point of the world coordinate system is defined as $M_{world} = [X_w, Y_w, Z_w]^T$. The transformation from a weak-calibration coordinate system to a world coordinate system is expressed by a transformation matrix using rotation matrix $\boldsymbol{R}$ and translation vector $\boldsymbol{t}$, represented by Eq. (2):

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

Here, 3D transformation matrix D is shown in Eq. (3):

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

Equation 4 is expressed using transformation matrix $\boldsymbol{D}$:

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

As shown in Fig. 3(a) and (b), the origin of the world coordinate system was defined to satisfy the following two conditions: the vertical intersection of two straight lines (edges: $\overrightarrow{X_0} \perp \overrightarrow{Y_0}$) from the capturing scene of the multi-view video, and the existence of an object whose size is known.

S_o corresponds to the origin of the world coordinate system in the weak-calibration coordinate system. Vector $\boldsymbol{t}$ is the parallel translation magnitude from point S_o to origin o_{sfm}. In addition, the scale is obtained from the size ratio of the weak-calibration coordinate system to 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 correspond to the points on the X-, Y-, and Z-axes of the world coordinate system. Rotation matrix $\boldsymbol{R}$ is obtained from the components of each vector e_i:

$$e_i = \frac{S_i - S_0}{|S_i - S_0|} \ (i = x, y, z). \tag{5}$$

Through transformation from a weak-calibration coordinate system into a world coordinate system, the authors calculated the 3D position of the subject in the 3D world coordinate system.

4. Accuracy Evaluation Experiment of Multiple Camera Calibration Method

As illustrated in Fig. 1, two fixed cameras were installed in a gymnasium to capture badminton scenes. Fig. 3(a) and (b) are 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 the standard badminton regulations. The distance between ① and ② is 6.1 m, and the distance between ① and ③ is 13.4 m, as shown in Fig. 3(a) and (b). The scale parameters are estimated based on these distances.

The two videos are captured using digital video cameras (Sony FDR AX-1) with 3,840 x 2,160 pixel resolution at 30 frames/second. Video of the same space was also

captured by moving (bridging) between two fixed cameras using a camera with identical specifications. The bridging images are obtained through extraction of the individual frames. In this experiment, the mobile camera was moved along the gymnasium's layout, as shown in Fig. 1.

To evaluate the relationship between the accuracy of the camera calibration and each frame's intervals (i.e., bridging gap), the bridging gap was adjusted to 1.0°, 1.5°, 2.5°, 6°, 12°, 21° and 26° for 300, 150, 75, 40, 20, 10, and 5 capturing images, respectively. The interpolation image of (a) - (g) in Fig. 4 is automatically acquired using frame extraction of the moving image. Additionally, based on the position and orientation information of the interpolation image estimated by (c), the interpolation image of (h) – (k) in Fig. 4 is manually acquired with equal spacing.

As displayed in Fig. 4(a)-(k), 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 and applied in Fig. 3(a) and (b). The authors further verified the estimation accuracy of the 3D position.

As detailed in Fig. 5, strong-calibration processing was conducted with the known badminton court coordinates to evaluate the accuracy of the camera calibration. The court's lines were defined based on badminton regulations (Fig. 5). The court coordinates of the world coordinate system and the pole tip position (Nos. 1-18 in Fig. 5) were calculated based on the origin's position defined in Fig. 3. Similarly, the specified position coordinates of each image were acquired, and strong calibration was performed using this information. Thus, the camera parameters, ① the origin of the world coordinate system: o_{world}, ② X_o, and ③ Y_o, presented in Fig. 3(a) and (b), were calculated and the estimation accuracy of the 3D position was verified.

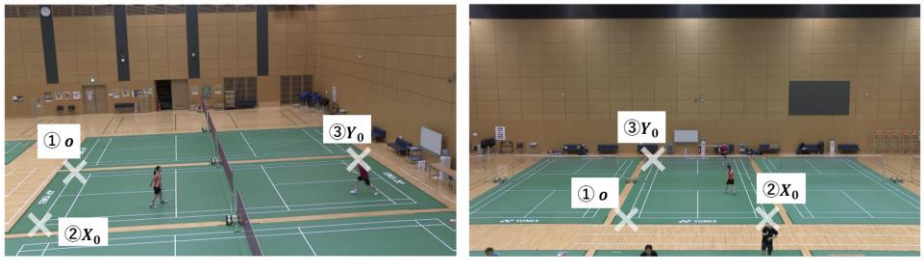


Figure 3: World coordinate system's image-capturing environment (a and b).

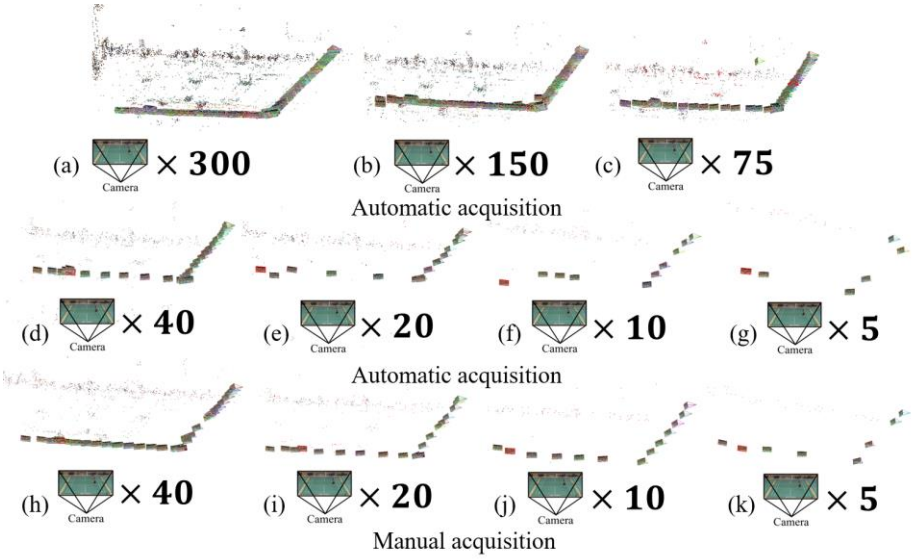
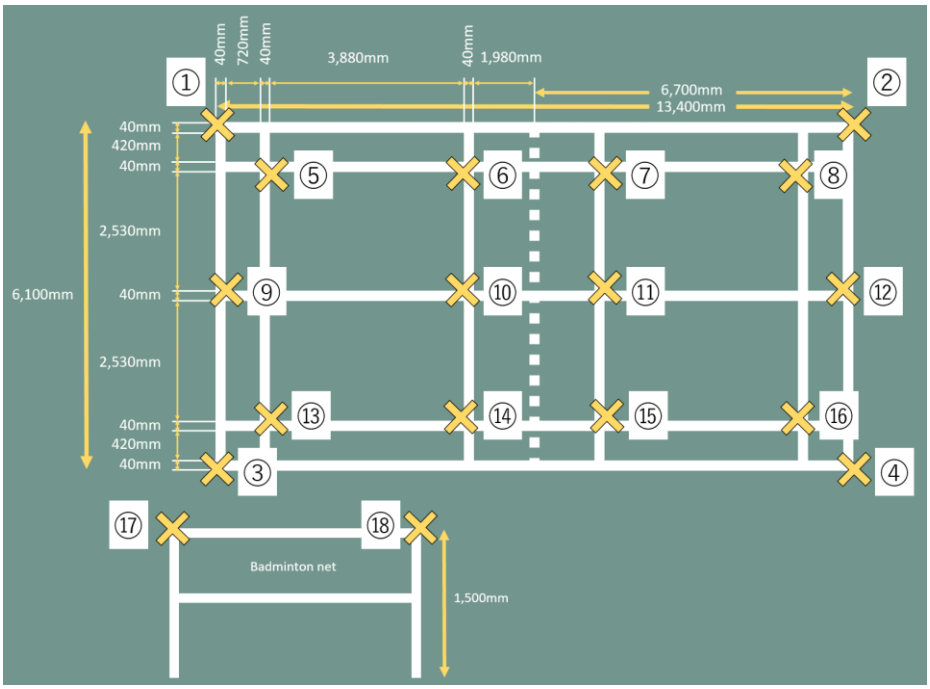


Figure 4: Results of estimating camera parameters using the mobile camera images outlined in Fig.1 (a-k). Number of bridging images for automatic acquisition (a-g) and manual acquisition (h-k): (a) 300, (b) 150, (c) 75, (d) 40, (e) 20, (f) 10, (g) 5, (h) 40, (i) 20, (j) 10, and (k) 5.



320 Figure 5: Badminton court coordinates of the world coordinate system and pole tip position (Nos. 1-18).

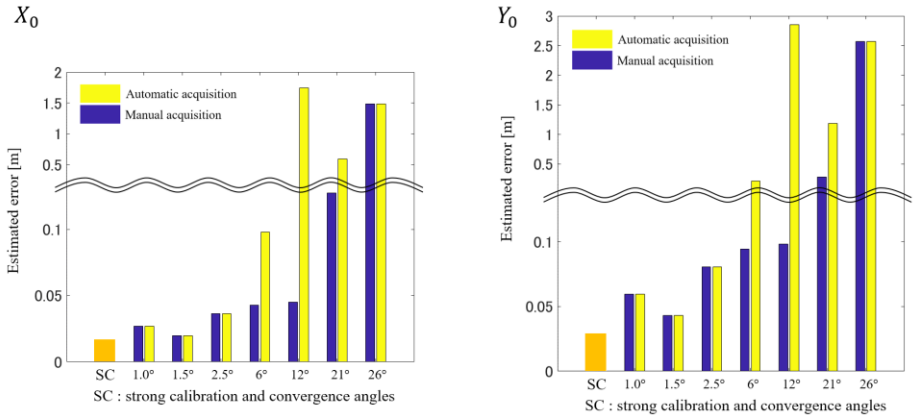


Figure 6: Calculated error of Euclidean distance by each bridging gap (Left: X_o and ground truth, Right: Y_o and ground truth)

4.1 Estimation error vs. number of bridging images

The authors compared the values defined in the world coordinate system (① o_{world}, ②X_o, and ③ Y_o) with those defined in badminton regulations (6.1 m between ① and ②, and 13.4 m between ① and ③). The calculated error of the Euclidean distance (X_o and ground truth/Y_o and ground truth) is delineated in Fig. 6. Estimated error of the strong calibration is indicated by the orange bar. Estimated error associated with manual and automatic acquisition is represented by blue and yellow bars, respectively. The average estimated error of strong calibration using the badminton court coordinates was 2.2 cm. The average error of the convergence angles 1.0°, 1.5°, and 2.5° is approximately 4.3, 3.1, and 5.8 cm, respectively. These results indicate that the proposed method has approximately the same accuracy as the strong calibration.

The estimation error from the automatic acquisition exhibits a drastic change from the convergence angle of 6° (approximate average error of 15 cm). Alternatively, the estimation error from manual acquisition exhibits a similarly abrupt change from the convergence angle of 21° (approximate average error of 20 cm).

For both manual and automatic acquisition, the estimation error monotonically increases as the convergence angle increases. For interpolation images acquired manually, error images that do not capture subjects can be excluded. Additionally, manual acquisition enables selection of only focused images. Consequently, selected images may exhibit many positive correspondences between adjacent images. Therefore, the convergence angle at which the estimation error drastically increased was wider than that of the automatic acquisition. However, manual acquisition of interpolation images requires significant labour.

As a result, to estimate the altitude and position of the fixed cameras with sufficient precision, the proposed method must sample bridging images at convergence angles less than 6 degrees. In the experiment, the total path length of the mobile camera motion was approximately 40 meters (20 m translation along the Y-axis followed by a 20 m translation along the X-axis). Accordingly, bridging images were deemed sufficient for this experiment sampled at one frame per second and captured by a moving camera at 1 meter per second.

4.2 3D skeleton estimation using the proposed method

3D pose estimation of badminton players is one of the promising applications of the proposed method. To explore this potential, an experiment was conducted to capture badminton scenes in a gymnasium. As shown in Fig. 8(a), two fixed cameras were installed parallel to the X-axis of the world coordinate system. As shown in Fig. 8(b), camera position and orientation were estimated by applying weak calibration (fixed camera 1, 2 and interpolation image). The distance between the two cameras was approximately 10 m. The positions of the origin, X-axis, and Y-axis were established as outlined in Section 3. Multi-videos were captured using digital video cameras (Blackmagic Studio Camera 4K) with 3,840 x 2,160 pixel resolution at 30 frames/second. These two cameras captured images using synchronizing signals. The authors also captured a video sequence of the same space by moving (bridging) between two fixed cameras using a camera with the same specifications as the fixed cameras. An interpolation image was acquired by sequencing the captured video into individual frames. As a result, 234 interpolation images were generated in this experiment. The badminton scene used for estimation of the 3D skeleton position totalled 16 frames (Nos. 111-127) from the start of hitting the shuttlecock to the end.

To estimate a subject's pose in the captured image, the pose estimation method of the convolutional neural network (CNN) was applied [16]. Pictured left in Fig. 7 is the result of applying convolutional pose machines [16] to the captured multi-view images. The resulting 3D pose position, estimated by the pose information detected at two viewpoints, is centred in Fig. 7. Accordingly, the projective transformation matrix for stereo processing was estimated by the proposed method.

As shown in Fig. 7 right, the trajectories of the wrist, elbow, head, and neck were estimated. The orange, green, purple, and yellow plots illustrate the trajectory of the right wrist, right elbow, head, and neck, respectively. As shown in the neck and head Z values in Fig. 7, the head estimate is never lower than the neck. Similarly, the right elbow and right wrist Z values represent parabola. This shows the swinging motion of the racket. In this way, the estimated skeleton does not reverse human structure. It was confirmed that the estimated value during the movement of the racket swing did not

lose continuity. Therefore, based on these results, the position of the skeleton from the first hit of the shuttlecock to the end was well-estimated.

Accordingly, the Euclidean distances of the right wrist and the right elbow were calculated as shown in Fig. 9 (Nos. 111-127). The average distance and standard deviation were approximately 21.5 cm and 1.9 cm, respectively. With this method, the 3D skeleton position can be calculated with less labour using the two fixed cameras to capture and produce interpolation images. The estimation data of the 3D skeleton position can contribute to improvement of an athlete's technical skills, through applications such as calculation of the skeleton's movement and corresponding data analysis.

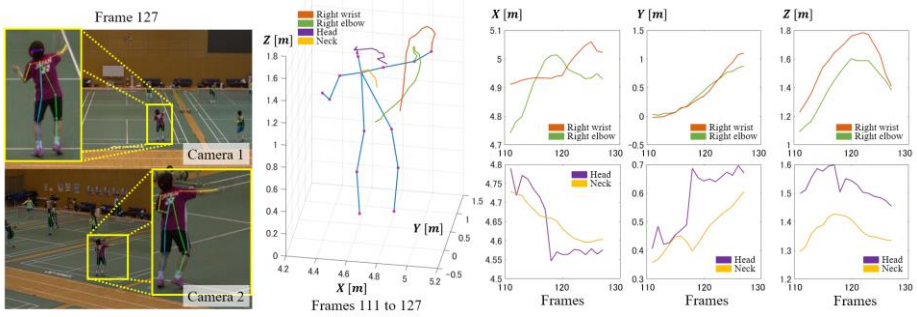
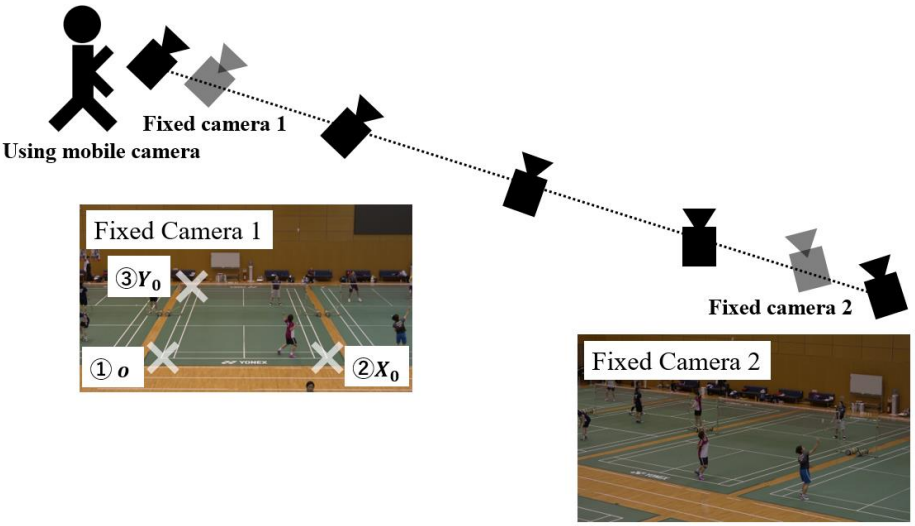
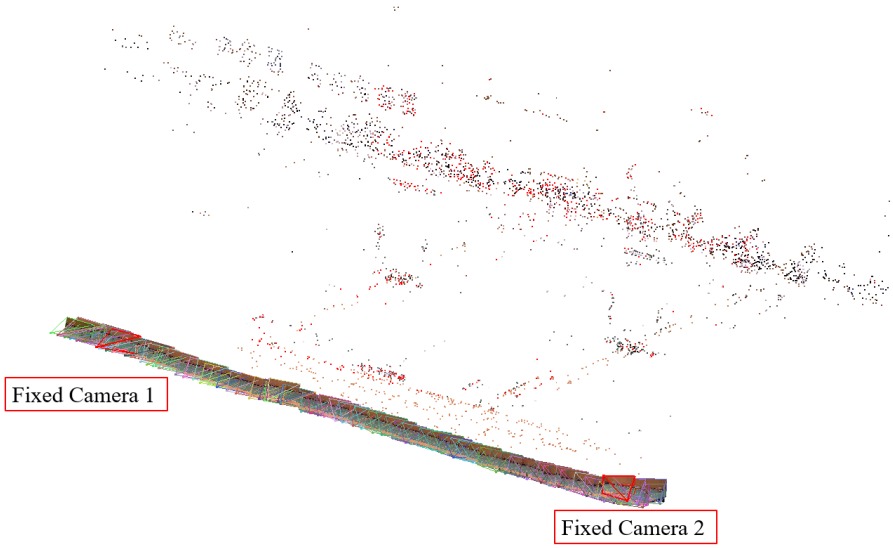


Figure 7: Left: the result of applying convolutional pose machines [16] to the captured multi-view images. Middle: The estimation result of the 3D pose position from the pose information detected at two viewpoints. Right: The estimated trajectories of the wrist, elbow, head, and neck.



(a)



(b)

Figure 8: (a) Two fixed cameras installed parallel to the X-axis of the world coordinate system. (b) Camera position and orientation estimated by applying weak calibration (fixed camera 1, 2 and interpolation image). Weak calibration estimates the position and orientation of the camera and at the same time estimates the 3D point cloud. Dots represent an estimated 3D point cloud.

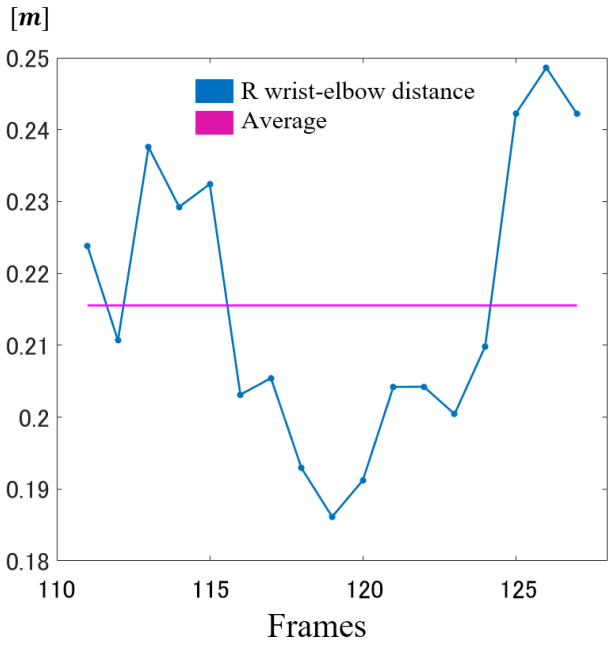


Figure 9: Euclidean distances of the right wrist and the right elbow (nos. 111 - 127).

Subsequently, a quantitative evaluation was conducted on camera calibration for the 3D skeleton. In this experiment, 3D key points were difficult to annotate, so the authors evaluated reprojection errors. The badminton scene used in this experiment yielded 17 frames (No. 111 to 127) from the first hit of the shuttlecock to the end. First, skeleton positions (2D) were acquired from the two viewpoints by convolutional pose machines [16]. Secondly, the proposed method was applied to the two skeleton positions (2D) in order to calculate the 3D skeleton position. Thirdly, the calculated 3D skeleton position was projected onto each camera image. Fourthly, the proposed method was applied again to the 2D skeleton position of the two projected viewpoints in order to calculate the reprojected 3D skeleton position. Finally, the Euclidean distance between the first and second 3D skeleton positions (calculated and reprojected, respectively) was

calculated to determine the reprojection error. The results of the calculation are reported in Table 1. The reprojection error specifies the average error of 17 frames for each part of the skeleton. As shown in Table 1, the average of the reprojection errors and standard deviations were approximately 2.72 mm and 0.81 mm, respectively. In effect, it was confirmed that the proposed method was highly accurate when applied to camera calibration.

Further, Fig. 10 presents the result of executing steps 1-3 in the reprojection error procedure described above, exemplified by frames 123 and 126. The yellow line indicates the 2D skeleton positions estimated by the convolutional pose machines [16], and the blue line represents the 2D coordinates projected onto each camera from the 3D position once estimated by the proposed method. Evidently, the lines in each frame (123 and 126) are nearly identical. However, the segment of the line from the left wrist to the left elbow does not overlap. This is due to the player's self-occlusion. Essentially, the whole body can be observed in the camera 1 image, but in the camera 2 image the left hand is hidden in front of the player's body and cannot be observed. Convolutional pose machines [16] estimate the skeleton position even if there is self-occlusion, but estimation accuracy is low for skeleton parts that cannot be observed. Therefore, the estimation precision of the 3D position by the proposed method is low in the 2D skeleton region where the estimation accuracy of convolutional pose machines [16] is low. A possible solution to this problem is positioning the camera so that self-occlusion does not occur, as opposed to the camera placement exhibited in Fig. 8.

Moreover, a quantitative evaluation was conducted on the proposed calibration method and a calibration method using the coordinates of eight manual points (hereinafter referred to as the 8 points calibration method). The positions of points 1, 2, 3, 10, 11, 16, 17, and 18, as shown in Fig. 5, were manually acquired in the 8 points calibration method. Camera parameters were calculated by corresponding the 8 points with the 3D field. As shown in the lower part of Fig. 11 (yellow plot), the data used for the quantitative evaluation were the racket positions manually acquired from the player's image (21 frames). The authors calculated the 3D racket length by applying both the proposed method and 8 points calibration method above and below the racket, acquired by the two viewpoints. The actual size of the racket was 674 mm. The estimated racket length is outlined in the upper part of Fig. 11. The 8 points calibration

and proposed methods are indicated by the blue and orange plots, respectively. Results indicate that the average racket length estimation error was nearly equivalent between the two methods: 12.69 mm and 11.86 mm for the 8 points calibration and proposed methods, respectively. Similarly commensurate, the standard deviations of the estimated racket length from 1 to 10 frames were 6.21 mm and 6.27 mm for the 8 points calibration and proposed methods, respectively. However, when comparing the standard deviations of the estimated racket length from 11 to 21 frames, the 8 points calibration and proposed methods were 12.30 mm and 8.47 mm, respectively. This result indicates that the estimation errors from 1 to 10 frames did not change between methods, but the estimation errors from 11 to 21 frames were less dispersed in the proposed method. This difference is explicated by the fact that frames 1 to 10 exhibit the racket at a height lower than the badminton court net, while in frames 11 to 21, the racket is above the net. Therefore, the accuracy of the 8 points calibration method decreases at a position higher than the height of the badminton court net. Alternatively, the proposed method provides stable calibration accuracy regardless of position relative to the net. The stabilization of the proposed method is due to uniformly distributed image features in 3D space. In conclusion, the proposed method stabilizes calibration accuracy in the vertical direction of the world coordinate system.

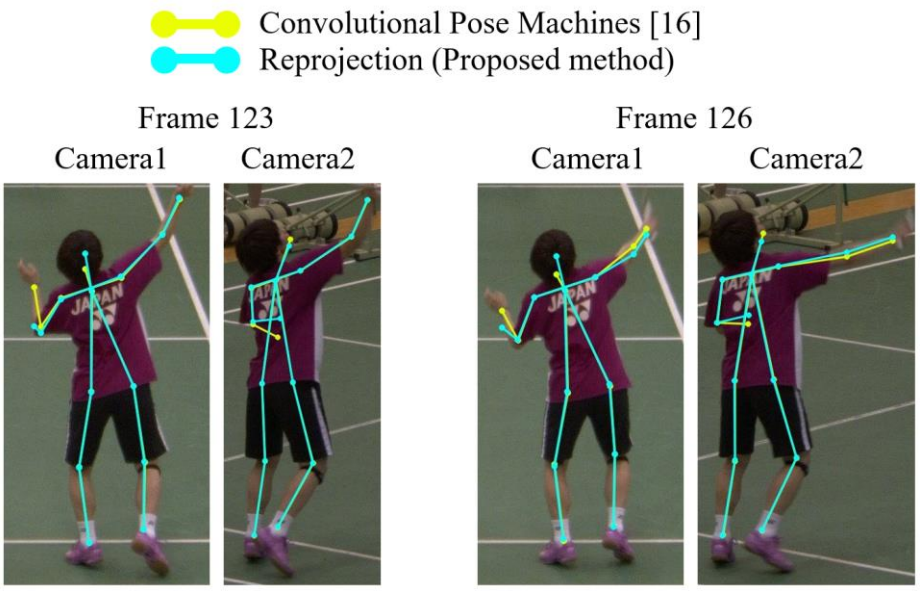


Figure 10: 2D Skeleton positions estimated from two viewpoints by convolutional pose machines [16] and by projection onto each camera from the 3D skeleton position estimated by the proposed method (frame nos. 123 and 126).

Table 1: The Euclidean distance between the calculated and reprojected 3D skeleton positions.

	Head	Neck	RShoulder	RElbow	RWrist	LShoulder	LElbow	LWrist	RHip	RKnee	RAnkle	LHip	LKnee	LAnkle	Total average
Reprojection error [mm] (Average:17 frame)	2.32	3.15	3.19	2.68	2.69	2.58	2.89	3.00	2.79	2.66	2.41	2.85	2.31	2.56	**2.72**
Standard deviation [mm]	1.17	0.84	0.58	0.94	0.86	0.93	0.76	0.85	0.82	0.81	0.62	0.79	0.67	0.74	**0.81**

Note: The reprojection error is averaged over 17 frames for each part of the skeleton.

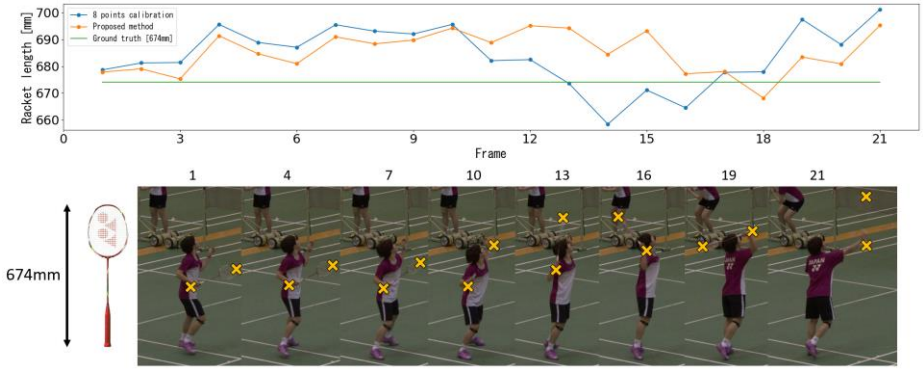


Figure 11: Quantitative evaluation of the proposed method and the 8 points calibration method. Upper: Estimation result of racket length using the proposed method and the 8 points calibration method. Lower: Racket positions manually acquired from the player's image (21 frames).

5. Conclusion

A method was introduced that achieves a calibration for multiple sparsely-distributed cameras by bridging them with mobile camera images. Experiments were also conducted to evaluate camera calibration accuracy by changing the convergence angles between each bridging image, effectively verifying the proposed method's effectiveness. When the distance between the sparsely-installed cameras increased, the proposed method performed with high accuracy and less labour.

In addition, the accuracy evaluation was executed using the interpolation images acquired manually and automatically in order to verify the effectiveness of the proposed method. Furthermore, the proposed method was applied to the 3D skeleton estimation of badminton players to confirm the possibility of the application. As a result of the study, the range of the proposed method's application expanded, demonstrating its effectiveness.

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] Kenichi Kanatani, Naoya Ohta and Yoshiyuki Shimizu, “3D reconstruction from uncalibrated-camera optical flow and its reliability evaluation," Systems and Computers in Japan, Vol. 33, No. 9, pp. 1-10, 2002.

[4] Hidehiko Shishido and Itaru Kitahara, “Calibration Method for Sparse Multi-view Cameras by Bridging with a Mobile Camera," 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 1-6, 2017.

[5] Davide Scaramuzza, Agostino Martinelli and Roland 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.

[6] Xida Chen and Yee-Hong 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.

[7] 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.

[8] Ian Schillebeeckx and Robert 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.

[9] 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.

[10] Rui Melo, João P. Barreto and Gabriel 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.

[11] 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.

[12] Changchang Wu, “Towards Linear-Time Incremental Structure from Motion," 2013 International Conference on 3D Vision - 3DV 2013, pp. 127-134, 2013.

[13] Changchang Wu, “Critical Configurations for Radial Distortion Self-Calibration," The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 25-32, 2014.

[14] 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.

[15] 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.

[16] 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.

[17] Eline Van der Kruk, Marco M. Reijne, “Accuracy of human motion capture systems for sport applications;state-of-the-art review," European Journal of Sport Science, Vol.18, No.6, pp.806–819, 2018.

[18] Jianhui Chen, Fangrui Zhu and James J. Little, “A Two-point Method for PTZ Camera Calibration in Sports,” IEEE Workshop on Applications of Computer Vision (WACV), pp. 287-295, 2018.

[19] Peter Carr, Yaser Sheikh and Iain Matthews, “Point-less calibration: Camera parameters from gradient-based alignment to edge images,” IEEE Workshop on Applications of Computer Vision (WACV), pp. 377-384, 2012.

[20] Konstantinos Rematas, Ira Kemelmacher-Shlizerman, Brian Curless and Steve Seitz, “Soccer on Your Tabletop,” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4738-4747, 2018.