

Paper

Method of Multiview Video Switching for Soccer Game Analysis in Large Scale Space

Hidehiko Shishido†, Yosuke Okada†, Yoshinari Kameda†, Masaaki Koido† and Itaru Kitahara†

Abstract Research on tactical and performance analysis utilizes videos of dynamic sports scenes. An effective multiview video switching method can support the analysis. Bullet-time video is a multiview video browsing approach. Because the image is presented almost as it is, it is suitable for high-quality observations of the subject from multiple directions. This paper proposes a multiview image switching method for understanding dynamic scenes in large-scale spaces such as soccer games. We develop a prediction model for the camerawork for shooting Bullet-time videos. The model using deep neural network, which can estimate a suitable viewpoint to observe the target scene from the position information of the soccer players, ball, and goals.

Keywords: camerawork, multiview cameras, Bullet-time video, dynamic scene, large-scale space, soccer video.

1. Introduction

With the development of computer processing capacity and the widespread use of high-quality cameras, the use of multiview videos is becoming prevalent [1]. In particular, in large spaces, such as soccer stadiums, more information can be obtained by observing multiview videos than a single-view video. Thus, research on multi-view videos that observe the subjects from various viewpoints is actively being conducted [2)-4)].

Bullet-time video is a method used to browse multiview videos. This method presents the captured image almost as it is. Because it is possible to observe the target object while maintaining the quality/resolution of the captured image, Bullet-time videos are suitable for detailed observation such as scene analysis. When the situation of target objects changes, it is necessary to dynamically set the following to obtain a Bullet-time video of a dynamic scene. (1) focus on the captured space (attention point). (2) the direction to be observed (observation viewpoint). (3) establish the size to be observed (zoom value). In the following, the attention point, observation viewpoint, and zoom value are referred to as the camerawork parameters.

In the previous Bullet-time video generation method [5], the camerawork parameters are manually set using some input device such as a mouse or keyboard. The observer must stop the video at the moment of multifaceted observation to set the camerawork parameters.

Moreover, to dynamically and appropriately setting the camerawork parameters requires the knowledge of algorithms for generating Bullet-time videos and video production. Therefore, it is too complex and difficult for the general audience. We aim to solve the problems of generating Bullet-time videos automatically.

This paper proposes an automatic multiview video switching method for dynamic scenes for large-scale spaces. As shown in the upper part of Fig. 1, under our shooting conditions, the cameras are arranged more

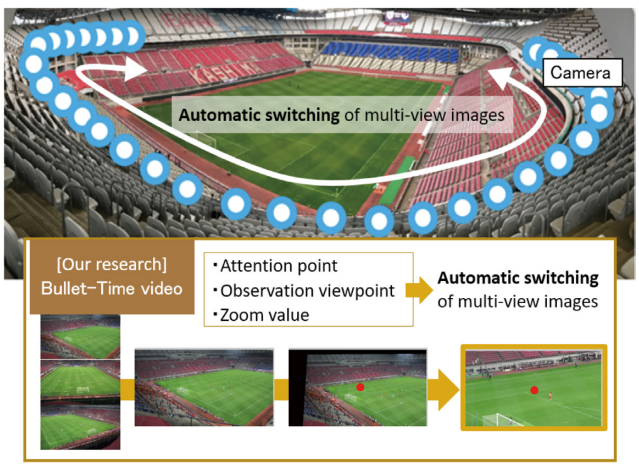

Fig. 1 Flow of our proposed method.

Received September 20, 2019; Revised December 2, 2019; Accepted December 30, 2019
†University of Tsukuba
(Ibaraki, Japan)

densely to generate high-quality Bullet-time videos that switches the observing view point smoothly. Owing to the camera layout, the camera parameters can be estimated accurately so that the system can correctly track moving objects such as soccer players and ball. As a result, even in a dynamic scene, the camerawork parameters can be obtained as shown in the lower part of Fig. 1.

2. Related Works

Bullet-time video is a multiview video presentation method, and it can reproduce the movement of the viewpoint while preventing the degradation of image quality. Akechi et al. [5] proposed a Bullet-time video generation method that can move the observation viewpoint. Bullet-time videos are generated by stopping the moment of interest (static scene) from the captured video of a large-scale soccer stadium. However, when observing scenes where the situation changes rapidly, such as in soccer, the changes cannot be understood only with still scenes. Ikeya et al. [6] have developed a multiview robot camera. The camera operator instructs all the robot cameras to focus and capture multiview images of the subject moving in the 3D space. Our research realizes the automatic generation of Bullet-time videos without camera operators.

To generate Bullet-time videos, it is necessary to select an observing camera from multiple cameras. Chen et al. [7] conducted research for automatically ranking videos used in live broadcast among multiple videos. We proposed a method to set the camera-work parameters by considering the ball position. Jiang et al. [8] tracked multiple subjects and formulated optimal viewpoint selection as a recursive decision problem. An experiment was conducted to select the optimal viewpoint from three videos of the movement of three people. We propose a viewpoint selection method that can deal with 22 soccer players as subjects. Wang et al. [9] proposed a viewpoint selection method that reflected user preferences and the number of subjects in a multiview video of a soccer game. However, the ball position must be entered manually. We propose a method to select an appropriate viewpoint from densely arranged multiple cameras using the soccer player position, ball position, and goal position.

In a research on estimating the position of soccer players, Iwase et al. [10] estimated the 3D position of a soccer player observed from multiple cameras. Kasuya et al. [11] proposed a method to stably estimate the player's position in the soccer field by using a video captured by two cameras using shadow information. Because our research aimed to generate a Bullet-time video, the number of cameras arranged to surround the target object can be considerably greater. We aim to realize robust and accurate player position estimation utilizing the rich visual information.

In research on estimating the position of the soccer ball which occupies a smaller region than humans, Ren et al. [12] proposed a soccer ball tracking method using multiple cameras. Although eight cameras were used to cover the entire field, the number of overlapping cameras was small and affects the estimation accuracy of the 3D position. Ishii et al. [13] detected a soccer ball from images captured by two fixed cameras and estimated the 3D ball position. When the player hides the ball, the Kalman filter was used to interpolate the ball position. However, if the ball cannot be sufficiently observed in the video, the prediction of the ball position fails. When generating a Bullet-time video, it is possible to observe the ball from multiview to utilize a large number of images to capture the same region. As a result, the ball can be detected in more frames even with a general object detection method.

Nagai et al. [14] have proposed a system that generates a Bullet-time video by automatic target point setting and viewpoint switching, and it achieves both real-time visual feedback and delayed visual feedback. Bullet-Time videos are effective for motion analysis; however, game video has not been verified thus far. In this research, Bullet-Time videos effective for game analysis are verified using Bullet-time video of soccer matches.

3. Method of multiview video switching for understanding dynamic scene

Fig. 2 shows the flow of the camerawork parameter estimation method for Bullet-time videos applied to dynamic scenes. At first, the multiview video is captured using multiple cameras in a soccer stadium. Then, the camera parameters of the multiview video are obtained by camera calibration. In this research, the position, and orientation of the camera and the focal length are estimated by using Structure from Motion (SfM) [15]. The position of the player and ball on the field are estimated based on the camera parameters. The camerawork parameters necessary for video switching are obtained from the observer. The relationship between the player position, ball position, goal position, and camerawork parameters generated by the observer is learned by a

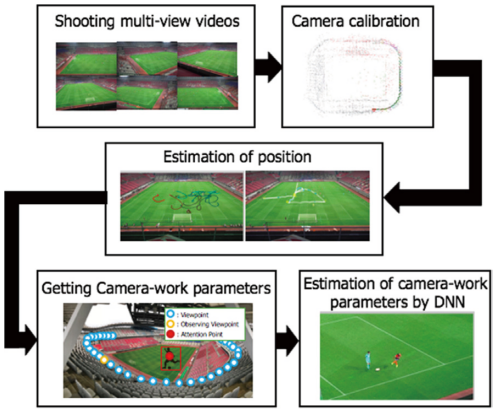

Fig. 2 Estimation camera-work of Bullet-time video for a dynamic scene.

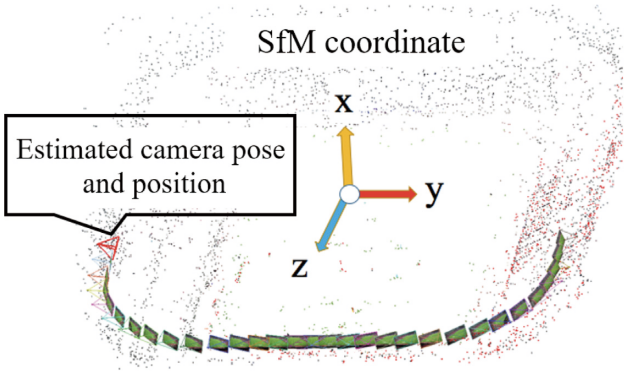

Fig. 4 Multiple captured images are applied to SfM. The position, pose, and focal length of each camera are estimated. At the same time, a sparse 3D point cloud is generated, which is not used in the proposed method. The acquired parameters (position, pose, focal length of each camera) are used to generate bullet time video.

deep neural network (DNN). A camerawork parameter estimation model based on the position is generated. Bullet-time videos applied to dynamic scenes are generated by estimating camerawork parameters using the generated model.

3.1 Shooting multiview videos and calibration

Our research assumes that multiview cameras are arranged densely for generating Bullet-time videos with smooth viewpoint switching. Fig. 3 shows an example of a camera arrangement method in a large-scale space such as a soccer stadium. The cameras are placed surrounding the target object as shown in Fig. 3 (red frame), and are fixed to the stadium structure such as the railing. As shown in Fig. 3 (yellow line), the cameras are aligned so that the optical axes intersect at the center of the area. The cameras are placed at the same angle between the optical axes of adjacent cameras. In addition, when the number of cameras is increased and the angle between the optical axes of adjacent cameras is narrowed, a Bullet-time video in which the viewpoint moves smoothly is generated.

To a generate Bullet-time video, the position, pose, and focal length of the multiview camera are obtained. Fig. 4 shows an example of applying SfM to the captured soccer field images, where the camera position, pose, and focal length are automatically estimated in the SfM coordinate system.

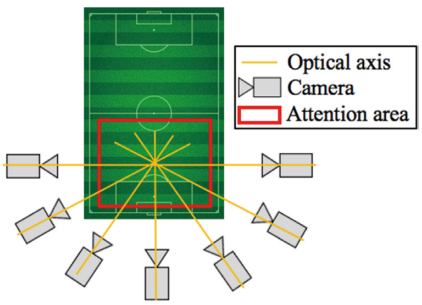

Fig. 3 Example of a camera arrangement in a large-scale space such as a soccer stadium.

3.2 Estimation method of player position

The proposed method assumes that many cameras are densely arranged. Therefore, the overlap region between the adjacent cameras is large, which is an excellent condition for obtaining volume information in a 3D space. This section proposes a 3D player position estimation method using background subtraction processing and voxel space voting.

The difference between the background image (upper left of Fig. 5) and the captured image (lower left of Fig. 5) is calculated to generate the difference image (right of Fig. 5). As shown in Fig. 6, we set a voxel space on the soccer field (i.e., we defined the origin, X axis, Y axis, and Z axis of the voxel space). Fig. 7 shows the relationship between the SfM coordinate system and the voxel coordinate system. To represent the cameras and soccer field in common 3D coordinates, we transformed the SfM coordinate system to the voxel coordinate system. The transformation matrix consists of a rotation matrix and a translation vector. As shown in Fig. 7, a region where two straight lines (edges) intersect perpendicularly and an object with a known size present in the capture scene of a multiview video is defined as the origin of the voxel coordinate system. The scale is determined by the ratio of the size in the SfM coordinate system corresponding to a known object in the voxel coordinate system (the length of the goal line and touch line of the soccer field).

The subtraction values obtained for all multiview images are projected onto the voxel space using the projective transformation matrix of each camera. This

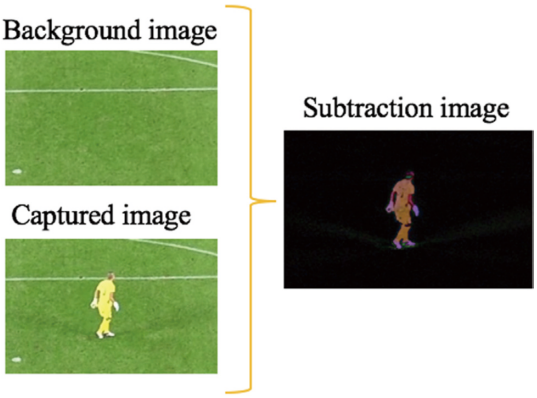

Fig. 5 Background subtraction.

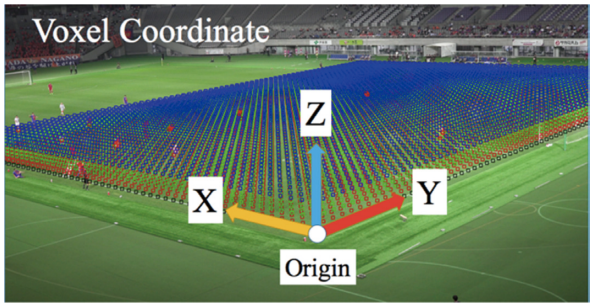

Fig. 6 Voxel coordinate.

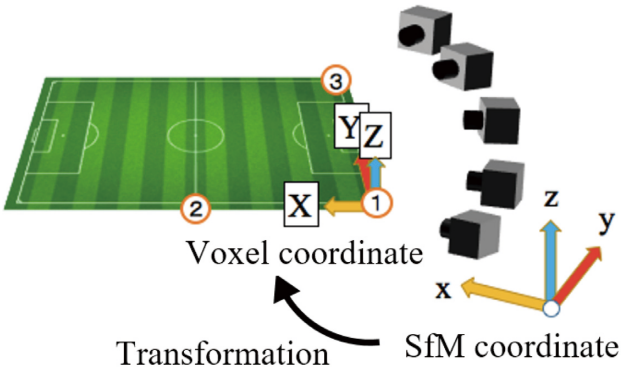

Fig. 7 SfM coordinate system and voxel coordinate system.

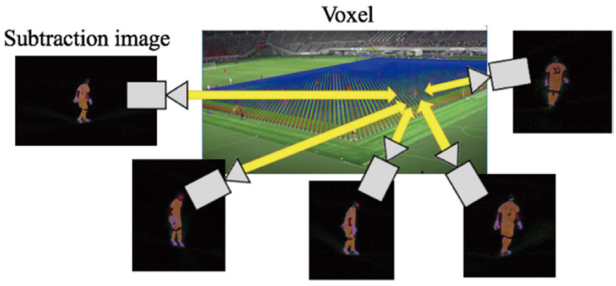

Fig. 8 Vote subtraction value to voxels as degree of foreground.

projection process is called voting. A voxel in which a foreground region such as a player exists has a high vote value, and a voxel in which no foreground region exists has a low vote value as shown in Fig. 8.

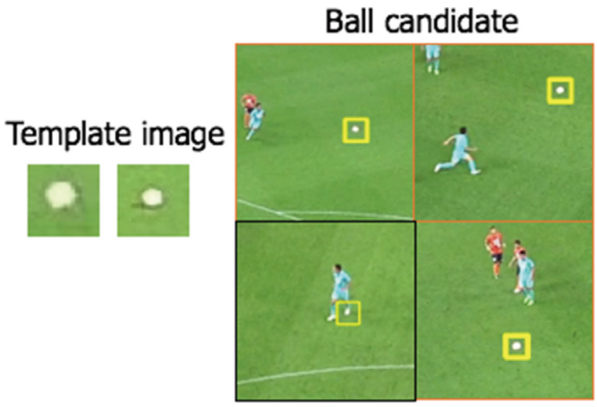

Fig. 9 Ball candidate position detected by the template image. The figure on the bottom left is false detection.

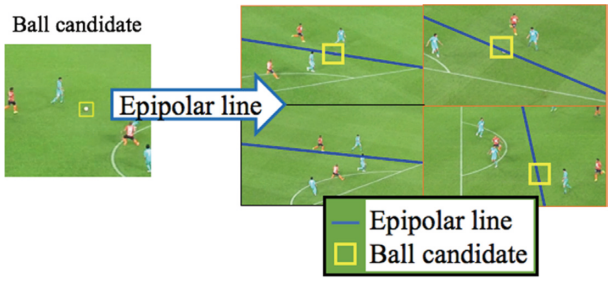

Fig. 10 Ball search on epipolar line. Balls can be detected with multiview.

3.3 Estimation method of ball position

A soccer ball is a sphere; therefore, it can be seen as a circle even when viewed from various angles. In addition, by utilizing the advantage of the dense camera layout, we propose a 3D position estimation method based on epipolar line search using the appearance of the ball shape.

As shown in Fig. 9, the ball candidate position in each image is detected by using the ball template image. The ball position is determined by judging the similarity threshold. The ball region can be detected correctly in the upper left, upper right, and lower right images as shown in Fig 9. However, the lower left image falsely detects the player's white spike as a ball. This problem is eliminated by the geometric restraint based on epipolar equations between multiple images.

The candidate ball position is searched on the calculated epipolar line. If the ball position is accurate, ball candidates are detected along the epipolar lines on multiview images as shown in Fig. 10. When there are two or more ball candidates, as shown in Fig. 11, all ball candidates are paired, and the 3D position by is estimated by the stereo vision. At this time, the median of all the estimated results is the 3D ball position. If no ball candidate is detected, interpolation is performed using the ball position estimated in the previous and

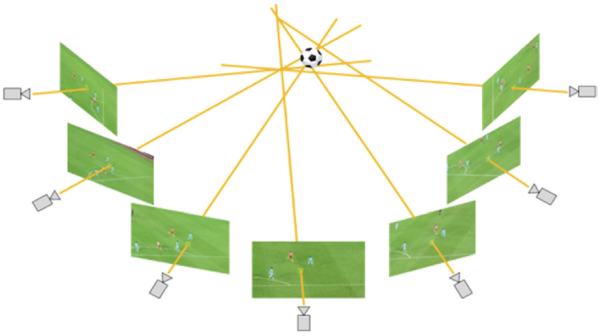

Fig. 11 3D position estimation by stereo vision.

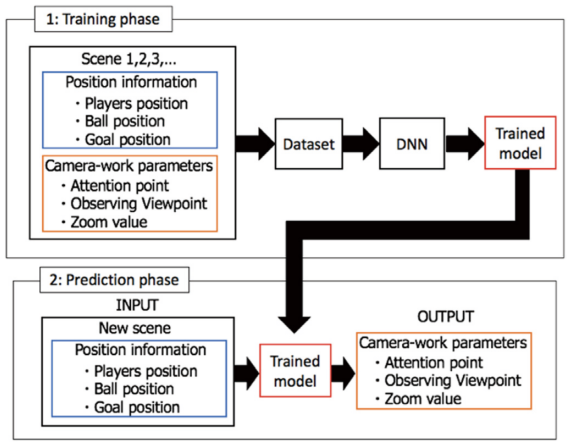

Fig. 12 Flow of camerawork generation method.

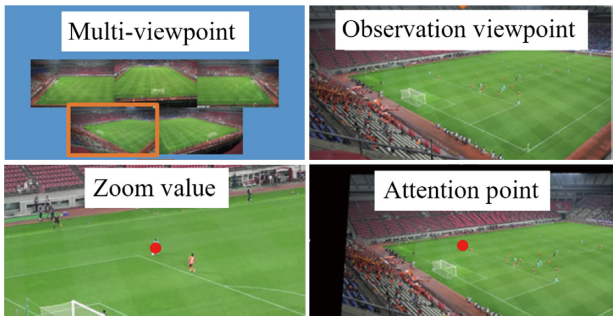

Fig. 13 Camerawork parameters given as the observation viewpoint, attention point, and zoom value.

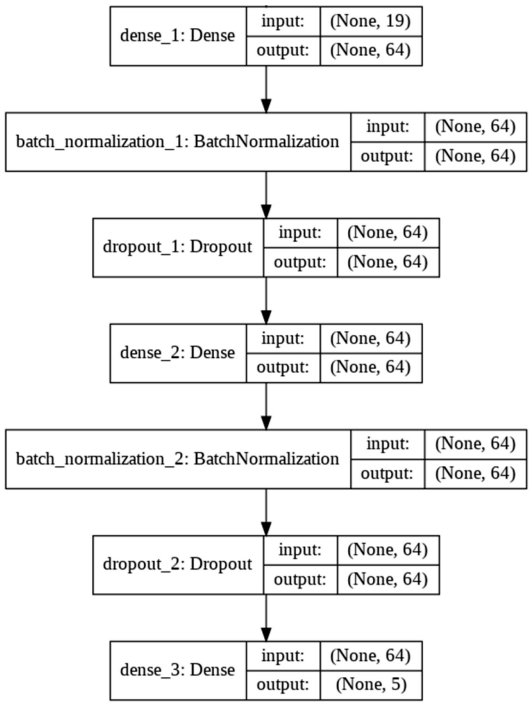

Fig. 14 Network structure.

next frames. A constant-velocity linear motion model is used in the XY plane, and a constant-acceleration linear motion model based on gravitational acceleration is used in the Z direction.

3.4 Camerawork estimation using DNN

This section proposes a camerawork parameter estimation method using a DNN-based on player, ball, and goal positions, as shown in Fig. 12. The camerawork setting for Bullet-time video should be customized for individual observers according to their preference. We propose a method using suitable DNN for estimating the camerawork parameters.

The upper part of Fig. 12 shows the learning phase. The training data uses position information and camerawork parameters. The position information provides the positions of players, balls, and goals. As shown in Fig. 13, the camerawork parameters are given as the observation viewpoint, attention point, and zoom value. The camerawork parameters used in the learning phase are the data obtained by the observer survey. The acquisition of ground truth data (attention point, observing viewpoint and zoom value) are as follows. In a participant set the camerawork parameters using the input devices. Multiview images were extracted every second and displayed on a monitor. The participant sets the camera-work parameters while checking the 5-s Bullet-time video.

Fig. 14 shows the network structure used. The network structure consists of three simple fully connected layers. Between each fully connected layer is a batch normalization layer and a dropout layer. The input is 19 dimensions, and the details are as follows. The number of players is multiply one place and one dimension. The position of the player closest to the left and right touch lines is multiply one place and three dimensions. The position of the player closest to the goal is multiply one place and three dimensions. The center position of all detected players is multiply one place and three dimensions. The above is player position information. Ball position information is multiply one place and three dimensions. Goal position information is multiply two place and three dimensions. The above

total is expressed in the input 19 dimensions. The output is 5 dimensions, and the details are as follows. The observing viewpoint is multiply one place and one dimension. The attention point is multiply one place and three dimension. The zoom value is multiply one place and one dimension. The above total is expressed in the output 5 dimensions. The positional information of player, ball, goal, and attention point are normalized by the size of the soccer field. The observation viewpoint is normalized by dividing with the number of cameras. The zoom value is normalized with the minimum and maximum values of the training data.

When evaluating machine learning models, the dataset is divided into three parts: training data, validation data, and test data. The machine learning model itself is adjusted (learned) using the training data, and the learning result is evaluated using the validation data that is not used for training. Finally, when the model learning is completed, the final prediction result of the model is evaluated using test data that is not used for training or validation. To prevent overlearning, the training data may be adjusted to learn with the number of epochs where the loss value is the lowest compared with the value of the past loss function.

The lower part of Fig. 12 is the prediction phase. The position information is input to the learned model. The output data are the predicted camerawork parameters. The automatic generation of the Bullet-time video adapted to the dynamic scenes is realized by using the obtained camerawork parameters.

4. Experiments and results

4.1 Environment for capturing images

In this research, we captured a video of the Japan Football Association (JFA) The 2017 Emperor's Cup. Fig. 15 shows the layout of the camera positions and the state of the installed cameras; 31 cameras were installed. They were 4K cameras (SONY FDR-AX 100). 30fps videos were captured. All cameras were fixed toward the marker (yellow circle) placed on the field as shown in Fig. 16. The angle between the optical axes of adjacent cameras was 6°. The video environment was processed using a PC with CPU: Intel Corei7-4770 3.40GHz and memory: 16.0GB. Bullet-time video generation was based on a program developed by Akechi et al. [5] The camera calibration used VisualSFM[16] of the public SfM library.

Fig. 15 Camera arrangement at Kashima soccer stadium.

Fig. 16 A camera installed at the handrail of the Kashima soccer stadium.

4.2 Results of player and ball position estimation methods

The position of the player and ball was within half the field depending on the camera installation conditions. For player position estimation, the voxel size was set to the half field size (57.5 m × 78 m), and the height was set to 2 m (greater than the height of a human). The voxel spacing was set to 0.2 m. After detecting the extreme value of the voxel value, if another extreme value existed in the surrounding region of 1 m × 1 m × 2 m, it was related to the same player. When the player position detected in the previous frame was within 0.4 m of the player position detected in the current frame, it was related between the frames as the position of the same player. This value was determined by referring to 19.19 s) of the 200 m run recorded by Usain Bolt [17], because a human cannot move at a speed greater than 0.4 m per frame. Fig. 17 shows an example of the tracking results of players. The movement trajectory of each player was divided into colors. The displayed trajectory was extracted at a height of z = 1 m. It can be confirmed that the player's position was acquired accurately.

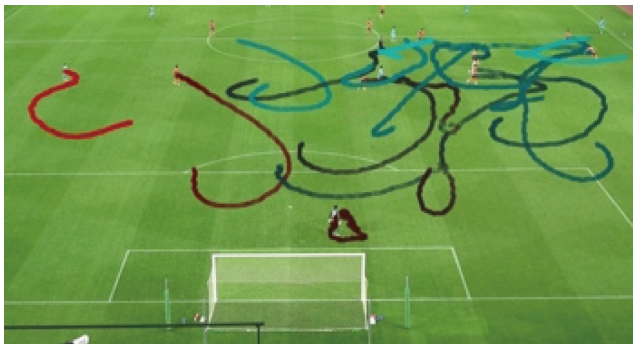

Fig. 17 Example of estimated player-trajectories.

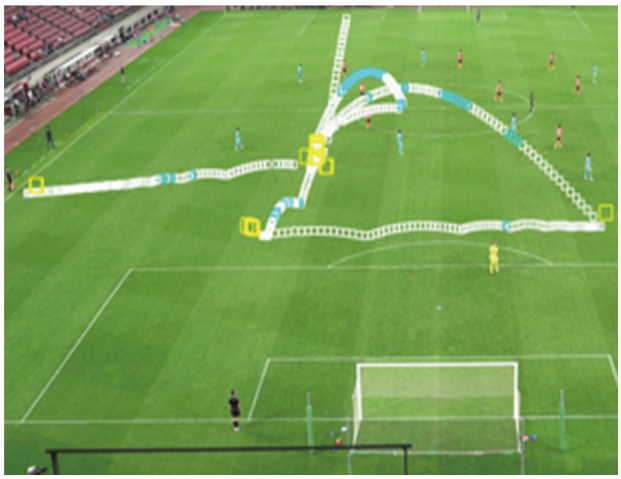

Fig. 18 Example of an estimated ball-trajectory.

Two types of template images were obtained for ball position estimation. The candidate ball position was detected. We used normalized cross-correlation (NCC) to evaluate the similarity of the template images and detected a value with an NCC value of 0.7 or greater as the ball candidate position. If there were no more than 15 viewpoints on the epipolar line, the ball was excluded from the candidate positions. The NCC threshold and the number of viewpoints were set by comparing the ball position detection results after extracting the four scenes of the ball's surroundings state (crowded / empty) with the ball's speed (fast / slow). Fig. 18 shows an example of the ball tracking results. The position interpolated by the player position is shown in yellow. The linearly interpolated position is shown in light blue. It can be confirmed that the position of the ball was estimated accurately.

4.3 Quantitative evaluation experiment

We generated a dataset to confirm the feasibility of the proposed method. A quantitative evaluation experiment was performed using the dataset to demonstrate its effectiveness with seven participants. The camera viewpoint is switched by a bullet time operation using video (each 5 seconds) captured by cameras of 31 units. A bullet time video including such a series of operations is defined as one scene. The data set is 5 scenes. As shown in Fig. 19 (a), a participant set the camerawork parameters using the input devices. Multiview images were extracted every second and displayed on a monitor. The participant sets the camera-work parameters while checking the 5-s Bullet-time video. In this research, 5 scenes of dataset were prepared. One scene is a video for 5 seconds. 30 images are played per second. That is, there are 150 images in one scene. The correct tag is attached to 150 images per scene. The image and correct tag are divided randomly

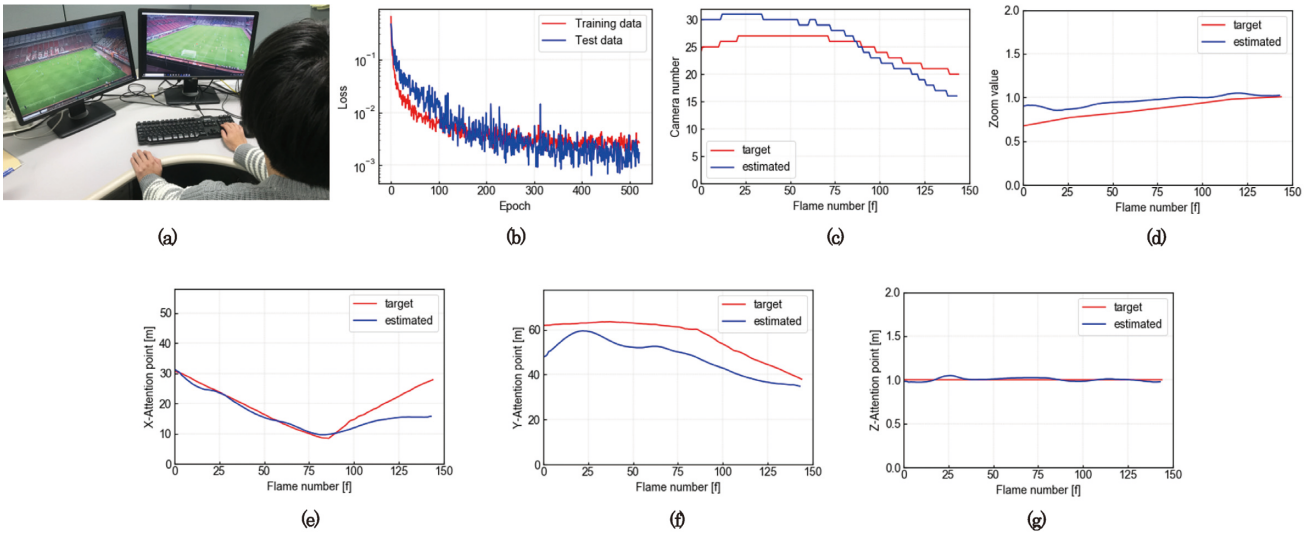

Fig. 19 (a) Acquisition of camerawork parameters from subject. (b)Learning result: Red uses learning data and blue uses training data. (c) Comparison of camera number. (d)Comparison of zoom value. (e) Comparison of X-attention point. (f) Comparison of Y-attention point. (g) Comparison of Z-attention point.

in one scene (150 images), but randomness across scenes is not executed. One of the five scenes was used for operation practice. Three scenes were used for model learning. The remaining 1 scene was used as an evaluation of the model prediction results. Fig. 19 (b) shows the learning results. Sum of squared differences was used as the loss function. The red line shows the result of using the learning data, and the blue line shows the result of using the test data. It can be confirmed that learning was completed before overlearning occurred.

Table 1 lists the average errors of each parameter generated by DNN (estimation) and each participant (correct data). The "No" column lists the participant number. X, Y, and Z are the errors in the attention point in the X, Y, and Z directions. "Cam" is the camera number error. For example, if an error of 1.0 occurs in "Cam," the viewpoint with the camera direction shifted by 6° is estimated. "Zoom" is the average of the zoom value errors (Range: 0.0-2.0). The result of the attention point in the X and Y directions is an average error value of 10 m or more. Recently, a 3D conversion system that can move the viewpoint of a monocular video of a soccer game has been proposed [18]. The input video is a general soccer broadcast video. In general soccer broadcast video, the touchline is divided into two so that the whole performance can be seen. In the user experiment reported by the method of [18], Viewpoint movement is performed so that the area where the touchline is divided into 4 parts can be seen from the video where the touchline is divided into 2 parts. Similarly, Viewpoint movement is performed so that you can see the area where the goal line is divided into three from the video that contains all the goal lines. The estimated average error of the proposed method is about 10.1 m for X and about 11.4m for Y. Therefore, the estimation error of the proposed method is not an error that greatly deviates from the region moving to the viewpoint the user wants to focus. Cam's errors are discussed in the next section, "Qualitative Evaluation Experiments". The average value of Zoom results is 0.156. This is a small error because the zoom value range is 0.0-2.0. The error rate is 7.8%.

Table 1 Average value of error.

No	X[m]	Y[m]	Z[m]	Cam	Zoom
1	2.7	8.9	0.016	2.9	0.097
2	24.3	2.1	0.044	5.8	0.155
3	16.3	8.2	0.040	15.5	0.063
4	9.4	14.9	0.089	11.0	0.348
5	6.9	23.1	0.110	13.3	0.194
6	6.2	14.4	0.119	9.7	0.264
7	4.9	8.0	0.057	4.8	0.078

X, Y, Z: Attention point errors
Cam: Camera number errors
Zoom: Average of the zoom value errors

By using the results of participant No. 1 in Table 1 as a representative example, Fig. 19 (c, d,e, f,g) shows the variation of each parameter generated by DNN (estimation: blue line) and each participant (correct data: red line). An example of a video at this time is shown in Fig. 20 and 21. Bullet-time video 1 for participant No. 1 in Table 1 is Fig. 20. Bullet-time video 2 obtained by using DNN is shown Fig. 21. Fig. 19 (c) shows the observation viewpoint estimation results. The estimated camera number is predicted to be the nearest camera number selected by the participant. Thus, the tendencies are the same, but not exactly the same. Fig, 19 (d) shows the estimation results of the zoom value, which were generally good. Fig. 19 (e, f,g) shows the estimation results of the attention points. In Fig. 19(e), there is a match in the first half and a difference in the

Fig. 20 Bullet-time video generated by participant (video 1).

Fig. 21 Bullet-time video generated by DNN (video 2).

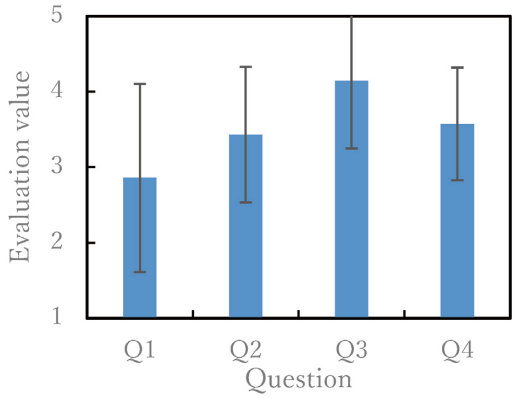

Fig. 22 Questionnaire results by testers.

second half. In Fig. 19(f), the waveform shapes match; however, there was a constant difference throughout. In Fig. 19 (g), the Z-axis direction did not change because the attention point was always set to 1 m of the player position. In the results of Fig. 19 (e, f,g), the estimation of the attention point can generally reflect the tendency of the value set by the participant.

4.4 Qualitative evaluation experiment

As a qualitative evaluation experiment, we compared Bullet-Time video 1 (Fig. 20) by tester No. 1 in Table 1 and Bullet-Time video 2 (Fig. 21) by DNN and asked them to answer questions. For questions, a five-point evaluation ("5: Extremely Satisfied", "4: Satisfied", "3: Neither", "2: Dissatisfied", "1: Extremely Dissatisfied"). The questions are shown below.

Q1. Is the attention point of video 2 the same as video 1?
Q2. Is the observation viewpoint of video 2 the same as video 1?
Q3. Is the zoom value of video 2 the same as video 1?
Q4. Is video 2 suitable for observing the game situation?

Fig. 22 shows the average and standard deviation of the evaluation values comparing Bullet-Time video 1 and 2. In Q1, the average evaluation value was 2.85, and the attention point could not be predicted well. However, the standard deviation is as large as 1.24, and the evaluation varies depending on the tester. Some testers gave a rating of 5. In Q2, the average evaluation value was 3.42. Since the viewpoints are densely arranged, the observation viewpoints do not change significantly if there are differences of several units, so the evaluation value may be relatively high. In Q3, the average evaluation value was 4.14, which was high. A good estimate of the zoom value with little error is obtained from the results in Table 1. In Q4, the average evaluation value was 3.57. Compared with the results for the attention points of Q1 and Q2, the evaluation value is high. Therefore, although it is different from the attention point and observation viewpoint set by the tester, videos suitable for observation of the game situation tend to be generated.

The previous method was a static bullet time video. The proposed method has been applied to dynamic bullet time video. Therefore, the efficiency of the situation grasp and the strategy analysis using the bullet time video is improved as compared with the previous method. This is because dynamic video confirmation has more information than static video confirmation. In the soccer scene example, a continuous flow from the center of the field to a shoot can be confirmed with a bullet time video. In addition, the proposed method generated bullet time video that can be applied to dynamic video by DNN. This realization eliminates the need for complicated manual operations by the user. Furthermore, when top athletes, their coaches, and directors create learning data, the bullet time video produced by the proposed method becomes a video with a high professional level. Therefore, the proposed method is a robust bullet time video generation method.

5. Discussion

5.1 The performance of the player extraction and the ball position estimation

In general, since the diameter of a soccer ball is 22 cm, the voxel spacing was set to 0.2 m in this research. Therefore, in the detection method using the voxel space, the voxel spacing is determined by the size of the object detection. In this research, the simplest template matching method is adopted as an object detection method. The similarity threshold at this time is set to 0.7. For this value, template matching was applied to four scenes in advance and the similarity threshold was investigated. According to the results of the investigation, the ball detection accuracy decreased when the similarity threshold was high. This is due to the slightly different appearance of the ball in the template image and the ball in the input image. When the similarity threshold is low, the ball is determined in the region other than the ball. The threshold of similarity was determined by such a preliminary survey. Next, the threshold of the viewpoints number was investigated. The threshold of the viewpoints number indicates how many images can be detected by re-searching on the epipolar line when the ball is correctly detected. When the threshold of the viewpoints number

is large, it cannot adapt to the situation where the ball is hidden by the player. When the threshold of the viewpoints number is small, a candidate region other than the ball may be detected. In this research, the threshold of the viewpoints number was set to 15 or more by such a preliminary survey. In the player region, when the camera arrangement is not set so as to surround the field or when the number of cameras is small, it cannot adapt to the occlusion between the players. For such problems, the number of cameras and the arrangement method were determined by a preliminary survey. When applying to other scenes, it is necessary to calculate an optimum parameter by such a preliminary investigation. When the detection accuracy of the player or the ball is low, it is difficult to always adjust the attention point to the player or the ball in the dynamic bullet time video. This is a factor that lowers the level of grasping the situation. Therefore, the easy-to-understand viewing method of dynamic bullet time video and the position estimation accuracy of players and balls are relative relationship.

5.2 Comparison with other related works

In the proposed method, it is possible to do the same as the method of akechi et al [5] by pausing while confirming the dynamic bullet time video. The proposed method adapts to dynamic bullet time video. The novelty is that the viewer can check the video of various viewpoints while checking the flow of the soccer scene. In this respect, the situation of the soccer scene can be grasped more than the method of akechi et al. In previous research, there were cases where multiple cameras were installed in a large stadium to generate bullet-time video, but the target scene was fixed [19]. For example, baseball is near the fielder or near the base. Basketball is near the goal. Thus, there is an example of bullet time video using competition characteristics. On the other hand, in soccer, players move across the field extensively. The previous method can be applied to shoots near the goal, but cannot be applied to long shoots or shoots from dribbling. In this case, if the camera's viewpoint is not fluid, the game situation cannot be followed. Therefore, in order to grasp the situation of soccer, it is required to grasp the situation using a dynamic bullet time video instead of a static bullet time video. The proposed method is suitable for grasping the situation of a fluid competition video such as a soccer scene because the competition continues even while the viewpoint is changed.

5.3 Limitation of the proposed method

In the proposed method, 31 cameras were installed in a large-scale soccer stadium to realize dynamic bullet time video generation. The convergence angle of the camera at this time is 6°. Therefore, as a condition for applying the proposed method, the same condition is required to maintain the video quality. Furthermore, high-resolution shooting is recommended to maintain video quality. In addition, it is desirable that the angle of view and the focal length of all cameras be unified. This is because correspondence information of image feature between adjacent images is required to calculate the camera position and pose. An environment that can be placed around a camera like a stadium is very important. It is a situation where the proposed method is highly applicable.

6. Conclusion

In this research, we proposed the automatic estimation of the multiview video switching method that can be applied to dynamic scenes in a large-scale space. We chose a soccer game as the dynamic/large scene. The camerawork parameters of the Bullet-time video were estimated by DNN using the player and ball positions. The results of our experiments showed that the proposed method tended to generate videos suitable for observing the game situation.

This work was partially supported by JSPS KAKENHI Grant Number 17H01772 and by JST CREST Grant Number JPMJCR14E2, Japan.

References

1) Jean-Yves Guillemaut and Adrian Hilton: "Joint Multi-Layer Segmentation and Reconstruction for Free-Viewpoint Video Applications", International Journal of Computer Vision Volume 93 Issue 1, pp.73-100 (2011)
2) Joel Carranza, Christian Theobalt, Marcus A. Magnor and Hans-Peter Seidel: "Free-viewpoint video of human actors", ACM SIGGRAPH 2003 Papers, pp.1-9 (2003)
3) Aljoscha Smolic, Karsten Mueller, Philipp Merkle, Christoph Fehn, Peter Kauff, Peter Eisert and Thomas Wiegand: "3D Video and Free Viewpoint Video - Technologies, Applications and MPEG Standards", IEEE International Conference on Multimedia and Expo, pp.2161-2164 (2006)
4) Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk and Steve Sullivan: "High-quality streamable free-viewpoint video", ACM Transactions on Graphics (TOG) Volume 34 Issue 4 (2015)
5) Nao Akechi, Itaru Kitahara, Ryuuki Sakamoto and Yuichi Ohta: "Multi-Resolution Bullet-Time Effect", ACM SIGGRAPH-Asia (2014)
6) Kensuke Ikeya and Yuichi Iwadate: "Multi-viewpoint robotic cameras and their applications", ITE Transactions on Media Technology and Applications Volume 4 Issue 4, pp.349-362 (2016)
7) Christine Chen, Oliver Wang, Simon Heinzle, Peter Carr, Aljoscha

Smolic and Markus Gross: "Computational Sports Broadcasting: Automated Director Assistance for Live Sports", IEEE International Conference on Multimedia and Expo (ICME) (2013)
8) Hao Jiang, Sidney Fels and James J. Little: "Optimizing Multiple Object Tracking and Best View Video Synthesis", IEEE Transactions on Multimedia, Volume 10, Issue 6, pp.997-1012 (2008)
9) Xueting Wang, Takatsugu Hirayama and Kenji Mase: "Viewpoint Sequence Recommendation Based on Contextual Information for Multiview Video", IEEE MultiMedia, Volume 22, Issue 4, pp.40-50 (2015)
10) Sachiko Iwase and Hideo Saito: "Parallel Tracking of All Soccer Players by Integrating Detected Positions in Multiple View Images", International Conference on Pattern Recognition, pp.751-754 (2004)
11) Nozomu Kasuya, Itaru Kitahara, Yoshinari Kameda and Yuichi Ohta: "Robust trajectory estimation of soccer players by using two cameras", International Conference on Pattern Recognition (ICPR2008), pp.1-4 (2008)
12) Jinchang Ren, James Orwell , Graeme A. Jones and Ming Xu: "Tracking the soccer ball using multiple fixes cameras", Computer Vision and Image Understanding, Volume 113, pp.633-642 (2009)
13) Norihiro Ishii, Itaru Kitahara, Yoshinari Kameda and Yuichi Ohta: "3D Tracking of a Soccer Ball Using Two Synchronized Cameras", Pacific Rim conference on Advances in multimedia information processing, pp.196-205 (2007)
14) Takasuke Nagai, Hidehiko Shishido, Yoshinari Kameda, Itaru Kitahara, "An On-site Visual Feedback Method Using Bullet-Time Video," ACM Multimedia Conference (1st International Workshop on Multimedia Content Analysis in Sports), pp.39-44 (2018)
15) Changchang Wu: "Towards Linear-time Incremental Structure from Motion", International Conference on 3D Vision, pp.127-134 (2013)
16) Changchang Wu: "VisualSFM: A Visual Structure from Motion System", http://ccwu.me/vsfm
17) J J Hernandez Gomez, V Marquina and R W Gomez: "On the performance of Usain Bolt in the 100 m sprint", European Journal of Physics 34(5), pp.1227-1233 (2013)
18) Konstantinos Rematas, Ira Kemelmacher-Shlizerman, Brian Curless and Steve Seitz: "Soccer on Your Tabletop", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4738-4747 (2018)
19) https://www.4dreplay.com/4dreplay

Hidehiko Shishido received his M.E. and Ph.D. degrees in Engineering from the University of Tsukuba, Japan, in 2013 and 2016, respectively. In 2016, he served as a researcher at the Japan Institute of Sports Sciences. Since 2017, he has been an Assistant Professor at the University of Tsukuba. In 2018, he was a visiting academic researcher at University of Surrey, UK. His research interests include computer vision and sports science.

Yosuke Okada received his Master of Engineering from the University of Tsukuba, Japan, in 2019, His research interests are computer vision.

Yoshinari Kameda received his B.E and M.E and Ph.D from Kyoto University in 1991, 1993, and 1999. He had a faculty position at Kyoto University in 1999-2003. He was a visiting scholar at MIT in 2001-2002. In 2003 he joined the University of Tsukuba and he is a professor at University of Tsukuba. His research interests include the enhancement of human vision, augmented reality, mixed reality, video media processing, computer vision, and sensor fusion.

Masaaki Koido received his Physical Education degrees in Faculty of Health and Sports Science from University of Tsukuba, Japan in 2000 and 2003, respectively. He has been an assistant Professor at the University of Tsukuba and head coach of football club since 2013. His research interests include sports coaching, coaches education.

Itaru Kitahara received his B.E. and M.E. degrees in Science Engineering from University of Tsukuba, Japan in 1994 and 1996, respectively. In 1996, he joined Sharp Corporation. 2000-2003, he was a research associate of University of Tsukuba. He received his PhD in 2003. 2003-2005, he was a researcher at ATR. 2005-2019, he was an assistant professor and associate professor at the University of Tsukuba. Since 2019, he has been a professor at the University of Tsukuba. His research interests include computer vision, mixed reality, and intelligent image media.