A Player Position Tracking Method Based on a Wide-Area Pan-Tilt-Zoom Video

Shunzo Yamagishi[1], Chun Xie[2][a], Hidehiko Shishido[3][b] and Itaru Kitahara[2][c]
[1]*Master's and Doctoral Program in Intelligent Mechanical Interaction Systems, University of Tsukuba, Ibaraki, Japan*
[2]*Center for Computational Sciences, University of Tsukuba, Ibaraki, Japan*
[3]*Faculty of Science and Engineering, Department of Information Systems Engineering, Soka University, Tokyo, Japan*

Keywords: Sports Analytics, Player Tracking, Wide-Area, Video Processing, Pan-Tilt-Zoom Camera.

Abstract: This paper proposes a method to estimate the posture of an athlete moving on a vast field in a sporting event using a pan-tilt-zoom camera. In order to estimate the posture of an athlete on a sports field from a dynamic video sequence, our method extracts image features to search correspondence points among successive frames and computes the homography transformation matrix that compensates for changes the camera parameters (e.g., angle of view and posture). The effectiveness of this method is qualitatively verified using MLB TV broadcast video. The accuracy and error factors are also quantitatively verified by CG simulation.

1 INTRODUCTION

The use of video data for player evaluation in sports is advancing. In team sports played on large fields, such as soccer and baseball, both individual performance and team coordination are analyzed. Detailed physical movement data is crucial for evaluating individual skills, such as dribbling or catching. Conversely, data on player positions and the positional relationship among the ball and players are used to evaluate passing and hitting techniques and to analyze tactics like formations.

To evaluate individual physical movements, local information around the players is needed, typically acquired using telephoto tracking shots. For tactical analysis, capturing wide-area information of the entire field with fixed wide-angle shots is more effective. However, due to limitation of camera resolution, it is challenging to capture both local and wide-area information with a single camera, while using multiple cameras introduces extra cost and sometimes unfeasible in competitive situations due to restrictions imposed by competition organizations.

A PTZ (pan-tilt-zoom) camera can change its field of view (zoom) and posture (pan and tilt), enabling both telephoto tracking and wide-area fixed shooting depending on the situation. Since the shooting region of a PTZ camera dynamically changes, it is necessary to estimate the shoot region of the field in every frame. Typically, standardized field markers set up in accordance with competition regulations, such as lines, are used for alignment. However, during telephoto shooting, only a small portion of the field is captured, making alignment difficult due to the lack of visible landmarks.

This paper aims to solve this issue by estimating changes in the camera's shooting area using natural feature points instead of relying on explicit landmarks. Instead of mapping player positions directly from a camera frame to the field, we transform them to a reference frame where landmarks are clearly observed in a sequential manner. Our method can map player positions to the filed coordinate using videos taken by a PTZ camera that switch between wide-angle and telephoto tracking. In the experiment, the effectiveness of our method and the errors associated with estimating player positions are discussed.

[a] https://orcid.org/0000-0003-4936-7404
[b] https://orcid.org/0000-0001-8575-0617
[c] https://orcid.org/0000-0002-5186-789X

Yamagishi, S., Xie, C., Shishido, H. and Kitahara, I.
A Player Position Tracking Method Based on a Wide-Area Pan-Tilt-Zoom Video.
DOI: 10.5220/0012920100003828

In *Proceedings of the 12th International Conference on Sport Sciences Research and Technology Support (icSPORTS 2024)*, pages 46-53
ISBN: 978-989-758-719-1; ISSN: 2184-3201

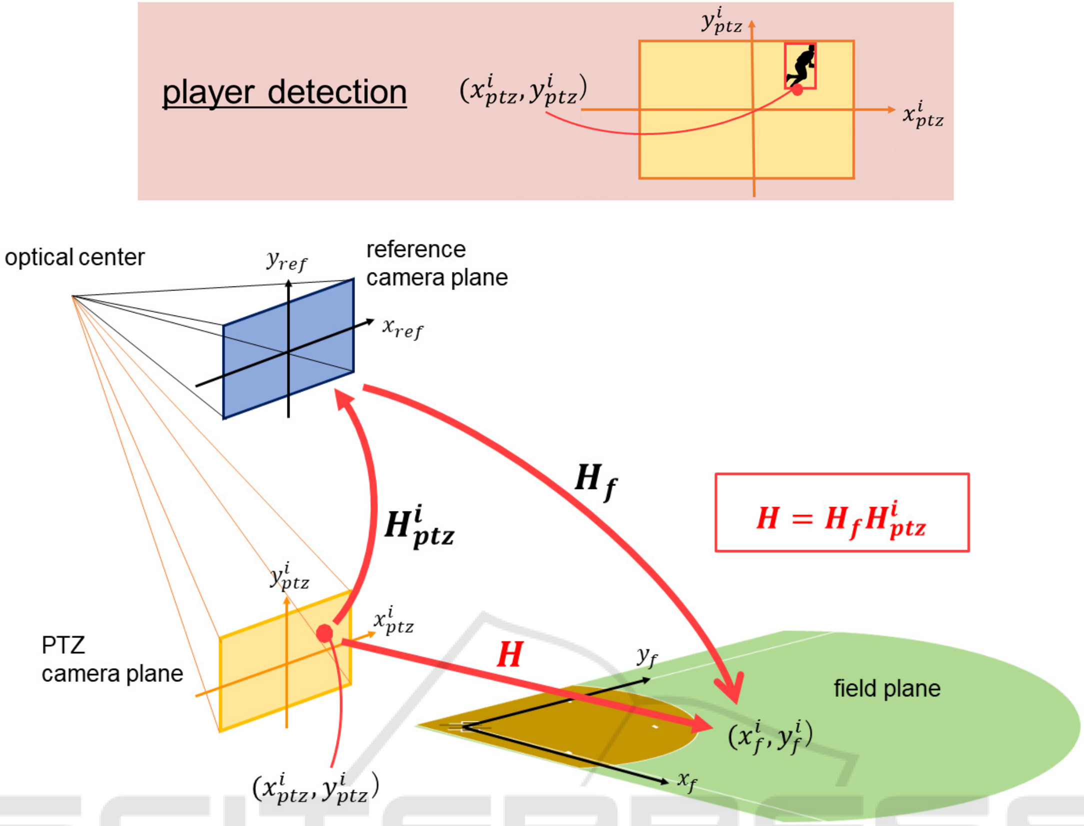

Figure 1: Overview of our proposed method: The geometric relationship between each frame of the PTZ camera and the field $\boldsymbol{H}$ can be determined by the homography matrix $\boldsymbol{H_f}$ between the reference frame and the field, and the homography matrix $\boldsymbol{H^i_{ptz}}$ between each frame i and the reference frame.

2 RELATED WORK

Player tracking has been applied to various kinds of sports. In basketball, Lu et al. proposed a method for tracking players using a single pan-tilt-zoom camera, and applied it to the analysis of player performance on the court by estimating the homography between video frames and the court (Lu et al., 2013). In football, a method was proposed to analyze the movement trajectories of players on the field by automatically estimating camera parameters using optical flow and detecting players with a Kalman filter from match video (Beetz et al., 2007). In ski racing, research visualizes the movement trajectory of players from camera video fixed to a tripod (Dunnhofer et al., 2023).

Due to the development of deep learning, object detection methods have been proposed that can extract effective features and perform class identification from small regions in low-resolution video or images (Redmon et al., 2016; Ren et al., 2016). As a result, in the field of sports, attempts have been made to detect and track small regions of players and balls in video from drone and fixed-camera video of the entire field (Katić et al., 2024).

There is also a lot of research being done on field alignment, which seeks to determine the geometric relationship between the field coordinate system and the camera image coordinate system. For example, a method has been proposed for adjusting the position of the parallel lines and vanishing points of a field using the energy maximization problem of a Markov field (Homayounfar et al., 2017). Additionally, there are research examples that use regression networks to align images of sports fields with template images of fields (Jiang et al., 2020; Shi et al., 2022). Another proposed approach is to create a database and query it with images of edges extracted from input images. (Chen & Little, 2019). These methods seek to establish a correspondence between frames and fields by directly matching field features for each frame. Therefore, they rely on accurately detecting field

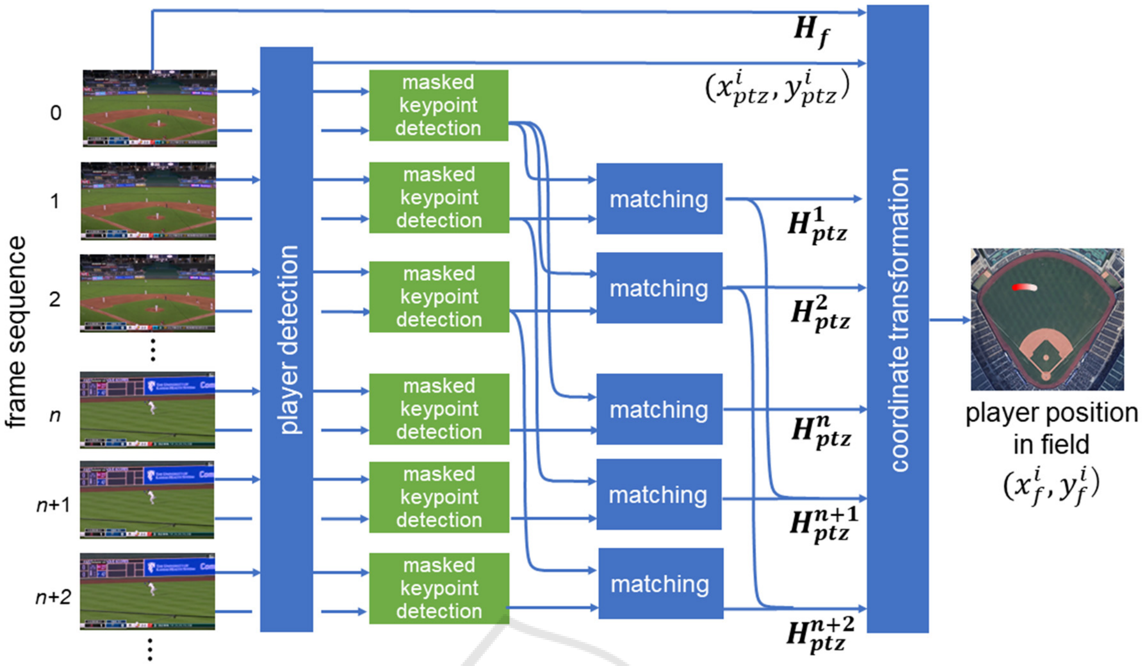

Figure 2: Process flow of the proposed method. Mask player regions in video frames based on player detection results. Feature points are detected from the masked frames. The frame of interest is compared with the previous n frames to obtain the transformation matrix $\boldsymbol{H}^i_{ptz}$ to the reference frame.

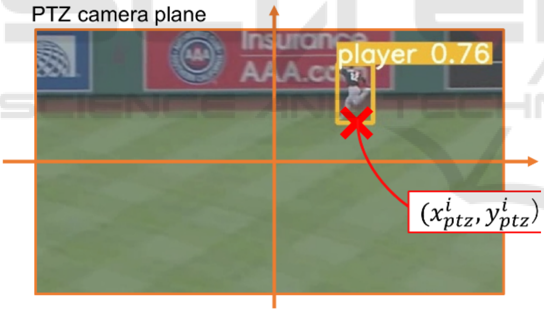

Figure 3: Detection of players by object detection network. The center of the lower edge of the resulting bounding box is used as the position of the player (x^i_{ptz}, y^i_{ptz}) in the PTZ camera frame.

features such as straight lines and circles.

In baseball, the image features available for matching between video frames are too sparse to estimate robust inter-frame transformations. To address this problem, we set a reference frame and calculate the transformation between each frame and the reference frame, as well as the transformation between the reference frame and the field coordinate system. This allows us to determine the position of the player on the field even when reliable image features are not available.

3 METHOD

Figure 1 shows the overview of our proposed method, and Figure 2 details the process flow. Our objective is to detect a target player in a video involving various camera motions and fields of view, and map the player's position to a field coordinate system via a homography transformation. Here, we denote points in three coordinate systems as follows:

- (x^i_{ptz}, y^i_{ptz}): coordinate in the i^{th} PTZ camera frame
- (x_{ref}, y_{ref}): coordinate in the specified reference frame of the PTZ camera
- (x_f, y_f): coordinate on the field

3.1 Player Detection and Tracking in a PTZ-Video Frame

An object detection neural network is used to detect and track the player in a frame of the shot PTZ-video. The center of the lower edge of the resulting bounding box, as shown in Figure 3, is used as the position of the player (x^i_{ptz}, y^i_{ptz}). We also apply a Multi-Object Tracking (MOT) technique to ensure a target player can be traced across different frames.

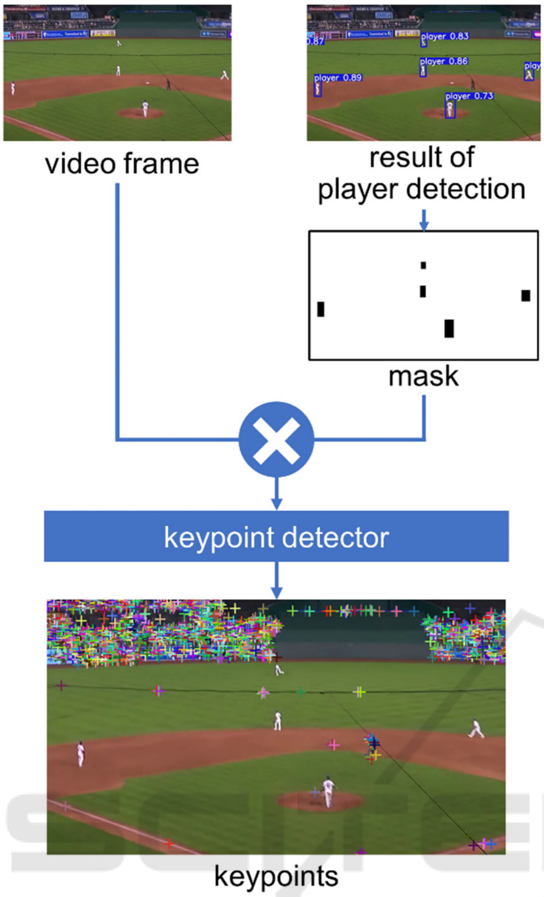

Figure 4: Masked feature detection: We use the bounding box of detected players to generate a mask image for each frame and remove the feature points in the masked area. The remaining features are consistent in terms of frame-to-frame transformation, making them robust for feature matching.

3.2 Player Position Mapping

Since the PTZ camera keeps stationary (e.g., fixed the environment using a tripod) and the shooting distance of sports scene is relatively long, it can be assumed that the optical center of the camera is also fixed while shooting. Therefore, the image coordinate system of each frame can be transformed using a homography matrix. The relationship between the coordinates in the PTZ camera plane $(x_{ptz}^{i}, y_{ptz}^{i})$ and the coordinates in the reference camera plane(x_{ref}, y_{ref}) can be described using the homography matrix $\boldsymbol{H}_{ptz}^{i}$ as shown in equation (1).

$$\begin{pmatrix} x_{ref} \\ y_{ref} \\ 1 \end{pmatrix} \sim \boldsymbol{H}_{ptz}^{i} \begin{pmatrix} x_{ptz}^{i} \\ y_{ptz}^{i} \\ 1 \end{pmatrix} \tag{1}$$

$\boldsymbol{H}_{ptz}^{i}$ can be estimated via feature matching between a PTZ camera frame and the reference frame. However, the camera field of view differs greatly between frames, and features can be extremely sparse in zoomed-in shots where the observed area is very narrow. As a result, it is usually difficult to find enough robust feature pairs to register a PTZ camera frame directly to the reference frame.

To address this problem, we match the target frame with the previous n frames, instead of the reference frame, to estimate a local transformation through Least Squares Method and Direct Linear Transform(Hartley & Zisserman, 2003). The local transformations are then incorporated back towards the reference frame to obtain the homography between the target frame and the reference frame $\boldsymbol{H}_{ptz}^{i}$. To make the estimation of local transformations more robust, we filter out unreliable features which majorly detected on the moving players from feature matching. This is done by using a mask image generated from the bounding boxes of players produced by an object detection network. The process is demonstrated in Figure 4.

Given the homography transformation $\boldsymbol{H}_{f}$ between the reference camera plane and the field plane, we can map the player coordinates $(x_{ptz}^{i}, y_{ptz}^{i})$ in the i^{th} frame to the field plane coordinates (x_f, y_f) via Equations (2) and (3). Note that $\boldsymbol{H}_{f}$ can be easily obtained by specifying the field coordinates of several representative points on the reference camera image.

$$\boldsymbol{H} = \boldsymbol{H}_{f}\boldsymbol{H}_{ptz}^{i} \tag{2}$$

$$\begin{pmatrix} x_f \\ y_f \\ 1 \end{pmatrix} \sim \boldsymbol{H} \begin{pmatrix} x_{ptz}^{i} \\ y_{ptz}^{i} \\ 1 \end{pmatrix} \tag{3}$$

4 EXPERIMENTS

To demonstrate the application of our method, two types of evaluations are performed. The first evaluation is qualitative, using real-world video of an outfielder catching a fly ball. The second evaluation is

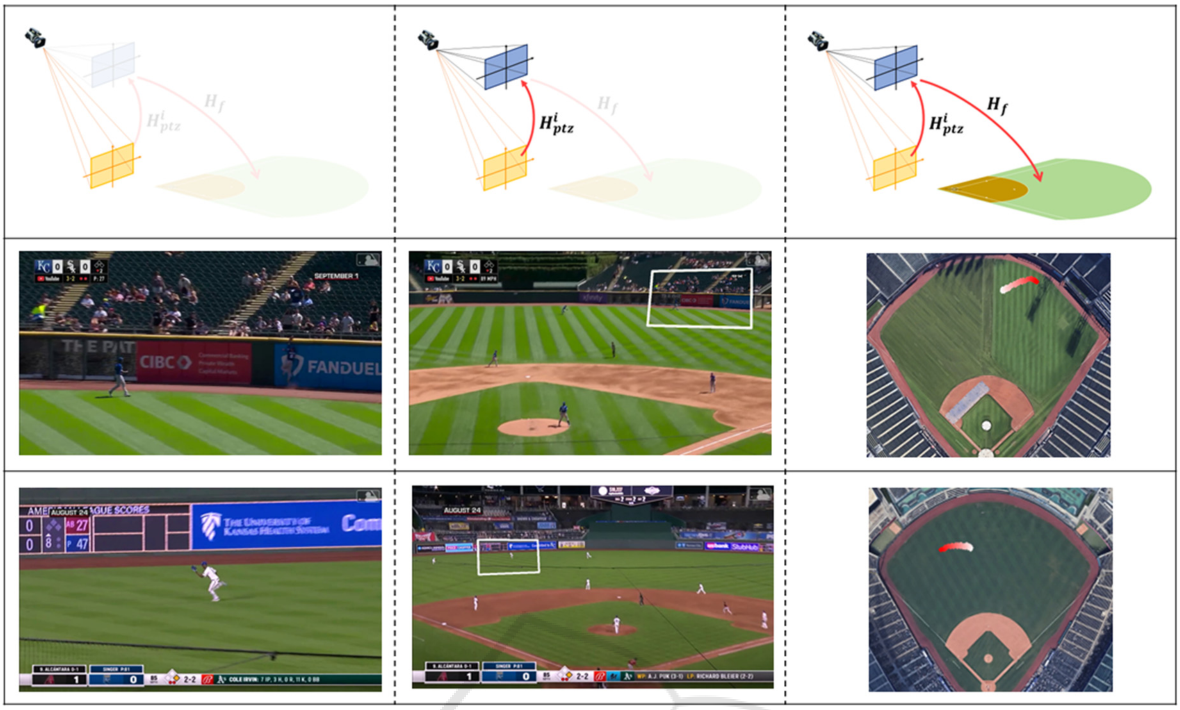

Figure 5: Results using MLB live video. Left: the frame of interest. Middle: the reference frame, with the frame of interest registered in the area marked by the white rectangle. Right: estimated player positions visualized on the aerial image. The player is moving from white point to red point.

quantitative, using a CG simulation. For object detection, we adopt YOLOv8 (Jocher et al., 2022/2023) and ByteTrack (Zhang et al., 2022) for multi-object tracking. SIFT (Lowe, 2004) is used for feature extraction for feature matching. In addition, the matching interval n is set to 30 frames.

4.1 Live-Action Video Evaluation

We conducted a qualitative experiment using PTZ video videos shot in a real-world baseball game. The video was sourced from MLB FILM ROOM (MLB.com, 2024) and primarily captured using PTZ cameras for TV broadcasts. We clipped scenes where the outfielder catches the ball from the cited video and used these as our input. We manually masked out the chyrons in the video as they contain undesired features.

To visualize player positions, we used aerial photographs from Google Earth (Google LLC, 2024). Figure 5 illustrates the results. The left column of Figure 5 shows a zoomed-in frame, and the middle column shows this frame registered to the reference frame using $\boldsymbol{H}^i_{ptz}$. The right column demonstrates that even when the camera's field of view differed greatly from the reference frame, we can still successfully map the positions of the players on the field.

4.2 CG Environment Evaluation

We use Unity (Unity Technologies, 2023b) to create the CG simulation and incorporated the Baseball Stadiums Pack (Distinctive Developments Ltd, 2017) and Starter Assets - ThirdPerson (Unity Technologies, 2023a) for the player models To improve object detection accuracy, we changed the color of the player model and fine-tuned the object detection mode. Using this environment, we rendered a video that includes PTZ camera work, reproducing an outfielder catching a ball. The video was rendered in Full HD quality at 30fps.

4.2.1 Decomposition of Error

The error in player position in the field plane is evaluated by breaking it down into depth and lateral directions as viewed from the camera. Figure 6 illustrates the overview. First, the camera optical axis is obtained from the true values of the camera parameters. Next, the camera optical axis is orthogonally projected onto the field plane. The error in the position of the player in the direction of the orthographic projection of the optical axis is the *D-error*, and the error in the component perpendicular to the *D-error* is the *L-error*. *D-error* represents the error in

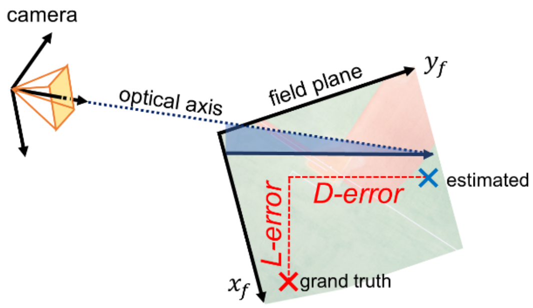

Figure 6: An overview of the decomposition of player position error, where *D-error* indicates the error in the depth direction as seen from the camera and *L-error* indicates the error in the lateral direction as seen from the camera.

the depth direction as seen from the camera, and *L-error* represents the error in the lateral direction as seen from the camera.

4.2.2 Player Position Error on the Field Plane

We analyze the errors in the estimated player positions (x_f, y_f) compared to the ground truth $(\overline{x_f}, \overline{y_f})$, as shown in the left column of Figure 7. The potential sources of error are:

- The error in the player's position (x^i_{ptz}, y^i_{ptz}) on the PTZ camera plane due to object detection.
- the estimated error in the homography transformation matrix $\boldsymbol{H}^i_{ptz}$ from the PTZ camera plane to the reference camera plane.

To further investigate the effects of each error source, we calculated the overall errors using the following setups:

- Use ground truth player positions in PTZ camera frames and map them to the field plane using the estimated homography transformation. The result is shown in Figure 7 (middle)
- Use estimated player positions in PTZ camera frames provided by object detection and map them to the field plane using the ground truth homography transform. The result is shown in Figure 7 (right)

As can be seen from the middle and right columns of Figure 7, the error is much more significant when using estimated player positions.

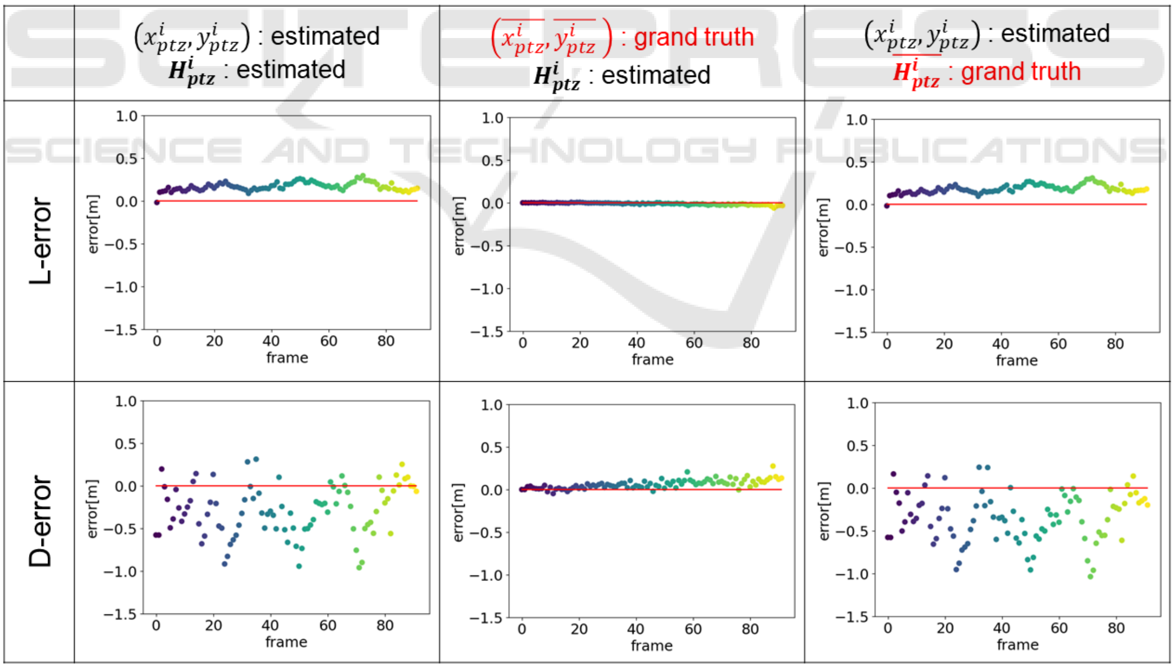

Figure 7: The player position error is shown decomposed into the camera depth direction (*D-error*) and the lateral direction as seen from the camera (*L-error*). (left) when object detection yields an estimate of the player's position on the PTZ camera plane and matching estimates the transformation to the reference camera plane; (center) when true values are given for the player's position on the PTZ camera plane from the camera parameters; (right) when true values are given for the transformation to the reference camera plane. If given. We see that the error in the player's position in the field coordinate system is greater in the depth direction of the camera. Comparing the rows also shows that most of the error is due to the detected position on the PTZ camera plane.

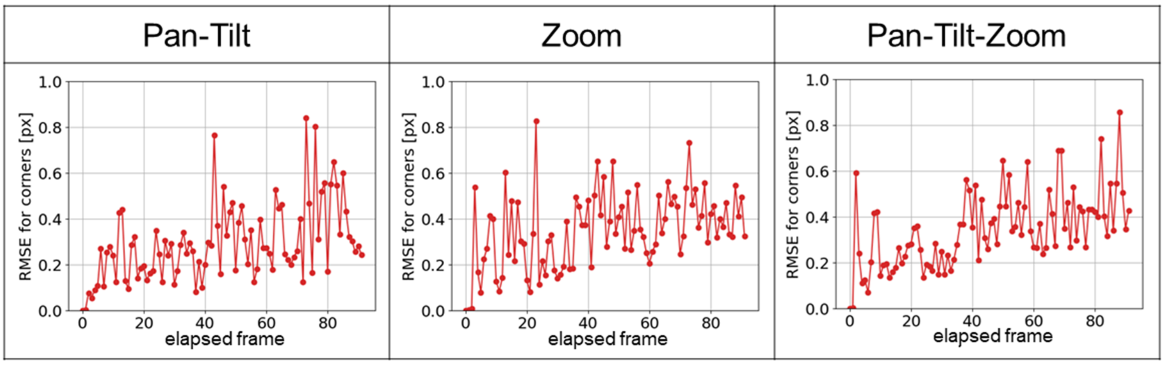

Figure 8: The relationship between the number of frames elapsed from the reference frame and the accuracy of $\boldsymbol{H}_{\boldsymbol{ptz}}^{\boldsymbol{i}}$ estimation (square root of the mean square error of the corners). Homographic errors accumulated as the number of synthetic transformations increased. However, no differences were found between the different camera work.

4.2.3 Effect of PTZ Camera Motion

We evaluate the accuracy of $\boldsymbol{H}_{\boldsymbol{ptz}}^{\boldsymbol{i}}$, by using root mean squared error (RMSE) of the four corners of the transformed camera frames on the reference frame. To investigate the effect of each camera operation and distance to the reference frame on $\boldsymbol{H}_{\boldsymbol{ptz}}^{\boldsymbol{i}}$, we use videos with only pan-tilt, only zoom, and all PTZ movements. The results, shown in Figure 8, indicate that homography errors accumulate as the number of composite transformations increases.

4.2.5 Discussion

The experimental results indicate that the position estimation error for players in short videos is primarily due to the object detection result $\left(x_{ptz}^{i}, y_{ptz}^{i}\right)$ on the PTZ camera plane. However, in longer videos, the errors caused by $\boldsymbol{H}_{\boldsymbol{ptz}}^{\boldsymbol{i}}$ accumulate as the transformation matrix is integrated from the PTZ frame to the reference frame.

Additionally, defining the player's position as the bottom-center of the bounding box is not strictly accurate since players often jump and land repeatedly. This is especially true for shallow tilt angles, which cause significant errors in the depth direction of the camera (*D-error*). Improving player position estimation accuracy requires developing more precise methods for detecting players on the PTZ camera plane.

5 CONCLUSIONS

This paper proposed a method to determine the positions of players on a large field during sporting events from videos captured using a PTZ camera where the camera pose and FOV change dynamically based on the context. Our approach estimates player positions by finding corresponding points between consecutive frames using image features and calculating a homography transformation matrix to map the player positions from the dynamic camera frame to the static field plane.

We validated the effectiveness of our method with MLB TV broadcast video and further verified its accuracy and error sources using a CG simulation with known true values. This research provides a valuable technique for accurately tracking player positions globally in dynamic sports settings.

REFERENCES

Beetz, M., Gedikli, S., Bandouch, J., Kirchlechner, B., Hoyningen-Huene, N. von, & Perzylo, A. (2007). Visually tracking football games based on TV broadcasts. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. https://mediatum.ub.tum.de/doc/1289990/document.pdf

Chen, J., & Little, J. J. (2019). Sports Camera Calibration via Synthetic Data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0. https://openaccess.thecvf.com/content_CVPRW_2019/html/CVSports/Chen_Sports_Camera_Calibration_via_Synthetic_Data_CVPRW_2019_paper.html

Distinctive Developments Ltd. (2017). *Baseball Stadiums Pack | 3D Environments | Unity Asset Store*. Baseball Stadiums Pack | 3D Environments | Unity Asset Store. https://assetstore.unity.com/packages/3d/environments/baseball-stadiums-pack-78197

Dunnhofer, M., Sordi, L., & Micheloni, C. (2023). Visualizing Skiers' Trajectories in Monocular Videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5187–5197. https://openaccess.thecvf.com/content/CVPR2023W/CVSports/html/Dunnhofer_Visualizing_Skiers_Trajectories_in_Monocular_Videos_CVPRW_2023_paper.html

Google LLC. (2024). *Google Earth*. Google Earth. https://www.google.com/intl/ja/earth/

Hartley, R., & Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge university press. https://books.google.co.jp/books?hl=ja&lr=lang_ja|lang_en&id=si3R3Pfa98QC&oi=fnd&pg=PR11&dq=multiple+view+geometry+in+computer+vision&ots=aUp1nrbf9P&sig=AumifQDqjW95sIbZidcJgOoSW6o

Homayounfar, N., Fidler, S., & Urtasun, R. (2017). Sports Field Localization via Deep Structured Models. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4012–4020. https://doi.org/10.1109/CVPR.2017.427

Jiang, W., Higuera, J. C. G., Angles, B., Sun, W., Javan, M., & Yi, K. M. (2020). *Optimizing Through Learned Errors for Accurate Sports Field Registration*. 201–210. https://doi.org/10.1109/WACV45572.2020.9093581

Jocher, G., Chaurasia, A., & Qiu, J. (2023). *Ultralytics YOLO* (8.0.0) [Python]. https://github.com/ultralytics/ultralytics (Original work published 2022)

Katić, A., Matić, V., & Papić, V. (2024). Detection and Player Tracking on Videos from SoccerTrack Dataset. *2024 23rd International Symposium INFOTEH-JAHORINA (INFOTEH)*, 1–6. https://doi.org/10.1109/INFOTEH60418.2024.10495998

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, *60*(2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

Lu, W.-L., Ting, J.-A., Little, J. J., & Murphy, K. P. (2013). Learning to track and identify players from broadcast sports videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(7), 1704–1716.

MLB.com. (2024). *Major League Baseball Video Search | MLB Film Room*. MLB.Com. https://www.mlb.com/video

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788. https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html

Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149.

Shi, F., Marchwica, P., Higuera, J. C. G., Jamieson, M., Javan, M., & Siva, P. (2022). Self-supervised shape alignment for sports field registration. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 287–296. http://openaccess.thecvf.com/content/WACV2022/html/Shi_Self-Supervised_Shape_Alignment_for_Sports_Field_Registration_WACV_2022_paper.html

Unity Technologies. (2023a). *Starter Assets—ThirdPerson | Updates in new CharacterController package | Essentials | Unity Asset Store*. https://assetstore.unity.com/packages/essentials/starter-assets-thirdperson-updates-in-new-charactercontroller-pa-196526

Unity Technologies. (2023b). *Unity (Version 2022.3.10f1)*. https://unity.com/

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., & Wang, X. (2022). ByteTrack: Multi-object Tracking by Associating Every Detection Box. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Computer Vision – ECCV 2022* (Vol. 13682, pp. 1–21). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-20047-2_1