Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 14, no. 3, March 2018, pp. 493-501
DOI: 10.23940/ijpe.18.03.p10.493501

3D Scene Recovery based on Multiple Objects Tracking in Sport Videos

Shihe Tian[a], Ming Huang[b], Yang Liu[c], and Chengxin Li[d,*]

[a]*Teaching and Research Department of Physical Education, Capital Normal University, Beijing, 100048, China*
[b]*Harbin No. 5 Middle School, Harbin, 150001, China*
[c]*Harbin Institute of Physical Education, Harbin, 150001, China*
[d]*The Affiliated High School of Harbin Normal University, Harbin, 150001, China*

Abstract

This paper proposes a new method for estimating the player's and ball's 3D position information from monocular broadcast videos. For players, the homography between image and playfield is used to estimate their positions. By analyzing the geometry relation between the ball, its "virtual" shadow and camera position, we derive equations for estimating the flying ball's 3D position. Moreover, we propose a method to predict the flying plane if it cannot be determined from images. This method designs a new cost function, which arrives at the minimum when the predicted flying plane is reasonable. This method has at least two merits. One is that it can estimate the flying ball's position without referring to other objects with known height; the other is that only one assumption is made and the camera is in a fixed position. Experimental results are satisfying.

Keywords: sport video analysis; multiple objects tracking; 3D reconstruction

(Submitted on December 29, 2017; Revised on January 30, 2018; Accepted on February 19, 2018)

1. Introduction

The generation of three-dimensional animation in football videos is not concerned in the computer vision field, which enables audiences to watch highlights from different view angles at different places. The reconstruction of 3D scenes in sport videos includes mainly the estimation of the 3D positions of players and the ball. Lots of methods have been proposed by former researchers as to the estimation of player position. [1] calculates correlative association between two planes with the use of relationship of field marking lines between the field plan and image plane; moreover, such kind of association of all frames in the sequence are computed by Mosaic method. Nam [5] improved on the basis of Choi's work [2] regarding the robust estimation of global motion; however, they didn't mention how the camera was setup and what influence was involved; instead, they took it for granted that all field sizes are constant, which proves unreasonable. Farin [12] focuses on detection of marking lines of the field. Watanabe [15] utilized camera position, angle and zoom level to realize the same function; but, the solution is not as flexible as the former three approaches. Yu [14] proposed to figure out position of players through variation of kick-off circle; as a matter of fact, his proposal has strict limits of application, that is, images of the midcourt line should be straight; otherwise, the result of calibration will go wrong. The major contribution here to the estimation of player position is to elaborately analyze the application scope of the method in [1] and discuss how to compute positions of players in different conditions of camera settings.

Three-dimensional trajectory estimation is important for 3D scene reconstruction. In recent years, many scholars have devoted their attention to the field of soccer video dimension information extraction and scene reconstruction. According to the processing of the video source, the current research can be divided into two categories. One kind is to deal with the special camera data and another class of research is mainly based on broadcast video data [8,13]. The research group led by the British ORWELL [7,13] and Japanese HIDEO [10,11] research group are multi camera based. Because of the multi camera fixed display, there is no camera movement, so it is relatively easy to track the players and the ball. Three-dimensional position information and the use of multiple view geometry knowledge are used to restore the players and the ball. Due to the

* Corresponding author.
E-mail address: chengxinlicxl@sina.com

limitations of the hardware conditions and the large number of broadcast soccer videos that need to be processed, more researchers will broadcast video or single camera data as the research background [2,6].

Yu et al. detected football pitch in every frame; they got football candidates by information like size, color and shape in the pitch to establish so-called candidate feature image. Next, they got the ball's alternative paths through Kalman filter. Finally, they chose the trajectory of the ball by judging alternative paths. Tong et al. detected football by a coarse-to-fine strategy. First, the pitch is detected and the area is clearly not in the field as the ball is removed through shape and color analysis, while the rest area is analyzed by color and ring to get the ball's position. They then use the CONDENSATION algorithm to track the ball. These methods find and track the ball in the video image but cannot get its real position in the world coordinate system. For the reconstruction of the football scene, it is necessary to have the three-dimensional information of the ball, rather than the position of the image. Reid [9] used the 3D position with the ball on the image position and the infinite point light shadow to estimate the position of the ball, which means that the ball detection and shadow are in the image. This work is more difficult, and in some games, there is no shadow. So, the application of this method is limited. This work is similar to that of Kim and Hong's work [3]. They use the corresponding relation between the image plane and the ground plane. The projection position of the ball in the ground plane is calculated, and then the 3D position of similar triangles is used to estimate the ball. This approach faces two difficulties.

To find a similar distance that is perpendicular to the plane of the object field and determine the projection plane of the camera position in the site, we need to know the player's height, which is determined by the position of the ball in the air. For the former, we manually mark two parallel straight lines through the stations of similar players, and through the corresponding relationship between the image and the ground plane obtained in the field on the plane corresponding to line, we determine their point of intersection. The latter assumes that the selected player height is 180cm. From the above analysis, we note that for broadcasting videos, too many assumptions and manual work restrict the application of the 3D position estimating method. Here, we propose to utilize a self-calibration technology of tilt and zoom camera to estimate the intrinsic parameters of the camera and further obtain the position of camera setup from the homography matrix between the field plan and image plane. Next, based on the physical limit of movement in the space, we induced the method for calculating ball's 3D information from camera position and homography. In order to improve the automation, we conducted a simulative experiment, which indicated the proposed method is correct; then, the method was applied for actual video sequences, examining the precision and effect of the algorithm by estimating the known object's height; lastly, 3D animation sequence is produced and that spectators can watch wonderful video clips from different perspectives.

2. Camera model

To restore 3D information regarding competition site, we introduce the imaging process of a pinhole camera and involved parameters. Figure 1 shows the imaging process at one point going across a pinhole camera in space and describes the relationship among world coordinate system, camera coordinate system and image plane coordinate system.

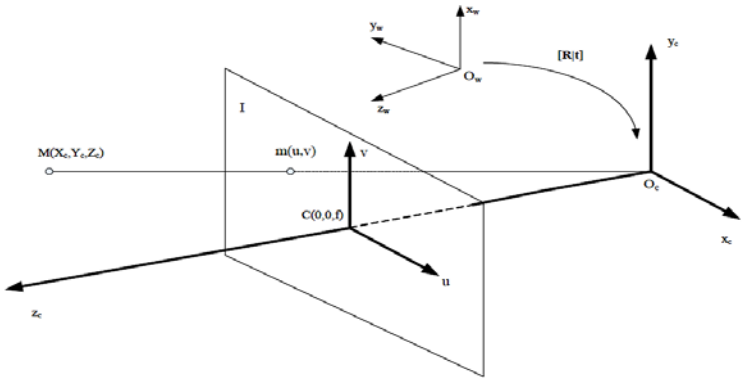


Figure 1. The pinhole camera model

In Figure 1, $O_c x_c y_c z_c$ refers to camera coordinate system, where O_c is the optical center of the camera and also the original point of camera coordinate system where camera main axis is regarded as z_c axis; plane I is image plane of the camera, which is vertical to main axis z_c ; its coordinate is Cuv , where C is coordinate origin and also the point of intersection between principal and image plane, which is called principal point in the paper; the coordinate of principal in

camera coordinate is (0,0,f) . F is the focal length. The coordinate axes v and u on the plane i are parallel to the x_c axis and the y_c axis. The coordinates of a point M in the camera coordinate system is $(x_c y_c z_c)^3$. According to the pinhole perspective projection, we can get point M in the image plane of the image point.

$$u = kf \frac{X_c}{Z_c}$$
$$v = lf \frac{Y_c}{Z_c} \quad (1)$$

We have pointed out that when the camera focal length becomes larger, the object image will be amplified. When the variable is an hour, the image of the object will be reduced, which is what is commonly known as the zoom process. K and L are the reciprocal of the length of each pixel of a photosensitive device in Equation (1).

In Equation (1), the coordinate of one point in space is presented with camera coordinate. Generally, we know only the world coordinate of spatial point; it requires the translation and rotation of one point's world coordinate to that in a camera coordinate system. For the convenience of expression, geometrical position of one point is expressed with a homogeneous coordinate (renamed as photographic coordinate), that is, to expand the coordinate by one dimension and complement it with 1. Now we use $M = (X_w, Y_w, Z_w)$ to represent world coordinate of point $\overline{M} = (X_w, Y_w, Z_w, 1)$; then, its homogeneous coordinate is n and that of the pixel point is $\overline{m} = (u, v, 1)$.

According to Equation (1), the relationship between the space points and the image points can be expressed in a simple matrix. It is shown in Equation (2).

$$\overline{m} = K[R \ \ t]\overline{M} \quad (2)$$

$\approx$ is that both sides of the vector into a ratio of s. K is a 3 * 3 matrix. It is shown in Equation (3).

$$K = \begin{bmatrix} \alpha & \gamma & \mu_0 \\ 0 & \beta & \nu_0 \\ 0 & 0 & 1 \end{bmatrix} \quad (3)$$

We will start from Equation (3) in order to use the background of soccer video, the position of the camera in the shooting location, the position of the players and the position of the ball by the geometric constraints. In particular this article only assumes that the camera is fixed to a certain location in the stadium and can be rotated and scaled. This assumption is reasonable because most of the radio cameras are satisfied with this condition. The extraction of three-dimensional information is carried out in the case of a single view to avoid the other problems caused by multiple cameras.

3. Player position and Homography transform

3.1. Homography transform

A point on the plane and its image has a corresponding relation. In general, the corresponding relation is one to one. Hence, it's an appalled homography conversion.

Figure 2 gives a visual representation of homography. A point on plane II is one-to-one correspondence to image point. When image is confirmed coming from a plane, then the position of one point on the plane II can be determined. Visual perception of such relation is decided by plane II, camera focal distance and camera position. An algebraic derivation of this transformation is given.

Without losing generality, the equation of plane II is denoted as Z = 0. The substitution of Equation (4) is

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \approx K[r_1, r_2, r_3, t] \begin{bmatrix} X_w \\ Y_w \\ 0 \\ 1 \end{bmatrix} \approx K[r_1, r_2, t] \begin{bmatrix} X_w \\ Y_w \\ 1 \end{bmatrix} \tag{4}$$

Given that M is a point on the plane of Z = 0, its coordinates can express two dimensions, $M = [X_w, Y_w]^T$. Equation (5) can be

$$\bar{m} = H\bar{M} \tag{5}$$

Equation (5) expresses the corresponding relation between a point on the plane and its image. The matrix H is called the homography matrix, which is a 3x3 matrix, whose scale is not variable so it includes eight individual components and is determined by at least four corresponding points. As mentioned previously, in general cases, one image point has only one object point corresponding to it; however, when the rule of the object plane being vertical to the image plane is broken, the determinant of H is 0.

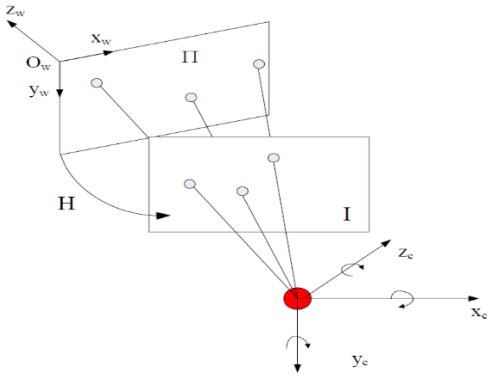


Figure 2. The homography between the points on a plane and their images

3.2. Homography calculation of arbitrary frames in a sequence

If homography transformation between one frame image and pitch is acquired, we can calculate the position of the player on the ground. Next, we discuss how to calculate H matrix with the use of information about image. It is shown in Figure 3.

Figure 3 is a football pitch model. There is a strict position provision for markings on a football playground. The points of intersection can be used to calibrate H. Those points are not strictly defined to highlight the width and length of football field in red color. If we want to use the midfield to demarcate, it needs to firstly mark the length and width of field. We can utilize the corresponding relation between red points and the field model in Figure 4 to calibrate. Yu suggested using cross-ratio invariability to calculate, but the method is applied only when midcourt line is straight in the image; otherwise, it's false. [4] stated the use of the kick-off circle to calibrate, but it needs the use of a tangent line of the kick-off circle for calibration. Since it's very difficult to find the tangent point in the image, the method proves to have less precision; however, it is a good calibration method that makes use of midfield information.

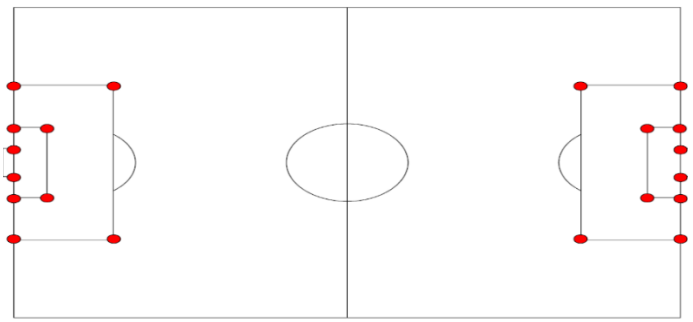


Figure 3. The model of soccer field

Figure 4. An image with sufficient corresponding points for calibration

Figure 5. An image with insufficient corresponding points for calibration

We are faced with a problem of how to estimate a homography matrix that does not contain enough of the corresponding point image. It is shown in Figure 5.

[1] shows how to deal with this situation, as shown in the case of different camera settings.

How to resolve this problem? The following mainly considers two kinds of camera settings:

(1) A rotating camera fixed to a certain position;

(2) A camera that is not subject to any action.

4. Three-Dimensional position estimation of ball

Recovering 3D information from one set of broadcasting video signals is a difficult and interesting issue. It has the following challenges:

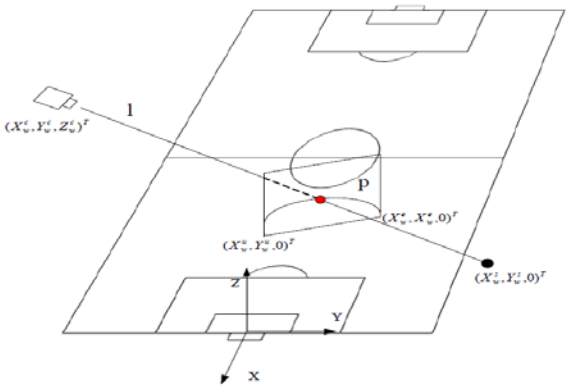
(1) Video streaming captured by fixed camera contains multiple frame information; but, camera position does not change, and it's impossible to form an optical parallax; thus, it's not probable to calculate 3D information of one point by means of a triangle.

(2) Two images can hardly be obtained at the same moment;

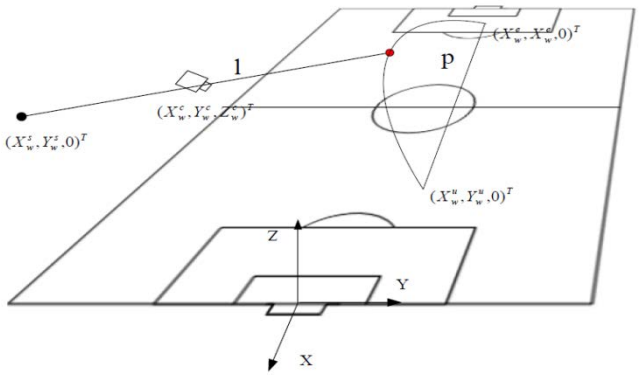
(3) Intrinsic and external parameters of camera need calibration. It's impossible to restore structure, especially European structure, from movement as done in a traditional way. But, with homography relation mentioned above, as well as imaging feature and the ball's flying physical properties, it's likely to restore the 3D position of the ball in the air.

4.1. Theoretical analysis of ball position estimation

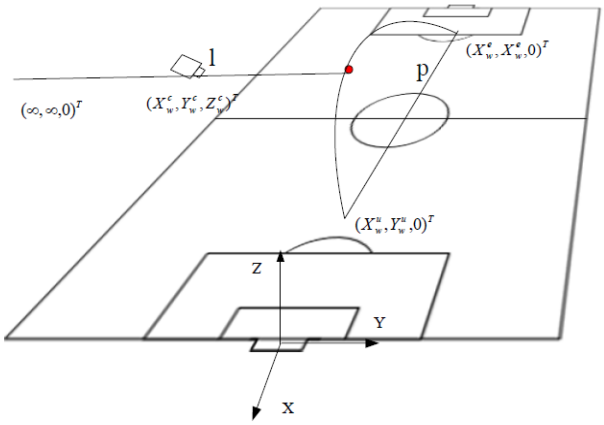
During football broadcasting, the camera used to shoot the main scene is often mounted on the grandstand. For video data coming from TV broadcasting, the priori knowledge of the camera is: the camera will rotate and zoom at a fixed but unknown place. Based on motion characteristics and geometric relation between objects, the 3D position of the ball can be confirmed after we make the following hypotheses: The camera is fixed at one unknown place. The ball's flying path is on a plane. The above hypotheses are rational because: the first point is generally applicable in most cases; then, during football match, due to ball's own rotation, the ball movement curve is not often on a plane but a banana kick. For movement not on a plane, no suitable solution has been sought until now; we can use the plane to approximate it, and the difference is tolerant.



(a) is the case of ball is lower than the camera position.



(b) is the case of ball is higher than the camera position.



(c) is the case of ball and the camera have the same height, in this case the ball's virtual shadow is at infinity.

Figure 6. The geometry relationship among camera position, ball and its virtual shadow

Figure 6 shows the schematic diagram of calculating the position of the ball below the height of the camera. (a) is lower than the height of the camera ball, (b) is higher than the camera height of the ball, and (c) is the height of the ball and the camera's high degree of consistency. For these three kinds of situations, the method proposed in this paper is applicable.

4.2. Algorithm discussion

From Figure 6, we can get results as below:

(1) If the ball is on the field, it's only necessary to know the projection transformation relation between the image plane and field plan; then, we can have the 3D position of the ball (3D coordinate is 0);

(2) If the ball flies to the sky, and if the virtual shadow of the camera, ball, position of plan π are all known, then we can calculate the value of the ball's 3D position through the intersection between straight line l and plan π;

(3) If straight line l is parallel to the field plan, the virtual shadow is infinitely far away; the ball's image point shows on the vanishing line of the field plan; in this case, we can still use Equation (5) to calculate the ball's 3D position because l and π still have an intersected point;

(4) If l is parallel to π in the whole sequence, it's not possible to compute the 3D position of ball;

(5) If the camera lies on the field plan, the virtual shadow can't be restored since the matrix itself is singular.

To sum up, we think in the case of (4) and (5) that it's improbable to restore the ball's 3D position; (5) is impossible and hardly found in reality and (4) is an extreme situation, also barely seen in reality. Through the above analysis, we can know that in order to estimate the three-dimensional position of the ball, we need to calculate the virtual image of the ball, the position of the camera in the world coordinate system, as well as the starting point and end position of the flying ball.

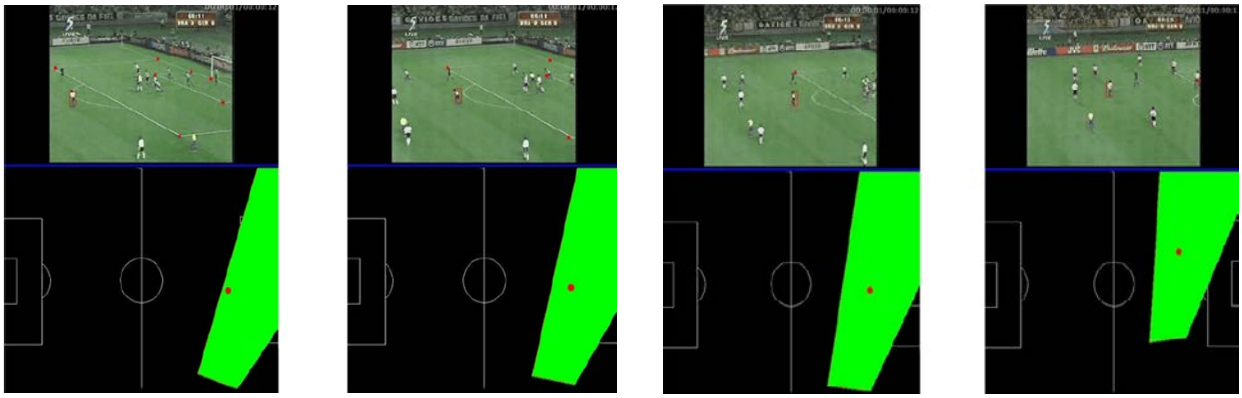
5. Experiment Design and Discussion

5.1. Player position and camera coverage

Now, we examine the player position restoring method with the use of actual video sequences. We give out field area of camera covering every frame image. Then, we get the precision degree with the help of points on the field plan.

Figure 7 lists a group of video sequences of a wonderful fragment. Every image consists of two parts: the first part is the original image of the camera shooting, and the second half is the field model. Since we don't measure the size of the football pitch in the experiment, the camera coverage and player position shown on the field model will deviate somewhat from their actual positions. But, that can be improved through online measuring of the field size. Red points in the video image are intersected points of marked lines on the field. Points in Figure 7(a) are determined by manual calibration. We can input the method proposed by Farin into our system; then, with those points, we can get a homography matrix of one frame image. Red points in the model graph are the positions of boxed players on the field in the image. Next, red points in Figure 7(b), Figure 7(c) and Figure 7(d) are parameters for estimating global movements among adjacent images through global movements. Then, by Equation (5), we calculate the homography matrix of everyone frame image; when correspondence matrices are obtained with the position of field marked line intersections in the world and homography matrix, we can figure out their locations in the image.

It can be found from the graph. The first few frames of the video stream can be better estimated. The image position of these points shows that the method is correct. The corresponding player position obtained by this method is also correct. But when the video stream is very long, due to the accumulation of errors, it will cause a great error in the calculation results. This requires us to find a more stable and accurate image matching algorithm.



(a) (b) (c) (d)

Figure 7. Estimation of player position and the covering range of the camera on playfield

5.2. Synthetic data experiments

During camera calibration, the central position of the image is usually considered as the camera's principal point, which is deflected from the center of image; so, we need to consider the influence on camera position. On the other hand, due to ubiquitous noises, we also need to consider if noises affect the estimation of the camera position. To find out the effect on main point position, we produce data about position of five different principal points. Principal point bias from central position becomes bigger, so we examined five groups of data of three levels. The camera model generates virtual data with 0 as the mean value and σ as noise of standard deviation, which are both added to the virtual data. Not only is the noise added to intra-frame matching point image coordinate, but the marked line intersected image coordinate on the field plan is also used to demarcate camera position. It is shown in Figure 8.

With regards to noise intensity, we investigate from 11 levels: variance growing by 0.2 pixels each step from 0 to 2 pixels. Figure 9(a) displays the impacts of both principal position and noises on camera position estimation. Five curves in the picture correspond to the offset of five different principal points. Each data of 55 coordinates in the figure is the average result after 100 iterations.

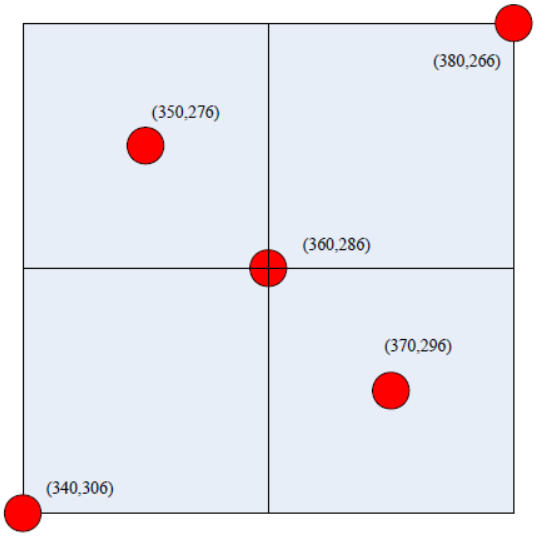
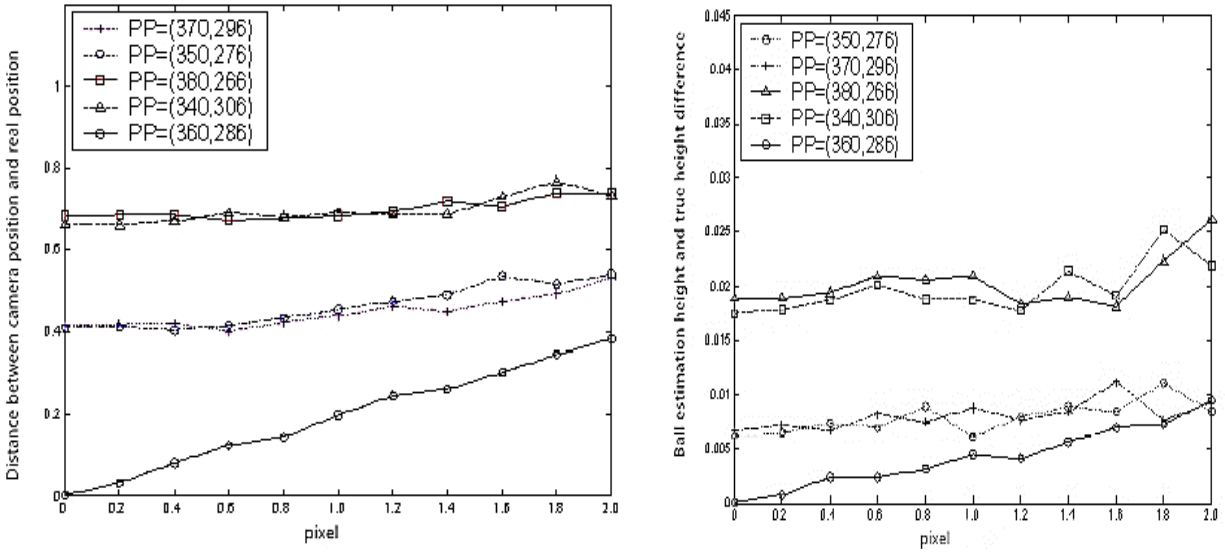


Figure 8. Principal point positions for generating virtual data



(a)the effect on camera position estimation (b) the effect on ball height estimation

Figure 9. The effect of principal point position deviation and noise on estimation

According to the distance from the center of the image, the 5 curves can be roughly divided into three categories, which are the main points of the image center. The greater the degree of deviation is caused by the greater the estimated error. But the size of the venue compared to the estimated error magnitude is relatively small. Now, we examine the effects of noise along with the noise intensity.

The estimation error increases with the increase in noise. For data without the main point offset, the influence of noise is more serious. With an increase in intensity, the estimation error grows rapidly. But, when the offset of the main point is increased, the effect is relatively lower. When the main point offsets reach the maximum, this effect can hardly be seen.

From the point of view of offset and noise, their effects are not very large. The key to the problem is the self-calibration of the camera and the radial distortion of the main camera.

Because of the difficulty in getting an accurate position of the ball, our evaluation algorithm meets troubles. But, ball height is known. By that, we can estimate its height to appraise the algorithm. The upper extreme points of post are manually annotated; the plane of the goal post is acquired by the vertical plane of the ball passing across base lines of the field and frame-to-frame corresponding relation is estimated by the method proposed in the paper. From the figure, we note the estimation of goal post height is close to its true height; but, compared with error from analyses of virtual data, the error based on results from actual videos is bigger.

6. Conclusions

The paper proposed an algorithm to restore 3D information of partial highlights of the European Cup. In that sequence, rubber ball rolls on the ground; then, the offensive player kicks it up off the ground; before the ball falls to the ground, it shoots at the goal. So, the video contains the ball on the ground and in the air; the flying plane of the ball flying in the air in two segments can't be manually calibrated; instead, it needs the flying planar prediction method stated here to foresee the plan position.

References

1. S. Choi, Y. Seo, H. Kim, K. S. Hong, "Where Are the Ball and Players? Soccer Game Analysis with Color Based Tracking and Image Mosaick", *International Conference on Image Analysis and Processing*, pp. 196-203,1997
2. S. Farin, P. Krabbe, "Robust Camera Calibration for Sport Videos Using Court Models", *SPIE Storage and Retrieval Methods and Applications for Multimedia*, pp.80-91,2014
3. H. Kim and K. S. Hong, "Robust Image Mosaicing of Soccer Videos Using Self-Calibration and Line Tracking", *Pattern Analysis & Applications*, vol.4, pp.9-19,2011
4. T. Kim, Y. Seo and K. S. Hong, "Physics-Based 3D Position Analysis of Soccer Ball from Monocular Image Sequence", *The International Conference on Computer Vision*, pp.721-726,1998
5. S. Nam, H. Kim and J. Kim, "Trajectory Estimation Based on Globally Consistent Homography", *International Conference Computer Analysis of Images and Patterns*, pp. 214-221, 2013
6. S. Iwase and H. Saito, "Tracking Soccer Players Based on Homography Among Multiple Views", *VCIP*, pp.283-292, 2013
7. S. Iwase, H. Saito, "Parallel Tracking of All Soccer Players by Integrating Detected Positions in Multiple View Images", *International Conference on Pattern Recognition*, pp.751-754, 2014
8. J. Ren, J. Orwell, G. A. Jones, M. Xu, "A General Framework for 3D Soccer Ball Estimation and Tracking", *IEEE International Conference on Image Processing*, pp.1935-1938, 2014
9. I. Reid, and A. North, "3D Trajectories from a Single Viewpoint Using Shadows", *British Machine Vision Conference*, pp. 863-872, 1998
10. H. Saito, N. Inamoto, S. Iwase, "Sports Scene Analysis and Visualization from Multiple-View Video", "IEEE International Conference on Multimedia & Expo", pp.1395-1398, 2014
11. X. Tong, H. Lu, and Q. Liu, "An Effective and Fast Soccer Ball Detection and Tracking Method", *ICPR*, pp.795-798. 2014
12. T. Watanabe, M. Haseyama, H. Kitajima, "A Soccer Field Tracking Method with Wire Frame Model from TV Images", *International Conference Image Processing*, vol.3, pp. 1633-1636, 2014
13. M. Xu, J. Orwell, G. Jones, "Tracking Football Players with Multiple Cameras", *IEEE International Conference on Image Processing*, 2909-2912, 2014
14. A. Yamada, Y. Shirai, and J. Miura, "Tracking Players and a Ball in Video Image Sequence and Estimating Camera Parameters for 3D Interpretation of Soccer Games", *International Conference on Pattern Recognition*, pp.303-306, 2012
15. X. Yu, "3D Reconstruction and Enrichment of Broadcast Soccer Video", *ACM Multimedia*, pp.260-263, 2014