Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 14, no. 8, August 2018, pp. 1774-1784
DOI: 10.23940/ijpe.18.08.p14.17741784

Object Tracking Method based on 3D Cartoon Animation in Broadcast Soccer Videos

Chunlong Xie[a], Zhiqian Zhang[b], Chunsheng Wang[b], and Zhengqing Liu[c,*]

[a]*Anta Sports Products Limited, Xiamen, 361008, China*
[b]*Sports Department of Harbin Engineering University, Harbin, 150001, China*
[c]*Beijing University of Technology, Beijing, 100048, China*

Abstract

In this paper, a system of broadcasting football video conversion into 3D cartoon animation is designed. When a sports event is broadcasted, multiple cameras are usually deployed around the field. However, at the same time, only one camera's video is available to viewers. Viewers hope to be able to watch the game from other viewpoints. Moreover, after a major sports game, some web portals provide cartoon animations of goal events. However, this is time-consuming and tedious, and only a single viewpoint is provided. Based on the proposed object tracking methods, this paper employs computer vision and computer graphics techniques to design a system that can generate 3D cartoon animations of soccer games. This allows users to watch the game from different viewpoints.

Keywords: moving object tracking; feature selection; broadcast soccer video; 3D cartoon animation

(Submitted on May 9, 2018; Revised on June 30, 2018; Accepted on July 30, 2018)

1. Introduction

As one of the most popular sports activities in the world, football attracts millions of people to watch and get involved in the sport. Each year, up to a thousand football competitions are held around the world [1-2]. A huge number of broadcast football videos are produced from those competitions. In most circumstances, recorded football videos display only one visual angle at a time, although a few cameras are set up around the playfield during the event [3-4]. Some audience members watch from the perspective of the goalkeeper, while others may hope to observe from above when a goal is scored. Additionally, after an important football match ends, some web portals may display moving pictures of scored goals. Such moving images are mostly manual works, which consume time and energy but provide another view angle of the competition. It is practically significant to develop a system that can convert broadcast football videos to 3D cartoon animations on a large scale, reducing the burden on animators and allowing watchers to view the game from many more perspectives [5].

Matsui et al. [6] stated an image integration system with broadcast football videos as input. The system can produce animations that are viewed from the angle of any player. However, in the system, football is not taken into consideration, even though football is the most concerned by audiences. Bebie et al. [7] developed a system named Soccer Man. This system considers a football video segment as input to generate a set of dynamic 3D scenes. Players are modeled with the use of a so-called dynamic texture object, i.e., a quadrangle surrounding the player region. Such a way of representation restrains the possible view range, because when perspective shifts from its original position, the shape of the texture object seems to change. In this system, football is ignored as well. Yu et al. [8] proposed a 3D rebuilding and augmentation system oriented to broadcast football videos [9]. The system can rebuild scenes of the goal region and midfield region. However, the system can provide only the view angle of the main camera. Overall, 3D reconstruction of dynamic scenes using several cameras is of interest to many researchers. For 3D reconstruction of scenes from multiple cameras, plentiful 3D information can be recovered, and the 3D scene after reconstruction looks more real. However, a multi-camera system is high in cost and not easily configured. Compared with videos that involve several cameras, broadcast football videos are more easily obtained.

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

This paper introduces a system named Video2Cartoon. It can convert broadcast football videos to 3D cartoon animations and allow users to watch football games through a virtual camera from any perspective. This system has the following features: it presents a workable method to detect and track players and the football; through global motion estimation, the camera can be calibrated despite very few characteristic points being utilized at the site; the ball's motion trajectory is a parabola; the 3D location of the ball can be calculated through self-calibration of the camera; modeling of players is achieved according to the International Standard for Human Body Modeling H-anim1.1 [10]; in order to augment the visual experience, modelling is executed by extracting image tiles from real images of the playfield and then performing texture mapping; and advertising boards and audience regions are similarly augmented.

2. Three-Dimensional Information Extraction

In general, it is difficult to recover 3D information from monocular videos, since there will be a loss of information when the scene is projected onto the two-dimensional image plane. To make this problem solvable, some prior knowledge is needed. In the football field, the length of the marking line in the forbidden area is known, and the camera can be calibrated with the help of this information. Then, the corresponding positions of the pixel points in the field plane can be calculated in the 3D world coordinate system.

2.1. Detection and Tracking of Players

2.1.1. Detection of Players

The detection of players is realized by the method in [11]. In football videos, players are moving in the field region, and in most cases, the color of the play court is the main color of video frames. Restricted by those conditions, the detection of players is simplified and accelerated. Firstly, use the strategy presented by Jiang et al. [12] to detect and segment the field region; in their method, the Gaussian mixture model (GMM) is employed to achieve modeling of field color. Parameters of GMM are acquired through the expected maximum (EM) algorithm estimation by means of online collected video frame samples. Next, use the region growth algorithm to withdraw connected components in the playfield region; after some noisy regions are filtered, the remaining regions are input to a classifier to determine whether they belong to the player region. This process is shown in Figure 1.

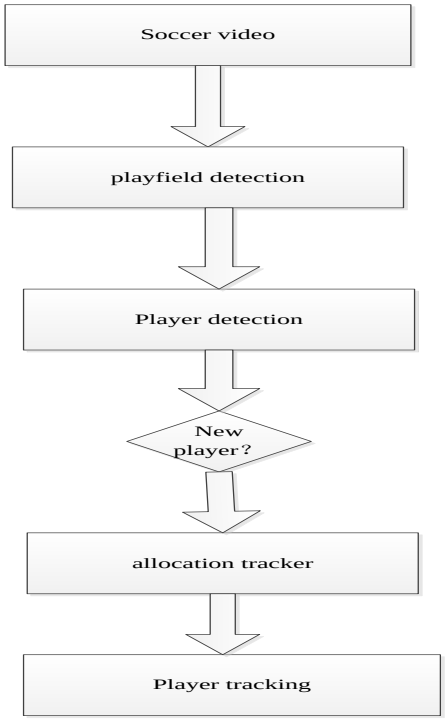


Figure 1. The schematic diagram of multiple player detection and tracking

Support vector machine (SVM) is widely used in the field of pattern recognition, and it shows good performance under the condition of small sample sets. Therefore, SVM is used as the classifier. SVM input is a region of color histogram in

HSV space. Only when the values of S and V are large enough will the color information H be stable.

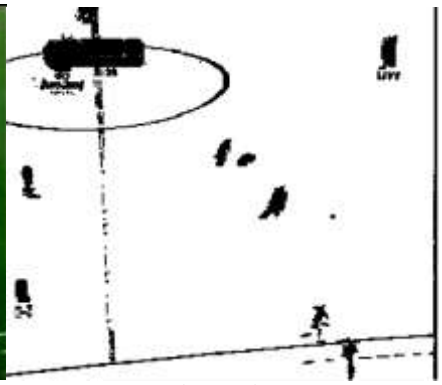
Therefore, only S and V values greater than a threshold pixel will contribute to the HS histogram; the remaining pixels contribute to the V histogram. Figure 2 gives a sample of training support vector machines, where the first line corresponds to the player area sample and the following line corresponds to the non-player area sample. Figure 3 shows the player detection process.

(a) Positive samples

(b) Negative samples

Figure 2. Training samples of SVM for player detection

(a) The original frame



(b) Results after playfield detection

(c) Results after post-processing

(d) Results after support vector classification

Figure 3. Player detection results

2.1.2. Tracking of Players

During the long-term tracking of football, the template matching method based on the Kalman filter is employed. Template matching adopts the operation of grayscale normalization cross-correlation. The Kalman filter is utilized to predict the position of the football in the following frame and rectify the current position of the ball. Template matching is used to obtain the observational value. The Kalman filter and template are initialized by the ball's short time tracking result. The Kalman filter solves the common problem of estimating the status X in the discrete time process. The process conforms to the following linear random difference equation, shown as Equation (1):

$$X_{k+1} = AX_k + w_k \tag{1}$$

The observation value of the system is given by *Z*. This is shown in Equation (2).

$$Z_k = HX_k + v_k \tag{2}$$

A is the state transition matrix of the system, and *H* is the measurement matrix. Random variables w_k and v_k respectively represent the process noise and measurement noise of the system. It is generally assumed that they are independent of each other and are subject to normal distribution. In this paper, the most commonly used first-order dynamic model is shown in Equation (3).

$$x = \begin{bmatrix} x \\ y \\ \dot{x} \\ \dot{y} \end{bmatrix},\ Z = \begin{bmatrix} x \\ y \end{bmatrix},\ A = \begin{bmatrix} 1\ 0\ 1\ 0 \\ 0\ 1\ 0\ 1 \\ 0\ 0\ 1\ 0 \\ 0\ 0\ 0\ 1 \end{bmatrix},\ H = \begin{bmatrix} 1\ 0\ 0\ 0 \\ 0\ 1\ 0\ 0 \end{bmatrix} \tag{3}$$

Here, we take a simple yet effective strategy to settle the ball size changing problem. For matching, a bigger block ($x1-\Delta$, $y1-\Delta$, $x2+\Delta$, $y2+\Delta$) must be generated in the region ($x1$, $y1$, $x2$, $y2$), where, ($x1$, $y2$) and ($x2$, $y2$) refer respectively to the coordinates on the upper left and lower right corner of the matching region. Then, use the method from the ball's short time tracking modules to extract the connected region from ($x1-\Delta$, $y1-\Delta$, $x2+\Delta$, $y2+\Delta$). Follow Equation (2) to judge if the region belongs to the ball's area. If Equation (2) is bigger than a given threshold, the ball is detected and the template needs to update. If the ball is not detected, count the video frames of the ball: if the counting is bigger than a predefined threshold, then reactivate the ball's short time tracking module.

2.2. Calibration of Camera

To obtain 3D information from a 2D image, cameras must be calibrated. With the increasing number of estimated parameters, the technology for camera calibration becomes complicated. In football videos, some a priori knowledge can be applied to simplify the process of camera calibration. First, the playfield is taken as a plane and the projection matrix of 3×4 is reduced to 3×3. Then, according to football game rules, the length of guidelines in the penalty area is known [13-14] and can help restore 3D information. Finally, the position of the main camera is approximately fixed, as this makes it easier to estimate the camera's 3D position. Based on the a priori knowledge introduced above, camera calibration requires the estimation of parameters including homography transformation between field model and field images and 3D position of the main camera.

Given the corresponding points of the four general positions on the model and the image, homography can be estimated directly. Since the length and width of the stadium are unknown, only the intersection of the labeled lines within the restricted area can be used to estimate homography. The positions of these intersections are shown in Figure 4.

2.2.1. Model to Image Homography

Given corresponding points of four general positions to respectively field the model and the image, homography can be directly estimated because the length and width of the playfield is unknown. Only intersected points of guide lines in the penalty area can be used to estimate homography, and the position of those intersected points is shown in Figure 4. Let $M_p = [X, Y]^T$ be a point on the ground plane, then $m_t = [u, v]^T$ represents the point of it.

$\bar{M}_p = [X, Y, 1]^T$ and $\bar{m}_t = [u, v, 1]^T$ are homogeneous coordinates. Therefore, $\bar{M}_p$ and $\bar{m}_t$ have the following relationship:

$$\bar{m}_t = H_t \bar{M}_p \tag{4}$$

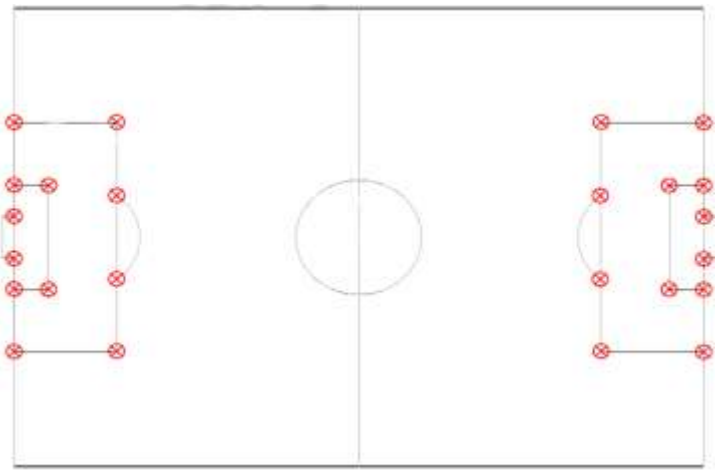


Figure 4. The playfield model of soccer game

If the number of corresponding points is less than 4, the model to image homography can be indirectly estimated, since the camera is fixed and just rotates and zooms. Two images of the pitch plane can likewise be associated through homography. To avoid confusion, we refer to it as homography from image to image, estimated through global motion estimation. To improve the robustness of global motion estimation, player regions are removed during parameter estimation and the detector and tracker of the KLT (Kanade-Lucas-Tomasi) feature point are used to initialize translation parameters for global movement.

Let P_t be the homography transform of *t*-1 video frames to *t*, and the relationships are represented by $\bar{m}_{t-1}$ and $\bar{m}_t$. This is shown in Equation (5).

$$\bar{m}_t = P_t \bar{m}_{t-1} \tag{5}$$

Equations (4) and (5) can be derived from the following Equation (6).

$$H_t = P_t \bar{m}_{t-1} \tag{6}$$

Equation (6) can be further manipulated, as shown in Equation (7).

$$H_t = P_{t1}, P_{t2}, \cdots, P_{t+k} H_{t-k} \tag{7}$$

2.2.2. Camera Position Calibration

Let $M = [X, Y, Z]^T$ represent a point in the scene, then its homogeneous coordinates are $\bar{M} = [X, Y, Z, 1]^T$. The imaging process is described in Equation (8).

$$K = \begin{bmatrix} \alpha & \gamma & \mu_0 \\ 0 & \beta & \nu_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{8}$$

Where s is an arbitrary non-zero scale factor and R and t are called external parameters, corresponding to rotation and translation respectively. They are related to the world coordinate system and the image coordinate system. *K* is called the camera intrinsic matrix. The scale factors α and β correspond to image coordinates on the u- and v-axis, respectively. $[\mu_0, \nu_0]$ are the coordinates of the principal point.

Set $H=[h_1,h_2,h_3]$ and $R=[r_1,r_2,r_3]$. Because Z=0 in the field plane, Equation (9) can be deduced from Equations (4) and (8).

$$K[r_1,r_2,r_3]=\lambda[h_1,h_2,h_3] \tag{9}$$

Equation (10) can be obtained from Equation (9).

$$r_1=\lambda K^{-1}h_1,\ r_2=\lambda K^{-1}h_2,\ r_3=\lambda K^{-1}h_3 \tag{10}$$

Because the main camera is seen as the only rotation and zoom operation, to calibrate the intrinsic matrix of K by the method described in [15-16], assume λ is zero and estimate α and β. They can be obtained by solving Equation (11).

$$K=\begin{bmatrix} p_{11}\ p_{21} & p_{12}\ p_{22} \\ p_{11}\ p_{31} & p_{12}\ p_{32} \\ p_{11}\ p_{31} & p_{22}\ p_{32} \end{bmatrix}\begin{bmatrix}\alpha^2 \\ \beta^2\end{bmatrix}=\begin{bmatrix}-p_{13}\ p_{23} \\ -p_{13}\ p_{33} \\ -p_{23}\ p_{33}\end{bmatrix} \tag{11}$$

3. 3D Position Estimation of Players and Football

3.1. 3D Position Estimation of Players

3D position estimation of players is rather direct, because in most circumstances, players are on the pitch plane. When a player jumps from the ground, it is difficult to estimate the player's 3D position from monocular videos. Let $\bar{m}_t$ represent the position of a player on the image plane, then the position $\bar{M}_p$ on the pitch plane is obtained by following Equation (12).

$$\bar{M}_p=H_t^{-1}\bar{m}_t \tag{12}$$

Generally speaking, it is a difficult task to estimate the 3D position of a football from monocular videos. However, if the motion trajectory of a football is a parabola, meaning the football moves on a plane that is vertical to the pitch plane, then the 3D position of the ball can be estimated. The 3D position of the football is determined using straight line 1 and plane π: straight line 1 is determined by camera position C and shadow position S of the football on the pitch plane, while π is determined by A and B, which represent the intersection point of the football trail and the plane of the play court. When the image positions of A, B, and S are given, they can be reached by Equation (1). Note that the image position of the shadow S is the image position of the football. When A or B cannot be acquired in certain conditions, e.g. the ball is kicked before it comes in contact with the pitch plane ground, π can be obtained by searching the plane of the parabola that best approximates the ball's trajectory.

3.2. 3D Position Estimation of Football

In general, it is difficult to estimate the 3D position of the ball from monocular videos. However, if we assume that the trajectory of the ball is a parabola, that is, the ball and the field plane are perpendicular to the plane of motion, then the three-dimensional position of the ball can be estimated. This is shown in Figure 5.

The 3D position of the ball is the intersection decision of l line and π. Linear l is the decision of the position of the camera C and the ball in the field plane of the virtual shadow position S. π is determined by A and B, which are the intersection of the trajectory of the ball and the pitch plane. When a given A, B, and S correspond to the image position, they can be calculated using Equation (12). Note that the image position of the virtual shadow S is the image position of the ball. When A or B cannot be obtained under certain conditions, π can be obtained by searching for a sphere with the trajectory closest to the parabola.

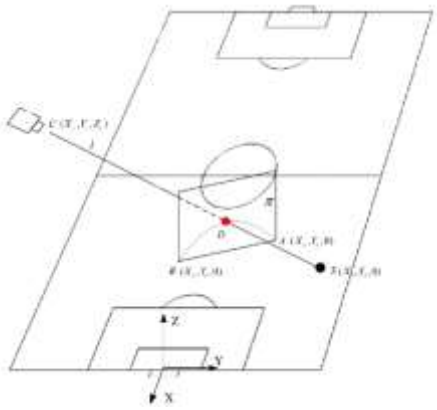


Figure 5. The geometrical relationship used for 3D position estimation of the ball

4. Generation of 3D Cartoon Animation

4.1. Field Modeling

Modeling of football pitch is very direct, because according to the rules of football games by the Federation Internationale de Football Association (FIFA), most parameters about playfield are known except for length and width, which can be estimated through homography conversion through online model-to-image and image-to-image. For an augmented visual experience, execute the modeling of playfield by extracting image tiles from real images and then performing texture mapping. Similarly, advertising boards and audience regions are augmented. Figure 6 shows several different perspectives through the course of the simulation using computer graphics technology.

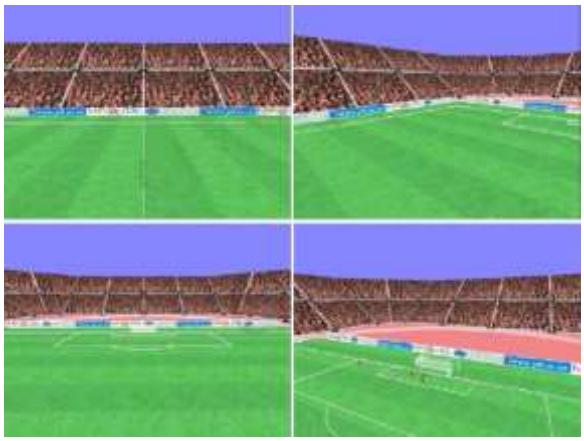


Figure 6. Several different views of the playfield

4.2. Player Modeling

Modeling of players is done according to the International Standard for Human Body Modeling H-anim1.1. Almost all types of actions of a player can be simulated. Furthermore, the model allows the colors of players' uniforms, socks, and shoes to be changed. Hence, users can choose to personalize the modeling of players based on their preferences. Figure 7 shows some of the different perspectives of players who have been modeled using graphics technology.

4.3. Generation of 3D Cartoon Animations

The generation of 3D cartoon animations includes the animation generation of the ball and players. Before the generation of animation, the trail of the football and players are smoothly processed using a Gaussian filter to eliminate noises. Compared with the players' animation, the ball's animation is generated more easily. A white globe is used to model the football, with its center on the 3D trajectory of the football. The players' animation is not so easily generated, because it is pretty difficult

to restore 3D articular movement information from monocular videos. Besides, the player region in long shot of low resolution raises challenges to solving this problem. The existing implemented system attempts to obtain a player's motion type based on the analysis of the player's path, such as running, walking, and stopping. Firstly, use a motion capture device to construct the player's motion pattern library, including running and walking; then, according to fixed length, e.g., one second, divide the player's trajectory into numerous non-overlapping segments. For each segment, calculate the average speed of each type of motion. If it is bigger than a given threshold, the motion type is running, and if it is smaller than a given threshold, the motion type is stopped. Otherwise, the motion type is walking. Lastly, for each segment, retrieve from the database the relative movement data to drive the player model. The moving direction of the player is determined by the direction of its velocity.

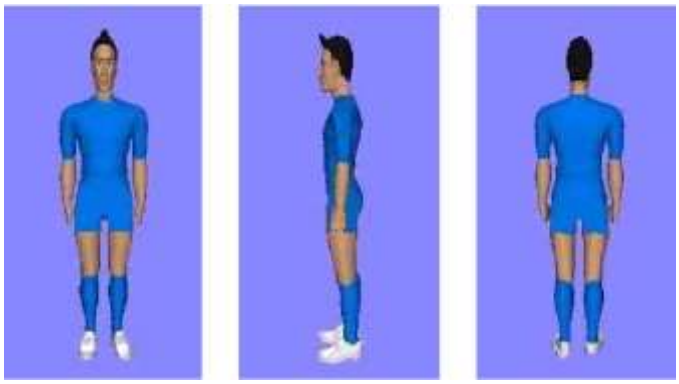


Figure 7. Several different views of a player

5. Experiment Design and Discussion

In this paper, the visual system that is implemented is Video2Cartoon in C++ 6.0 and OpenGL. This system allows the user to rotate and scale the operation by controlling the virtual camera as well as change the angle of view. In addition, by using the direction key and the mouse, the user can roam on the field plane. The video data are from the 2014 European Cup TV, using the standard MPEG-II compression. The frame rate is 25 frames per second, and the resolution is 352×288.

5.1. Player Detection and Tracking

The experiment on player detection and tracking is conducted with five segments of long-shot videos. Table 1 lists experimental results. The five video segments contain a total of 1405 frames. Since the SVM detector and particle filtering tracker perform well, the recall ratio is very high. Therefore, only the recall ratio is given here. From Table 1, we see that the overall recall ratio reaches up to 90.4%. In some cases, players are hardly detected and tracked, due to either of the following reasons:

(1) Player areas are so small that they are screened out by being mistaken as noises.

(2) Player areas are close to/overlapping with words, station captions, and guidelines, so they are removed during pre-treatment after the detection of football pitch.

Table 1. Results of player detection and tracking

Video clip	Video frames	Number of players	Detect and track the number of players	Recall (%)
pclip1	399	1232	1146	93
Pclip2	271	795	713	89.7
Pclip3	315	1241	1053	84.9
Pclip4	230	763	694	91
Pclip5	190	644	619	96.1
Total	1405	4675	4225	90.4

5.2. Football Detection and Tracking

To evaluate the effectiveness of the proposed algorithm for detecting and tracking the leather ball, here we choose to test

two sets of typical videos that contain a total of 1369 frames. In one segment, the pitch apparently looks worse; its color is far from green and close to the soil color, and one part of the pitch is shining under the sun while the other lies in the shadow of the stadium. In the other video segment, the pitch color is green and no shadow found in the pitch. We mark the true value by hands. Suppose one video frame contains the football, which can still be found if human eyes do not depend on the information in the fore-and-aft frames. This is shown in Table 2.

Table 2. Results of ball detection and tracking

Video clip	Video frames	Number of players	Detect and track the number of players	Number of false check	Precision (%)	Recall (%)
bclip1	650	600	460	53	89.7	76.7
bclip2	719	631	533	65	89.1	84.5
total	1369	1231	993	118	89.4	80.7

The precision ratios of the two video segments are close, but the first video's recall ratio is lower than that of the second one because terrible illumination makes it harder to detect the ball. Moreover, the football is sheltered by players or overlaps with marking lines on the pitch. The virtual shadow appears because the football is completely sheltered by players or overlaps with marking lines. On this regard, players' socks or marking lines are visually more similar to the football.

5.3. 3D Position Estimation

The height of the goal defined according to FIFA's rules is known. Here, we use it to evaluate the effectiveness of 3D position estimation. The plane employed to estimate the 3D position is determined by two goal posts. Straight lines taken to estimate the 3D position are decided by the 3D position of the camera and the virtual shadow position of the goalpost top. The height of the goal, which is acquired by estimation from three video segments, is given in Figure 8. In the picture, the red straight line represents the true value (2.44m) of goal the height. Only the matrix from the first frame to image homography is obtained by estimating intersected points of marking lines. The homography matrix of the following video frames is obtained by estimating global motion. The difference between the estimated value and the true value arises out of the error accumulation of global motion estimation, the quantization error of video frames, and the inaccuracy of the pinhole camera model. However, for the generation of cartoon animations, the proposed 3D position estimation method in this paper is acceptable.

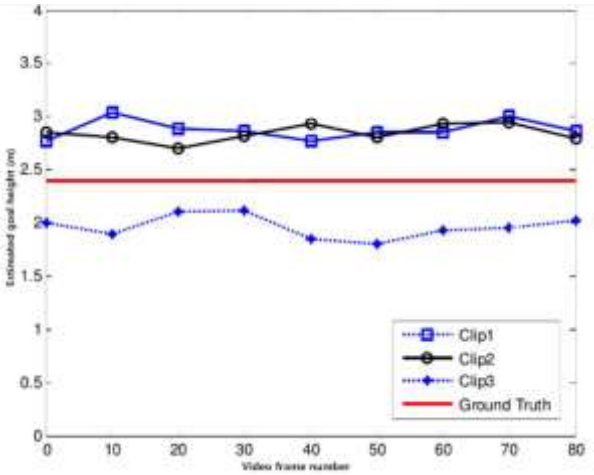


Figure 8. The estimated height of the goalpost on three video clips

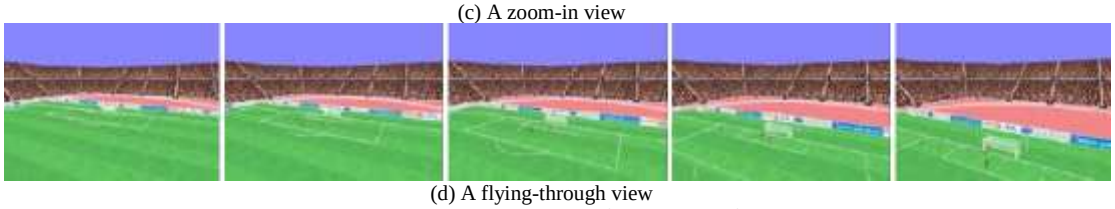
5.4. 3D Cartoon Animation Generation

In this paper, we choose five segments of football video clips for our experiment. Here is the result of a 3D cartoon animation displaying a goal. The goal came from a television broadcast of the 2012 European Cup in Portugal and Russia. Some representative video frames are shown in Figure 9.

(a) The original soccer video

(b) A view from the right side of the playfield

(c) A zoom-in view



(d) A flying-through view

Figure 9. 3D cartoon animation of a video clip of a goal

6. Conclusion

This paper presents a system that converts the broadcast soccer videos into three-dimensional cartoon animations. With the help of computer vision and computer graphics technology, the system allows users to change the perspective, allowing them to use the direction keys and mouse to roam the field plane. This system has a number of potential applications. Content service providers can use the system to generate a score from any visual angle of the cartoon animation and release this information on their website or pass it to mobile device users.

References

1. D. Qin, "Study on the Algorithm for Automatic Classification of Video Content and Multi Feature Combination based on SVM," *Shanghai Jiao Tong University*, 2009
2. S. Yang, "Research on Digital Media and Pan Animation," *Yunnan Normal University*, 2014
3. Y. Lu, " Research on Video Shot Detection and Classification based on Content," *Shandong Normal University*, 2010
4. R. R. Wang, "Generation of 3D Character Animation based on Motion Database," *Graduate University of Chinese Academy of Sciences*, 2006
5. J. Tan and L. D. Wu, "Video Summarization Method based on Feature Animation," *Computer Application*, Vol. 10, pp. 3960-3962, 2011
6. K. Matsui, M. Iwase, M. Agata, T. Tanaka, and N. Ohnishi, "Soccer Image Sequence Computed by a Virtual Camera," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 860-865, 1998
7. T. Bebie and H. Bieri, "Soccer Man-Reconstructing Soccer Games from Video Sequences," in *Proceedings of IEEE International Conference on Image Processing*, Vol. 1, pp. 898-902, 1998
8. X. Yu, X. Yan, T. S. Hay, and H. W. Leong, "3D Reconstruction and Enrichment of Broadcast Soccer Video," in *Proceedings of ACM International Conference on Multimedia*, pp. 260-263, 2014
9. T. Kanade and P. J. Narayanan, "Virtualized Reality: Perspectives on 4D Digitization of Dynamic Events," *IEEE Computer Graphics and Applications*, Vol. 27, No. 3, pp. 32-40, 2012
10. Humanoid Animation Working Group, "Specification for a Standard Humanoid Version 1.1," (http://h-anim.org/Specifications/H-Anim1.1/)
11. G. Zhu, C. Xu, Q. Huang, and W. Gao, "Automatic Multi-Player Detection and Tracking in Broadcast Sports Video using Support Vector Machine and Particle Filter," in *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 1629 -1632, 2013
12. S. Jiang, Q. Ye, W. Gao, and T. Huang, "A New Method to Segment Playfield and its Applications in Match Analysis in Sports Video," in *Proceedings of ACM International Conference on Multimedia*, pp. 292-295, 2014
13. R. Hartly and A. Zisserman, "Multiple View Geometry in Computer Vision," 2nd Edition, Cambridge University Press, 2013

14. Z. Zhang, "A Flexible New Technique for Camera Calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 11, pp. 1330-1334, 2012
15. P. Pérez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based Probabilistic Tracking," in *Proceedings of European Conference on Computer Vision*, pp. 661-675, 2012
16. V. Vapnik, "The Nature of Statistical Learning Theory," Springer-Verlag, New York, 1995

Chunlong Xie received his B.S degree from Harbin Sport University and his M.S degree from Beijing Sport University. His research focuses on sports humanistic sociology.

Zhiqian Zhang received his M.S degree from Harbin Engineering University. He is an associate professor at Harbin Engineering University. His research interests include physical education and training.

Chunsheng Wang received his M.S degree from Harbin Engineering University. He is a lecturer at Harbin Engineering University. His research interests include physical education and training.

Zhenqing Liu received his M.S degree from Beijing Sport University. He is an associate professor at Beijing University of Technology. His research interests include sports education and management.