Characteristics Extraction of Behavior of Multiplayers in Video Football Game

Zhiwen Wang, Hao Ouyang, Canlong Zhang, Bowen Tang, Zhenghuan Hu, Xinliang Cao, Jing Feng, and Min Zha

Abstract In the process of behavior recognition of multiplayers for soccer game video, various features of athletes need to be extracted. In this paper, color moments extracted by using color classification learning set are regarded as color feature. Contour features of athletes are extracted by utilizing players silhouettes block extraction and normalization. Hough transform is used to extract the features of coordinates of pitch line, which can be used for camera calibration, rebuilding the stadium, and calculating the coordinate of players in the real scene. The trajectories of players and ball are predicted by using Kalman filter, while trajectories characteristics of player and ball are extracted by using the trajectory growth method. Temporal and spatial interest points are extracted in this paper. Experimental results show that the accuracy of behavior recognition can be greatly improved when these features extracted are used to recognize athlete behavior.

Keywords Behavior recognition · Feature extraction · Color moment · Hough transform · Trajectory growth method · Temporal and spatial interest points

1 Introduction

In order to detect whether each pixel represents a feature, the extraction of behavior characteristics of multiplayers in soccer video is a primary operation in the process

Z. Wang (✉) · H. Ouyang
College of Computer Science and Communication Engineering, Guangxi University of Science and Technology, LiuZhou, Guangxi Province, People's Republic of China

C. Zhang
College of Computer Science and Information Technology, Guangxi Normal University, Guilin, Guangxi Province, People's Republic of China

B. Tang · Z. Hu · X. Cao · J. Feng · M. Zha
College of Electrical and Information Engineering, Guangxi University of Science and Technology, Liuzhou, Guangxi Province, People's Republic of China

E. T. Quinto et al. (eds.), *The Proceedings of the International Conference on Sensing and Imaging, 2018*, Lecture Notes in Electrical Engineering 606,
https://doi.org/10.1007/978-3-030-30825-4_11

of behavior recognition [1]. The result of feature extraction is that the different subsets divided by the points on the image, which often belong to isolated points, continuous curves, or contiguous regions. Usually, there are many features that can be used for feature recognition behavior in soccer video game, such as the color feature of the stadium, ball, athletes and referees, the moving trajectory (spatial relationship) features of ball and athletes, the contour feature of players, the shape features of ball pitch line features, etc.

2 Related Work

On the feature extraction of athletes, in recent years, many researchers have carried out extensive research on it. Among them, the color features of the clothing of athletes, the contours features of the athletes, and the characteristics of the field line are extracted. In the past 10 years, the motion analysis based on contour features has developed into an important feature extraction technique. Recently, a lot of research has been done in this area. In [2, 3], novel approaches for indexing and evaluating sketch-based systems were proposed. A recent work in sports analytics [4] uses sketch-based search techniques to analyze player movements in rugby matches. It uses multiple distance-based similarity measures, which compares the user sketch against video scenes. Beside the traditional image and video search, sketch-based search techniques can also be applied to other complex data sets. In [5, 6], 2D sketches are used to search for 3D objects. Lee et al. [7] make use of synthesized background shadows to help users in sketching objects. Further applications of sketch-based search include retrieval in graphs [8], or for bivariate data patterns in scatter plots [9, 10]. Regarding data generation, in [11] 2D shape sketches are used to create and manipulate high-dimensional data spaces. A novel model-based technique that is capable of generating personalized full-body 3D avatars from orthogonal photographs [12]. The proposed method utilizes a statistical model of human 3D shape and a multiview statistical 2D shape model of its corresponding silhouettes.

3 Characteristics Extraction of Behavior of Multiplayers

Color Feature Extraction of Players and Referees

Since soccer is an ornamental sport, the playing field, ball pitch line, the clothing of referee and players are designed to have unique visual effects. From visual features, these color differences are one of the best messages [13–16]. The field is green and the field lines are white. Judges and players must wear as high a contrast as possible. Color characteristics can be used not only to improve tracking ability, but

also to distinguish players which belong to different teams. Therefore, the color feature of video images can be extracted for behavior recognition [2]. In this paper, color classification learning sets is used to find interesting regions by mapping the image pixels to their color classes, and then morphological operators is used to group pixels. The mixed color space is used to detect the best distinction of the space produced by the pixels of the opponent team and the referee [17, 18].

Color Classification, Segmentation, and Feature Extraction

The first step in dealing with video images of a soccer game is to apply color classification and segmentation to the image. Assume that the previous color classification learning set includes the green field set, clothing set of team, and other set, color image pixels are firstly mapped to the respective classification in visual perception module, and then morphological operations is used to find the region of interest by denoising of the same type of color grouping region [3, 4].

In addition, the object of interest can be characterized by the characteristics of the image patch. In particular, the assumption of using the player's upright position to contact the pitch plane (if considered in a particularly long image sequence, at least) [19, 20]. The size of every patch can be estimated with much accuracy by using image patch coordinates. In addition, the interest objects must meet certain compact relations (the ratio between area and perimeter), such as the players and soccer. These assumptions can be used to filter data and more reliably extract related objects.

Existing color regions and color patches can be used to define the interest complex regions that are handled by certain image processing operations. For example, in order to find the line of field, the green centralized area in the field image (not including the player or referee occlusion area) are considered to get the field line. This area can be represented by using Eq. (1).

$$\mathrm{L} = \left(\theta \cap \complement\right) - \alpha 1 - \alpha 2 - \alpha 3 \tag{1}$$

where, L is the area of the field line, θ is non-green area, $\complement$ is playfield detection, $\alpha 1$ is the area where the team 1 is located, $\alpha 2$ is the area where the team 2 is located, and $\alpha 3$ is the area where the referee is located.

After obtaining the color regions of interest, the color moments of the region can be extracted as color features. In the HIS color space, the central moment (the first three order color moments) of each component can be calculated by Eq. (2).

$$\begin{cases} M_1 = \frac{1}{N}\sum_{i=1}^{N} X(p_i) \\ M_2 = \left[\frac{1}{N}\sum_{i=1}^{N}(X(p_i) - M_1)^2\right]^{1/2} \\ M_3 = \left[\frac{1}{N}\sum_{i=1}^{N}(X(p_i) - M_1)^3\right]^{1/3} \end{cases} \tag{2}$$

Among them, X represents the H, I and S components in the HIS color space. $H(p_i)$ represents the X value of the ith pixel of the image p. N is the number of pixels in the image. Figure 1 is the image of the RGB distribution, the HIS color space, H, I, and S components and histogram. The calculated center moment of image and seven characteristic values are 1.0e + 004 * [1.5278 0.000 0.0001 0.0000 0.0004 0.0002 0.0003 0.0009 0.0005 0.0007].

Robustness of Color Classification

Lighting conditions change when the camera sweeps from one side of the court to another, clouds change, rain begins, and so on. For these reasons, reliable color segmentation cannot be achieved by prior learning of color categories and remains unchanged during the game [21–26]. On the contrary, color segmentation must adapt to the change of the sky, especially the segmentation of green sites in color categories. Therefore, the expectation maximization method is used to segment the green sites in color classes. By giving the stadium model and estimating camera parameters, it is determined that these areas must be green. The related area is expressed by Eq. (3).

$$\mathrm{A_R} = A_\mathrm{P} - A_\mathrm{T1} - A_\mathrm{T2} - A_\mathrm{Re} - A_\mathrm{L} \tag{3}$$

where, A_R is the related area, A_P is the playfield detection, A_T1 and A_T2 represent the area in which the team 1 and team 2 is located respectively, A_Re is the region of the referee, and A_L is the area where the pitch of a course line is located. Morphological operations are used to eliminate holes in the region in this paper. Then, the pixels in these regions are extracted to estimate the color classes of green regions. Finally, we use this color model to estimate camera parameters. In practice, the estimation of the classification model is much lower than that of the camera parameter estimation.

Contour Feature Extractions of Players

Given a video $v = \{I_1, I_2, \cdots, I_T\}$ which contains T frames of soccer game behavior, the related behavior contour sequence $S_s = \{s_1, s_2, \cdots, s_T\}$ can be obtained from

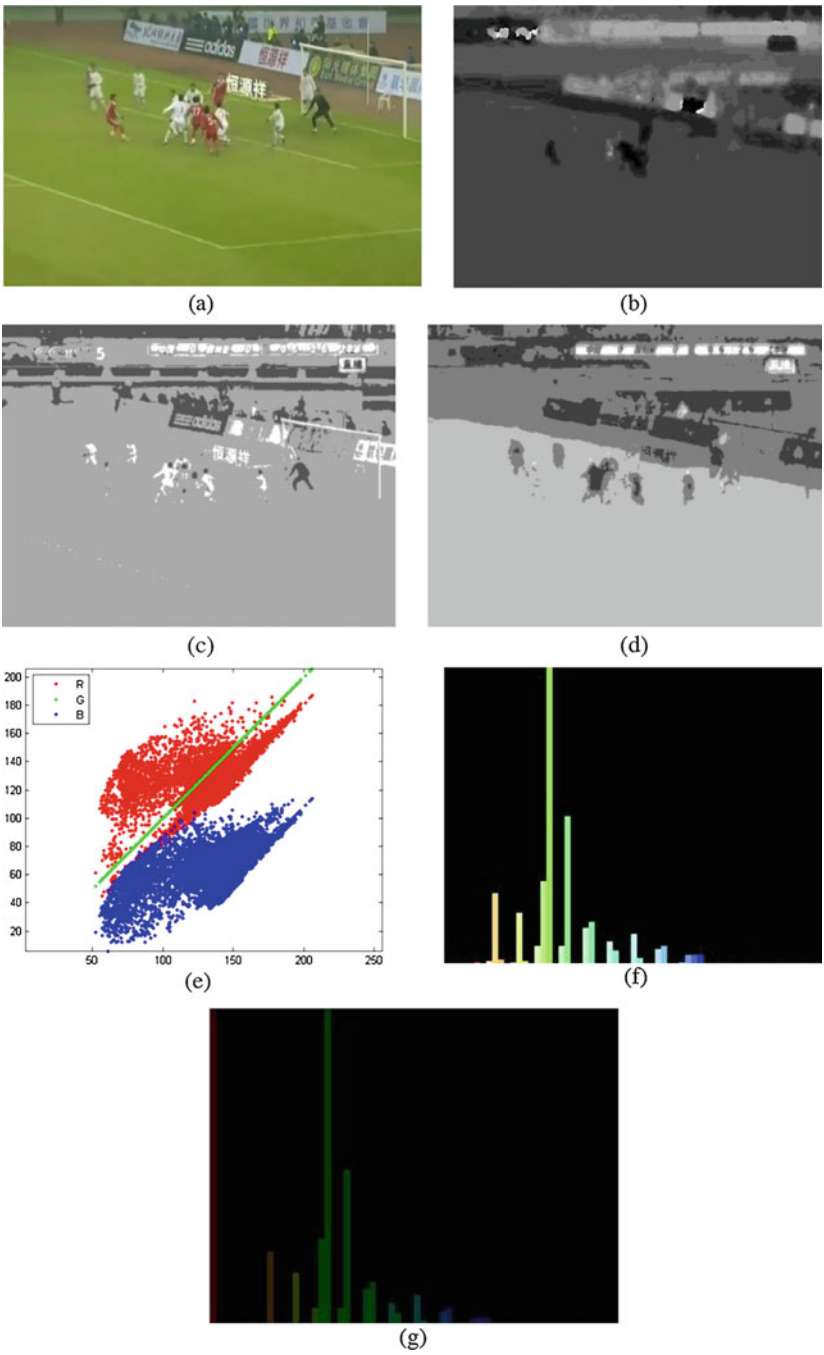


Fig. 1 Color component and histogram of video football match image. (**a**) Original image, (**b**) *H* component of image, (**c**) *I* component of image, (**d**) *S* component of image, (**e**) Color distribution of image, (**f**) *H*-*S* histogram of image, (**g**) *H*-*I* histogram of image

Fig. 2 Contour sequence and block feature representation of a player

the original video. The size and position of the foreground region changes with the distance between the player and the camera, the size of the target, and the behavior that have been completed changes. On the basis of keeping the contour ratio of the player unchanged, the contour image of the player is centralized and normalized, so that the resulting image $\mathrm{RI} = \{R_1, R_2, \cdots, R_T\}$ contains as much foreground as possible. All input video frames have the same dimension $r_i \times c_i$ without making the motion deform. The normalized motion contour image of player is shown in Fig. 2. If the original contour image R_i of the player is represented by the vector r_i in $\Re^{r_i \times c_i}$ space by using line scanning mode, the outline of the player in the whole football game will be represented as $v_r = \{r_1, r_2, \cdots, r_T\}$.

In order to improve the computational efficiency, the contour image of each player is equidimensional divided into $h \times w$ sub blocks without overlap in this paper [5]. Then, the normalized values of each sub block are calculated by using Eq. (4).

$$N_i = b(i)/mv, \quad i = 1, 2, \cdots, h \times w \tag{4}$$

Among them, $b(i)$ is the number of foreground pixels of the ith sub block and mv is the maximum of all $b(i)$. In $\Re^{h \times w}$ space, the descriptor of the silhouette of a player in the tth frame is $f_t = [N_1, N_2, \cdots, N_{h \times w}]^T$ and the outline of the player in the whole video can be correspondingly represented as $vf = \{f_1, f_2, \cdots, f_T\}$. In fact, the original player contour representation v_r can be considered as a special case based on block features, that is, the size of sub block is 1×1 (one pixel).

Feature Extractions for Stadium Line

The field line contains important coordinate information of the stadium. The results of line extraction can be directly used for camera calibration, stadium reconstruction, and calculation of the coordinates of players in the real scene. In the process of extracting court line feature, the video image of ball game is translated into binary image, then the feature of coordinate parameters of court line is extracted preliminarily using Hough transform, and at last the accurate linear coordinates is computed by using gray fitting. When using Hough transform to extract line feature, distance vector which can be calculated by Eq. (5) can be used to represent.

$$d = x \cos\theta + y \sin\theta \tag{5}$$

Among them, the value range of d is the diagonal length l of the video image, that is $d \in [-l, l]$, θ is the intersection angle between the vertical axis y and the level axis x and $\theta \in [0, \pi]$. x and y represent the two-dimensional coordinates of image pixels. We define the parameter space with the subscript d and θ using an array of integers k, and set the threshold th. When we use Hough transform to run statistical calculation, if $k > th$, the curve with the subscript d and θ is determined a straight line. Because the calculated value d may be negative, the subscript d is rewritten as $d + l$. The specific steps are as follows [6].

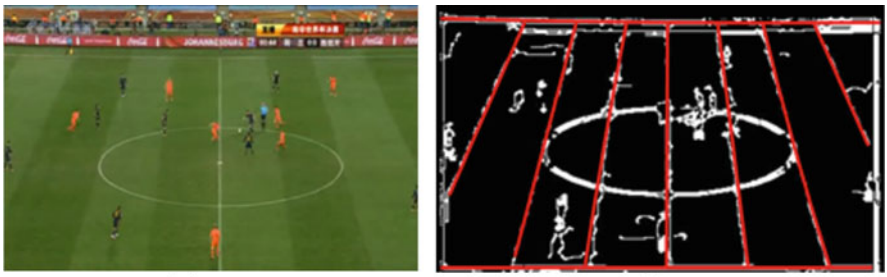
Step 1: construction and initialization of the lookup table of function *sin* and *cos*.

Step 2: for each non-background point (white dot) of a binary image, each value of d corresponding θ is computed by using Eq. (4) and the values of $d + l$ and $k + 1$ are meanwhile computed.

Step 3: when the values are greater than th, all subscripts d and θ corresponding to array k are found out, and then $d - l$ is calculated.

Step 4: because of the field line has not been refined in the binary image to be detected, there is a certain width. In the case of Hough transformation, it leads to several groups of similar d and θ for a straight line at the same time. If there is similar d and θ, only one group is retained.

Step 5: the coordinates of the two ends of the field line obtained by Hough transformation are (x_0, y_0) and (x_1, y_1).The mean gray value $Mean_{i,j}$ of the pixels passing through the line between the point i and the point j in the gray image are calculated by using $((x_0 + \sin\theta(f - m \times i), y_0 - \cos\theta(f - m \times i))$ and $((x_1 + \sin\theta(f - m \times j), y_1 - \cos\theta(f - m \times j))$, respectively. Where, f expresses the range of fitting and m is expressed as the fitting step length, $i, j \in \left[0, \frac{2f}{m}\right]$. When the calculated $\text{Mean}_{i,j}$ is maximum, the determined line segment $i^{'}, j^{'}$ corresponding to maximum $\text{Mean}_{i,j}$ can be considered as the optimal course line. Figure 3 is the characteristic of the field line determined by using Hough transform.



(a) (b)

Fig. 3 Features extraction of stadium line. (**a**) Original image, (**b**) Extracted field line

Feature Extractions for Motion Trajectory of Player and Ball

In the process of feature extraction for motion trajectory of players and ball, the video game is divided into small segments containing a specific number of video frames. When we extract the motion trajectory of players and ball, the segment is regarded as the basic unit, that is, the length of the processed trajectory is not more than the number of frames in the video fragment. After obtaining the candidate area of the motion target of each frame of soccer game video, first, we locate the moving objects in the three consecutive frames of the video in a spatial-temporal domain. The position of second frames in three consecutive frames is centered, and the candidate areas falling near the position are found in the front and back frames. After finding such a continuous three frame image, determine whether the moving target is included in the existing trajectory. If there is no, the new trajectory is initialized with the moving target in the three consecutive frame, and the location of each trajectory is recorded. After getting the new trajectory, the Kalman filter is used to predict the trajectory. The prediction Eq. (6) is used to predict new trajectory.

$$\begin{cases} X_t = AX_{t-1} + \gamma_t \\ O_t = BX_t + \kappa_t \end{cases} \tag{6}$$

Among them, $X_t = AX_{t-1} + \gamma_t$ is the motion equation of the system, $O_t = BX_t + \kappa_t$ is the observation equation of the system, X_t and O_t is system state vector and system state measurement vector of time *t,* respectively, γ_t and κ_t is the vector of the motion which is the normal distribution and measurement noise, and they are mutually independent. A and B express the state transfer matrix and the measurement matrix. The center position of the moving target is chosen as the measurement vector of the state of system, and the center position of the moving target and its motion speed are regarded as state vectors of the system. It can obtain Eq. (7) [7].

$$X = \begin{bmatrix} x \\ v_x \\ y \\ v_y \end{bmatrix}, \quad O = \begin{bmatrix} x \\ y \end{bmatrix}, \quad A = \begin{bmatrix} 1\,1\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,1 \\ 0\,0\,0\,1 \end{bmatrix}, \quad B = \begin{bmatrix} 1\,0\,0\,0 \\ 0\,0\,1\,0 \end{bmatrix} \tag{7}$$

Among them, (x, y) is the center position of the moving target, v_x and v_y express the motion speed of the moving target in the direction x and direction y. Prediction results by using Kalman filter are shown in Fig. 4, '+' is true value and '∘' is predictive value in the prediction process.

The location of the trajectory of the moving target in the new frame is predicted through Kalman filter, and the candidate moving targets near the location are searched in the frame image. If the motion target exists, the trajectories of the motion target are extended to the frame, and the center of the candidate moving target is used as the trajectory in the frame. If the corresponding candidate moving

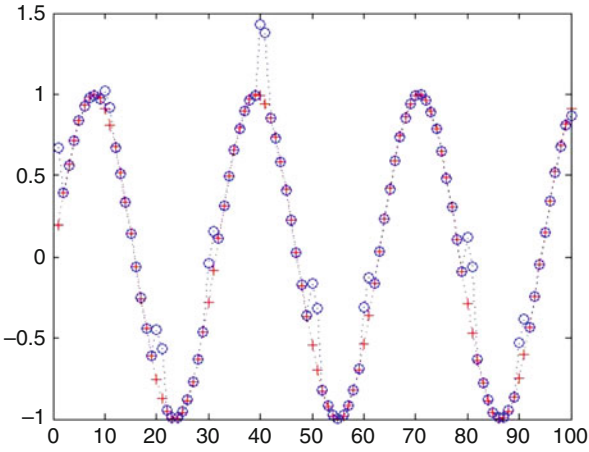


Fig. 4 Prediction process of Kalman filter

target is not found, the moving target corresponding to this trajectory is missing, blocked, or disappearing. When the false or occluded frames do not exceed the threshold, the trajectory of moving target is extended to this frame and the position of trajectory in this frame is instead of prediction value predicted by using Kalman filter. When the false or occluded frames exceed the threshold, we think this path has disappeared in the video, and then stop the growth trajectory. By means of trajectory growth, we can get a plurality of trajectory generated by candidate moving objects from the segments of soccer game video (including the moving target and noise), as shown in (a) in Fig. 5. The trajectory of red and blue is the track of two teams. Green trajectory is the track of soccer. When the soccer leaps and bounds in the air, its trajectory is mapped to the field. The reddish brown trajectory is the referee trajectory, and the red blue track is the track generated by noise. Because part of the generated trajectory is generated by noise, it is necessary to select the corresponding trajectories of the real moving target from these trajectories.

The set of trajectories generated by the real moving target is defined as C_{t}. The initialization set is C_{t}, the element in the set C_{t} is all the trajectories, that is $C_{\mathrm{t}} = \{T_i, i \in [1, N]\}$. Among them, T_i is the ith track of the video clip in the current football game. N is expressed as the number of the tracks in the video clip of the current football game. We randomly select two trajectories T_{u} and T_{v} with different starting frames in soccer video clips, $K_{\mathrm{s,\,u}}$, $K_{\mathrm{e,\,u}}$, $K_{\mathrm{s,\,v}}$, and $K_{\mathrm{e,\,v}}$ are corresponding to their starting frames and ending frames respectively and $K_{\mathrm{s,\,u}} \leq K_{\mathrm{s,\,v}}$. When the end frame of the trajectory T_{u} is larger than the starting frame of the trajectory T_{v}, that is $K_{\mathrm{e,\,u}} \geq K_{\mathrm{s,\,v}}$, the trajectories T_{u} and T_{v} intersects in the space-time domain, $T_{\mathrm{u}} \cap T \neq \phi$. On the other hand, the two trajectories are considered to be separated. In the video segment of a football game, the trajectories of the moving targets are usually long, and the track produced by the noise is shorter. Therefore, when the two trajectories cross, the trajectory of the moving target takes a longer trajectory.

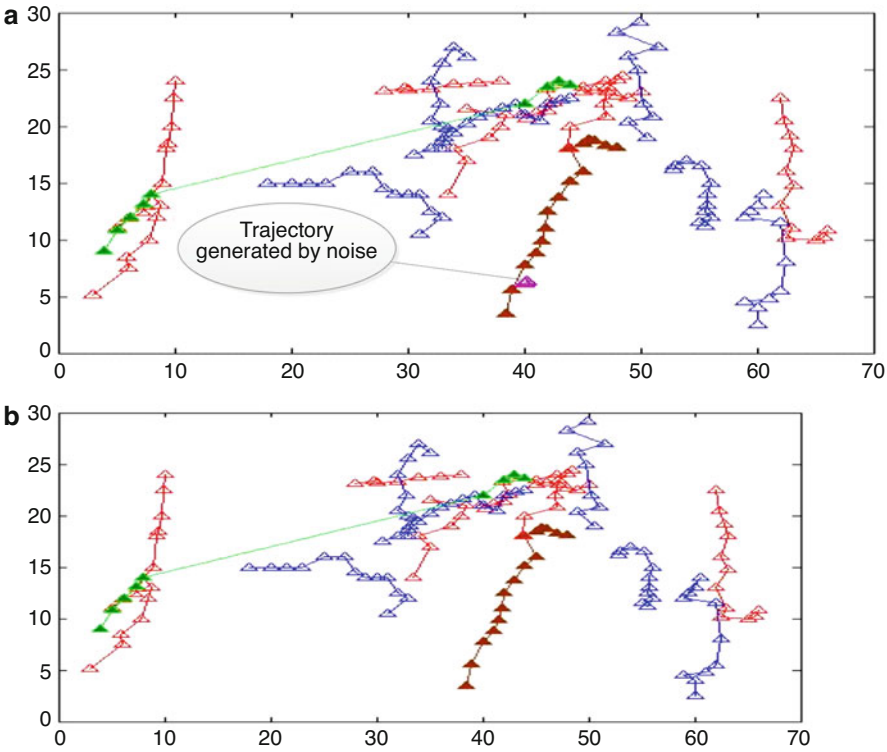


Fig. 5 Trajectory generation diagram for video football game fragment. (**a**) Trajectories of moving objects with noise trajectories. (**b**) Trajectories of moving objects without noise trajectories

We can use the Eq. (8) to calculate the set of trajectories generated by the moving target.

$$C_t = \begin{cases} C_t - \{T_u\} \text{ if } \left(\left(K_{e,u} - K_{s,u}\right) < \left(K_{e,v} - K_{s,v}\right)\right) \wedge (T_u \cap T_v \neq \phi) \\ C_t - \{T_v\} \text{ if } \left(\left(K_{e,u} - K_{s,u}\right) \geq \left(K_{e,v} - K_{s,v}\right)\right) \wedge (T_u \cap T_v \neq \phi) \end{cases} \tag{8}$$

By choosing the trajectory, the set C_t of the trajectories of the moving target is finally obtained, as shown in (b) in Fig. 5. The separate trajectories are included in set C_t, that is, the missing frames exist between the trajectories. The main reason for the missed frames is the mutual occlusion between the moving targets on the field and the sudden change in the direction and speed of the motion. In order to get the full trajectories of the video clip of the football game, it is necessary to connect these separate tracks.

First, the predicted values $\hat{p}_{k,u}$ and $\hat{p}_{k,v}$ of the trajectories T_u and T_v in the interval $[K_{e,u}, K_{s,v}]$ are calculated. Then we search for two points closest to the distance between the two trajectories in the predicted interval $[K_{e,u}, K_{s,v}]$,

corresponding to the frame a and frame b in the trajectories T_{u} and T_{v}. In the prediction process, Eq. (9) is used as constraint condition.

$$\begin{cases} (a,b) = \arg\min\limits_{a,b} dist\left(\hat{p}_{a,\mathrm{u}}, \hat{p}_{b,\mathrm{v}}\right) \\ a \le b, \\ K_{\mathrm{s,u}} \le a \le K_{\mathrm{e,v}}, \\ K_{\mathrm{s,u}} \le b \le K_{\mathrm{e,v}}. \end{cases} \tag{9}$$

When $a \ge b$, the location of the moving target which is missed before the frame a and the location of the moving target is not detected after the frame a were expressed by the predicted value of trajectory T_{u} and track T_{v} in this interval, respectively. Their mean is used to represent the position of moving target in the frame a. We can use Eq. (10) to calculate this mean.

$$p_k = \begin{cases} \hat{p}_{k,\mathrm{u}} & K_{\mathrm{e,u}} \le k < a \\ \left(\hat{p}_{k,\mathrm{u}} + \hat{p}_{k,\mathrm{v}}\right)/2 & k = a \\ \hat{p}_{k,\mathrm{v}} & a < k \le K_{\mathrm{s,v}} \end{cases} \tag{10}$$

When $a < b$, the location of the missed moving target around frame a can be calculated as same as the time $a \ge b$. The motion of the moving target is smaller in the two frames a and b, and the more accurate position of the moving target is obtained by using the linear interpolation method (Eq. 11).

$$p_k = \begin{cases} \hat{p}_{k,\mathrm{u}} & K_{\mathrm{e,u}} \le k \le a \\ (k-a)\left(\hat{p}_{b,\mathrm{v}} - \hat{p}_{a,\mathrm{u}}\right)/(b-a) & a < k < b \\ \hat{p}_{k,\mathrm{v}} & b \le k \le K_{\mathrm{s,v}} \end{cases} \tag{11}$$

The location of the missed moving target can be accurately filled through the connection, and the complete track of moving target can be generated.

Extraction of Temporal and Spatial Interest Points of Players and the Referee

The temporal and spatial interest points refer to the relatively large intensity changes in time and space. Temporal and spatial interest point representation is a relatively new underlying representation method of characteristics for frame sequences. The video sequence of football game video is set as $v : \Re^2 \times \Re \to \Re$. Linear scale space representation $R : \Re^2 \times \Re \times \Re_+^2 \mapsto \Re$ is constructed by convolution with anisotropic Gaussian kernel with different spatial and temporal variances σ_{r}^2 and τ_{r}^2 of v. The relationship between them can be expressed by using Eq. (12).

$$R\left(\cdot;\sigma_{\mathrm{r}}^2,\tau_{\mathrm{r}}^2\right)=g\left(.;\sigma_{\mathrm{r}}^2,\tau_{\mathrm{r}}^2\right)*v\left(\cdot\right) \tag{12}$$

The space-time separable Gauss kernel is defined by Eq. (13).

$$g\left(x,y,t;\sigma_{\mathrm{r}}^2,\tau_{\mathrm{r}}^2\right)=\frac{\exp\left(-\left(x^2+y^2\right)/2\sigma_{\mathrm{r}}^2-t^2/2\tau_{\mathrm{r}}^2\right)}{\sqrt{(2\pi)^3\sigma_{\mathrm{r}}^4\tau_{\mathrm{r}}^2}} \tag{13}$$

It is of great concern to use separate scale parameters for both temporal and spatial domains, since the temporal and spatial components of events are generally independent. Moreover, the detection of events by temporal and spatial interest point operators depends on the observed time and space scales, so it is necessary to deal with the corresponding temporal and spatial scale parameters σ_{r}^2 and τ_{r}^2 separately.

In this paper, we consider a 3×3 space-time second-order moment matrix consisting of first-order space and time scales, and Gauss weighting function is used to the average of Gauss kernel by Eq. (14).

$$g\left(.;\sigma_{\mathrm{r}}^2,\tau_{\mathrm{r}}^2\right)\mu=g\left(.;\sigma_{\mathrm{r}}^2,\tau_{\mathrm{r}}^2\right)*\begin{pmatrix} {r_x}^2 & r_xr_y & r_xr_t \\ r_xr_y & {r_y}^2 & r_yr_t \\ r_xr_t & r_yr_t & {r_t}^2 \end{pmatrix} \tag{14}$$

The scale parameters σ_i^2 and τ_i^2 in Eq. (14) are fused into the local scale parameters σ_{r}^2 and τ_{r}^2 by using $\sigma_i^2=s\sigma_{\mathrm{r}}^2$ and $\tau_i^2=s\tau_{\mathrm{r}}^2$, the first derivative is defined by Eq. (15).

$$\begin{cases} r_x\left(x,y,t;\sigma_{\mathrm{r}}^2,\tau_{\mathrm{r}}^2\right)=\partial x\left(g*v\right) \\ r_y\left(x,y,t;\sigma_{\mathrm{r}}^2,\tau_{\mathrm{r}}^2\right)=\partial y\left(g*v\right) \\ r_t\left(x,y,t;\sigma_{\mathrm{r}}^2,\tau_{\mathrm{r}}^2\right)=\partial t\left(g*v\right) \end{cases} \tag{15}$$

In order to detect interest temporal and spatial points, we search for the regions with significant eigenvalues λ_1, λ_2 and λ_3 in μ in video. In different methods of searching region, spatial domain Harris corner function is defined as Harris corner function in space-time domain by combining determinant and tracking extension of μ in Eq. (16).

$$H=\lambda_1\lambda_2\lambda_3-k(\lambda_1+\lambda_2+\lambda_3)^3,\quad(\lambda_1\le\lambda_2\le\lambda_3) \tag{16}$$

In order to show that the positive local maximum of H corresponds to the high value point λ_1, λ_2, and λ_3, the ratio is defined as $\alpha=\lambda_2/\lambda_1$ and $\beta=\lambda_3/\lambda_1$, and the Eq. (16) can be rewritten as Eq. (17).

$$H=\lambda_1^3(\alpha\beta-k\left(1+\alpha+\beta\right) \tag{17}$$

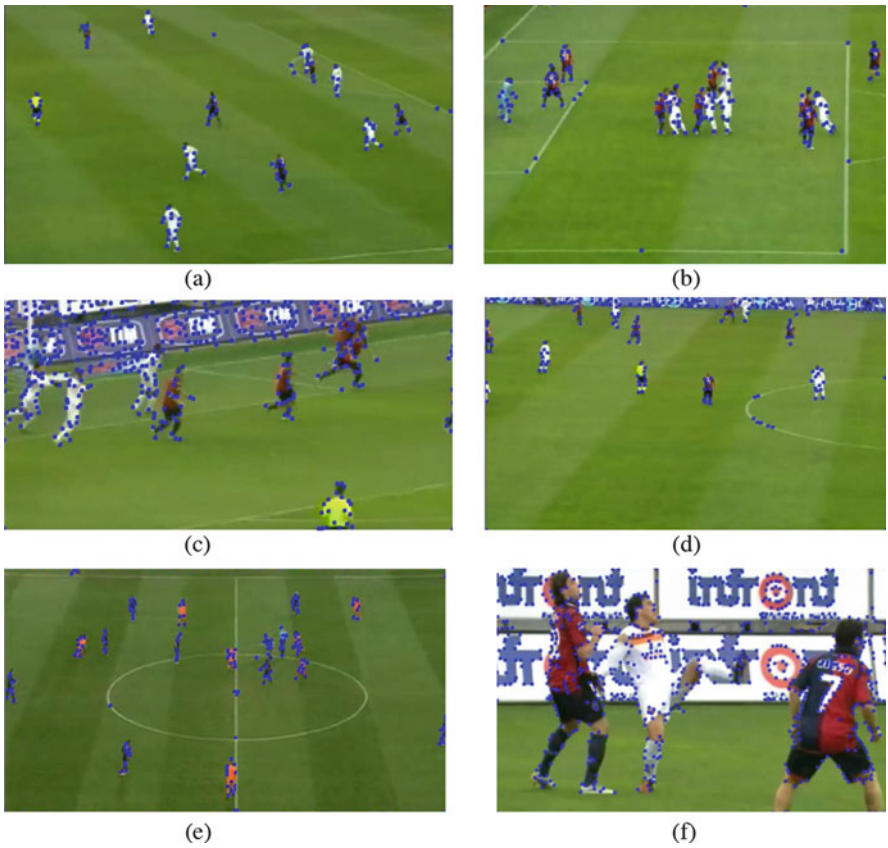


Fig. 6 Detection results of temporal and spatial interest points. (**a**) 52nd frame, (**b**) 88th frame, (**c**) 137th frame, (**d**) 223rd frame, (**e**) 76th frame (another video), (**f**) 361st frame (close-up frames)

If $H \geq 0$, then $k \leq \alpha\beta/(1+\alpha+\beta)^3$. Assume $\alpha = \beta = 1$, the maximum possible value of k is 1/27.When the value k is large enough, the point corresponding to the positive local maximum H varies sharply along the time and space directions. Especially when the maximum value of α and β in the spatial domain is 23, $k \approx 0.005$. Therefore, we can detect the temporal and spatial interest points in the video sequence v of football matches by detecting the positive local temporal and spatial maximum of H. Time and space interest point detection results are shown in Fig. 6.

4 Conclusion

Because single feature is difficult to effectively describe the behavior characteristics of multiple athletes, it is not reliable to use single feature to identify the behavior of multiple athletes. According to the characteristics of various features of athletes needed to extract in the process of multiplayers behavior recognition in soccer game video and feature extraction has a great influence on the final recognition results, in this paper, we extract the contour features, clothing color histogram, temporal and spatial interest points, and color moment features of players and referees. We also transform the soccer video images into binary images and extract feature of coordinate parameters of field line using the Hough transform, and calculate accurate linear coordinates using gray fitting. We use Kalman filter to track the moving object and predict its motion trajectory. In order to reduce the burden of high-level recognition algorithm, we propose using the method of growth trajectories to extract low-level features, such as trajectory features of the moving object. The experimental results show that it can greatly improve the accuracy of the behavior recognition by using these features to identify the behavior of the athletes.

Acknowledgments The authors are very grateful for the support provided by the National Natural Science Foundation of China (61462008, 61751213, 61866004), the Key projects of Guangxi Natural Science Foundation (2018GXNSFDA294001, 2018GXNSFDA281009), the Natural Science Foundation of Guangxi (2017GXNSFAA198365), 2015 Innovation Team Project of Guangxi University of Science and Technology (gxkjdx201504), Scientific Research and Technology Development Project of Liuzhou (2016C050205).

References

1. Lazebnik, S., & Raginsky, M. (2009). Supervised learning of quantizer codebooks by information loss minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(7), 1294–1309.
2. Cao, Y., Wang, H., Wang, C., Li, Z., Zhang, L., & Zhang, L. (2010). Mindfinder: Interactive sketch-based image search on millions of images. In *Proceedings of the international conference on multimedia* (pp. 1605–1608). New York: ACM.
3. Eitz, M., Hildebrand, K., Boubekeur, T., & Alexa, M. (2011). Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE Transactions on Visualization and Computer Graphics, 17*(11), 1624–1636.
4. Legg, P. A., Chung, D. H. S., Parry, M. L., Bown, R., Jones, M. W., Griffiths, I. W., & Chen, M. (2013). Transformation of an uncertain video search pipeline to a sketch-based visual analytics loop. *IEEE Transactions on Visualization and Computer Graphics, 19*(12), 2109–2118.
5. Eitz, M., Richter, R., Boubekeur, T., Hildebrand, K., & Alexa, M. (2012). Sketch-based shape retrieval. *ACM Transactions on Graphics, 31*(4), 31:1–31:10.
6. Lee, J., & Funkhouser, T. (2008). Sketch-based search and composition of 3D models. In *EUROGRAPHICS workshop on sketch-based interfaces and modeling*, June 2008.
7. Lee, Y. J., Zitnick, C. L., & Cohen, M. F. (2011). Shadowdraw: Real-time user guidance for freehand drawing. In *ACM SIGGRAPH 2011 papers* (pp. 27:1–27:10). New York: ACM.

8. von Landesberger, T., Bremm, S., Bernard, J., & Schreck, T. (2010). Smart query definition for content-based search in large sets of graphs. In *Proceedings of the International Symposium on visual analytics science and technology* (pp. 7–12). Geneva: Eurographics Association.
9. Scherer, M., Bernard, J., & Schreck, T. (2011). Retrieval and exploratory search in multivariate research data repositories using regressional features. In *Proceedings of the 11th Annual International ACM/IEEE joint conference on digital libraries* (pp. 363–372).
10. Shao, L., Behrisch, M., Schreck, T., von Landesberger, T., Scherer, M., Bremm, S., & Keim, D. A. (2014). Guided sketching for visual search and exploration in large scatter plot spaces. In *Proceedings of the EuroVA International workshop on visual analytics*. Geneva: The Eurographics Association.
11. Wang, B., Ruchikachorn, P., & Mueller, K. (Dec 2013). Sketchpadn-d: Wydiwyg sculpting and editing in high-dimensional space. *IEEE Transactions on Visualization and Computer Graphics, 19*(12), 2060–2069.
12. Michael, N., & Lanitis, A. (2014). Model-based generation of realistic 3D full body avatars from uncalibrated multi-view photographs. In L. Iliadis, I. Maglogiannis, & H. Papadopoulos (Eds.), *Artificial intelligence applications and innovations. AIAI 2014. IFIP advances in information and communication technology* (Vol. 436). Berlin: Springer.
13. Zhiwen, W., & Shaozi, L. I. (2014). Adaptive fractal-wavelet image denoising based on multivariate statistical model. *Chinese Journal of Computer, 37*(6), 1380–1389.
14. Ashok, V., Amit, R. C., & Rama, K. C. (2005). Matching shape sequences in video with applications in human movement analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(12), 1896–1909.
15. Wang, L., & Yung Nelson, H. C. (2010). Extraction of moving objects from their background based on multiple adaptive thresholds and boundary evaluation. *IEEE Transactions on Intelligent Transportation Systems, 11*(1), 40–51.
16. Michael, N., Drakou, M., & Lanitis, A. (2017). Model-based generation of personalized full-body 3D avatars from uncalibrated multi-view photographs. *Multimedia Tools and Applications, 76*(12), 14169–14195.
17. Fuchs, Z. E., Casbeer, D. W., & Garcia, E. (2016). Singular analysis of a multi-agent, turn-constrained, defensive game. In *2016 American Control Conference (ACC) Boston Marriott Copley Place*, July 6-8, 2016. Boston, MA, USA (pp. 4705–4712).
18. Lin, S., Dominik, S., Benjamin, N., Manuel, S., & Tobias, S. (2016). Visual-interactive search for soccer trajectories to identify interesting game situations. *Electronic Imaging, Visualization and Data Analysis*, 510.1–510.10.
19. Boiman, O., & Irani, M. (2007). Detecting irregularities in images and in video. *International Journal of Computer Vision, 74*(1), 17–31.
20. Shintani, T., Nobata, M., Muneyasu, M. (2016). A-15-14 a Silhouette extraction method for moving objects based on image characteristics. In *IEICE Engineering Sciences Society/Nolta Society Conference*, 2016.
21. Ahmed, H. A., Rashid, T. A., & Sadiq, A. T. (2016). Face behavior recognition through support vector machines. *International Journal of Advanced Computer Science and Applications, 7*(1), 101–108.
22. Alba-Cabrera, E., & Godoy-Calderon, S. (2016). Generating synthetic test matrices as a benchmark for the computational behavior of typical testor-finding algorithms. *Pattern Recognition Letters, 80*(1), 46–51.
23. Heng, F., Jun, X., Yong, D., & Jinhai, X. (2016). Behavior recognition of human based on deep learning. *Geomatics and Information Science of Wuhan University, 41*(4), 492–497.
24. Kim, H., Lee, S., Kim, Y., Lee, S., & Lee, D. (2016). Weighted joint-based human behavior recognition algorithm using only depth information for low-cost intelligent video-surveillance system. *Expert Systems with Applications, 45*(C), 131–141.
25. Jiang, Q. (2016). Research of multiple-instance learning for target recognition and tracking. *EURASIP Journal on Embedded Systems, 2016*(1), 1–6.
26. Wang, X., Gao, B., Masnou, S., Chen, L., & Theurkauff, I. (2016). Active colloids segmentation and tracking. *Pattern Recognition, 60*, 177–188.