Automatic tracking of indoor soccer players using videos from multiple cameras

Erikson Morais, Siome Goldenstein, Anselmo Ferreira, Anderson Rocha
Institute of Computing – IC
University of Campinas – UNICAMP
Campinas, SP, Brasil
{emorais,siome,ra023169,anderson.rocha}@ic.unicamp.br

***Abstract*—Indoor soccer has been of tactical and scientific interest, with applications dedicated to analyze tactical and physiological factors and also physical training. In both cases, the analysis is based on player tracking, done with human supervision. This paper presents an automatic tracking method which shows the trajectories of indoor soccer players during the game and saving skilled labor during the process. For this, we use a predictive filter to model the motion and the observation of multiple stationary cameras, strategically positioned around the court. We associate a particle filter to a robust probabilistic observation model with the measurement in court coordinates. The observation model proposed is based on data fusion across multiple camera coordinates and projected onto the court plane, creating a multimodal and bidirectional probability function, which represents the potential localization of players in the court plane. The probability function uses an appearance model to observe player's location, distinguishing very close players and yielding good weights in the observation model. The experimental results show tracking errors below 70 centimeters in most cases and indicate the potential of the method to help sports teams.**

***Keywords*-Tracking, Soccer, Particle Filter, Data Fusion.**

I. Introduction

Indoor soccer plays an important role in sports nowadays due to its dynamic nature and reduced space to play, which often force the teams to think better and practice their tactics to the extreme. For this reason, this sport has been the target of technical and scientific interest, showing several applications dedicated to tactical, physical and physiological analysis [1].

The tactical staff is interested in player trajectories to verify the positioning efficiency in the game. The technical staff, responsible for the players fitness, analyzes the trajectories to verify data such as speed reached, acceleration peaks and distance traveled to establish the physical training. The physiological analysis uses the player trajectory to evaluate the stress levels to help in physical training. A Computer Vision system able to retrieve the player trajectories in a game can be useful to help sport teams to improve all of these aspects [2].

With the advance of technology, the video cameras and recording media became very accessible. Nowadays, it is cheap to acquire multiple recordings of an indoor soccer game with multiple high-resolution cameras. With these videos, the teams can review their moves, the opponent moves and identify important characteristics regarding the opponents.

Multiple recordings, taken from different observation points, can store important redundancies for automatic processing and further analysis. Estimations of position can be improved by using multiple observations and a data fusion process, resulting in more reliable estimations. The use of Computer Vision methods in indoor soccer are also important in a scientific point of view. The dynamic environment of a collective game allows us to try and validate new methods, not only for applications focused on sports but also for other cases, such as security, pedestrian counting, and surveillance monitoring.

Most proposed methods to solve the player tracking issue in literature are partially automated and have assisted tracking steps. In this context, this paper discusses a method to retrieve the trajectories of indoor soccer players during a game automatically with human intervention only at startup. The principal advantage is its automation, saving skilled labor during the process with great potencial to help sports teams analisys. The contribution is in observation method with fusion of data found in multiple camera projected onto the court plane.

The method consists of a tracking system based on a particle filter, in which the observation stage combines the data found in different cameras projected onto the court coordinates. Initially, for each video on each camera, the method uses a computer vision standard detector trained to find people and, therefore, the players. A representation of these players is projected onto a tridimensional virtual plane that represents the court using homography. To deal with possible projection errors, the method characterizes the position of players in court coordinates using a probabilistic representation. Such representation is done by a bidirectional and multimodal Gaussian function, whose output is the probability of a player being found in a given position $S(x, y)$ of the court. The multimodal function is used with an appearance model to strengthen the result, distinguishing close players on the court. Finally, the proposed method uses this multimodal function in conjunction with the appearance model to weight the particles used to compute the trajectory in the court plane coordinates.

II. Related Works

In recent years, some researchers have turned their attention to the detection and tracking of indoor soccer players. In [2] and [1], the authors have presented an approach that deals with the problem of visual tracking as a graph shortest path problem. In the graph, the nodes represent the blobs detected

1530-1834/12 $26.00 © 2012 IEEE
DOI 10.1109/SIBGRAPI.2012.32

during the analysis of the game videos. Each edge links a blob to another blob of the next frame from the same video sequence. Each blob is connected to every other blob on the next frame. By determining which player in the first frame must be tracked, the method finds the shortest path to a final node and maps it to indoor soccer court coordinates.

Usually, tracking methods need an observation step which consists in the detection of the tracked objects. Some studies have used the separation of the background from the object of interest. In [2], [1] the background is found by averaging the frames of an excerpt of the video sequence and periodically recalculated. In [3], the background model is represented by a color histogram. Considering that the majority of the image is occupied by the court, the predominant value in the color histogram represents the area of interest that eliminates the other regions and leaves the players as evidence. In [4], the authors focus on the tracking of hockey players.

Detecting indoor soccer players is a similar problem to the localization of people in images. Viola and Jones [5], for instance, have presented an efficient method for detecting objects with a focus on face detection. This approach is easily adaptable to the localization of other objects. It is based on machine learning and uses simple features such as Haar filters arranged in a cascade. A sliding window runs through the input image, searching for an object of interest. The cascade is organized in order to quickly remove, in the early stages, a window with low chance of containing an object. The hypotheses maintained until the end are considered positive detections.

In [6], Felzenszwalb et al. have proposed a method based on a mixture of models in multi-scale deformable parts to represent objects with high variability. The method is trained using a discriminative procedure that requires only a set of images containing rectangular markings for training. Each part captures local appearance properties of an object, while the deformable configuration connects the parts to the pairs. Different from previous approaches, Khan and Shah [7] have presented a method based on [8] to separate the background, followed by homography to find the locations of people in a 3D scene.

Silhouette-based techniques have also been used. In [9], a method was proposed to detect basketball players and also track them based on the shortest distance between the detections of adjacent frames. The method presented in [10] aims at tracking soccer players in the field using stationary cameras. Players are found in images using *blob* detection by means of background separation, using a statistical model. The authors use homography to map 2D positions in the image plane onto 3D position in the plane of the (virtual) game field, and then the tracking is performed by a Kalman filter. However, detection of *blobs* does not separate objects very close or partially occluded. At critical moments, in which there are high proximity between the tracked objects, two or more objects of interest can be detected as one. This can occur as a result of occlusions caused by the camera perspective or proximity of similar objects that cause failures in the appearance model used. In addition, the Kalman filter works only with Gaussian functions and therefore can not deal with multiple hypotheses, resulting from players proximity. This often happens in indoor games such as indoor soccer.

III. Basic Concepts

The problem of visual tracking consists of processing a sequence of images to describe the movement of one or more objects in a scene [11] or, also, generating inferences about the movement of an object from a sequence of images [12].

To track an object, we must detect it on the image. Some methods in the literature process the whole image to locate the object [1], [10]. At each frame of a video sequence, all existing blobs are detected prior to deciding which one will be tracked. Most recent techniques use predictive filters that use a model for the object motion called *dynamics* to estimate the state of a tracked object in the next frame of a video sequence. The prediction is adjusted by a local observation, without taking into account all of the input images. Further details about predictive filters can be found in [13].

In a nutshell, two filters widely used in the literature are the Kalman and Particle filters. Kalman filters use Gaussian functions to estimate the next state of a tracking. The advantage is that the Gaussian function is a parametric function and has a simple representation [14], [13]. Such filters can estimate past, present or future states in a sequence or a function, which gives them power to infer lost states or the ones that could not be detected. This occurs, for example, when the tracked object is occluded by another element of the scene causing an occlusion. The problem with this approach is that the Gaussian function is unimodal, which means the filter cannot represent multiple hypotheses simultaneously.

Particle filters, such as those discussed in [15] and [16], on the other hand, represent functions of probabilities in a non-parametric form, using sets of weighted samples, called particles. This type of representation allows multimodal and unknown functions to be represented. Overall, each particle X_t is associated with a weight W_t, that represents the function value at the point determined by the particle. The higher the number of particles, the better is the representation of the desired function. This model of no parametric functions is used to represent the probability function of the object occurs in the next frame of the sequence, before the frame is observed. This function is called *a priori* probability function and is used to estimate the object state in the next frame.

In a particle filter, a state is the representation of the object tracked in each frame of an observed sequence. A state may contain any information considered relevant. Usually, we represent the states in a vector form. For instance, if the task is to track the position and velocity of an object, each state can be composed by an array of two values that corresponds to these informations.

The *a priori* probability function is an estimation of what should happen in the next frame, based on the dynamics of motion known. When the filter has access to the new frame, we need to adjust or confirm the *a priori* function.

At this point, each particle passes through a phase of *observation* that verifies its representation in that frame. The new weighted representation of the probability function represents the probability of a particle X_t, given an observation Z_t and is denoted by $P(X_t|Z_t)$. This new function is called *a posteriori* probability function.

The estimation for the next step is done by sorting particles from the current set, allowing repetition and giving preference to particles with higher weights. The new set of particles has the same size as before, but there are repetitions that can cause the system to collapse. Thus, the values of new particles are adjusted, considering the known motion dynamics and adding to each one a random error, representing the level of uncertainty in the process. This step is called *prediction* because, based on knowledge of the tracked object, it estimates the next set of particles, adding a random error to ensure the distinction among samples. The new function represented by the new set of particles is an *a priori* function $P(X_{t+1}|Z_t)$ for the new frame of the sequence.

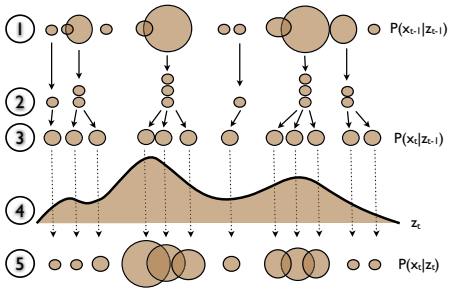

Fig. 1. Particle Filter steps. (1) *a posteriori* function $t-1$; (2) sampling; (3) propagation, resulting in an *a priori* function $P(X_t|Z_{t-1})$; (4) observation in t; and (5) *a posteriori* function $P(X_t|Z_t)$ in t.

Figure 1 shows the steps of one cycle of the particle filter. The first layer shows a set of particles represented by ellipses of different sizes. In the drawing, the size of an ellipsis represents the weight of the particle in the process. The second layer shows a set of particles that are repeated; this is the result of the sampling process. At this point, the weights of the particles lose their meaning and must be measured again. The third layer of the figure shows the result of the spread, where the used dynamics and a random error are applied to the sorted particles. At this point, the set of particles represents an estimation for the next frame of the video sequence. In the fourth layer, the measurement is represented as a function where the height represents the weight of the measurement at that point. After the measurement step, the set of particles of the fifth layer represents the *a posteriori* probability function, with the particles weighted and ready for a new iteration. This process is repeated until the end of the video sequence is reached.

IV. Proposed Method

In this paper, we discuss a method for tracking indoor soccer players based on particle filters with a robust probabilistic observation model. The contribution is in the proposed observation model, which consists of the fusion of data found in multiple camera coordinates projected onto a virtual plane representing the court.

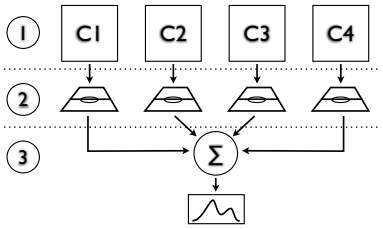

Fig. 2. Diagram showing the model stages. In Stage #1, players are detected in the image plane using, in each of the cameras, a detector trained to find indoor soccer players. In Stage #2, the observations are projected onto a virtual plane representing the court by homography. In Stage #3, the projected data are combined in a two-dimensional and multimodal probability function, which gives the probability of finding a player in a given position of the court.

For the observation of multiple cameras, players are detected using an object detector trained specifically to locate indoor soccer players. The positions found by the tracking system in each camera are projected onto a virtual plane representing the court plane coordinates (3D), with the aid of a homography operation. The method keeps track of players through a particle filter associated with each player, but with an observation function shared by all filters. The observation function used by the system consists of a multimodal and bidirectional function used in conjunction with an appearance model to map the probability of finding a player in a given position of the court. In the following, we give more details of each step of the method.

A. Particle Filter Motion Dynamics

As we described in Section III, each particle of a set is a possible state that the object may or may not take. In our case, the tracked object is a player in the game and its state consists of its $S(x, y)$ position in court coordinates. To describe how the object moves, we adopt a uniform motion dynamics as described in Equation 1 and 2. In this motion dynamics, the time variation is equivalent to the frequency of the video used (30 frames per second in this article).

$$S_t = S_{t-1} + V\Delta(t) \tag{1}$$

$$S_t = \begin{bmatrix} 1 & 0 & 1/30 & 0 \\ 0 & 1 & 0 & 1/30 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} x \\ y \\ v_x \\ v_y \end{bmatrix}_{t-1} \tag{2}$$

Besides the dynamics of the system, we need to weight the particles. For this, we use a multimodal and bidirectional function that gives the probability of finding a player in a given position on the court. Section IV-B describes this observation function.

B. Observation Function

The observation function attached to the particle filter contains three stages, showing in Figure 2:

1) **Stage** $\#1$ – Detection of the players in image coordinates. In this paper, we use the detector proposed by Viola and Jones [5].
2) **Stage** $\#2$ - Projection of detections onto court coordinates. The data found by the detector are projected onto court coordinates using homography.
3) **Stage** $\#3$ – Fusion of multi-camera data. The 3D projections are combined into a multimodal function composed by a bidirectional Gaussian with known covariance. The resulting function gives the probability of detecting a player in a given position of the court.

*1) **Stage** $\#2$ — **Projection of Detections onto Court Coordinates**:* The objects of interest move on the plane of the court and are situated on a plane in a 3D world. We can use the homography of specific points (*e.g.*, feet of the players) to find their locations in court coordinates. Each player found by the detector is represented by a rectangle in the image plane. In our work, we consider the middle basis of this rectangle as a good representation of the location of the player's foot in the image plane.

To perform homography, we can use points of the court whose coordinates we know *a priori*. The points already known, such as the penalty mark and court corners, serve to find a mapping function between points in image plane and points in court coordinates. For this, we can find a matrix H that maps points of the plane a to their counterparts in a plane b, starting from a known set of matches. Using the image as the plane a, we find the matches to known points described in the plane b (in this case, the plane of the court). These matches are obtained using a frame of the captured video to mark the correspondences between points and, consequently, to find the transformation matrix for the camera before the video processing.

*2) **Stage** $\#3$ — **Multi-Camera Data Fusion**:* After the projection, we have a global set of detections described in world coordinates. Due to possible errors accumulated by both the detection and the projection, these points are only an estimation for the location of each player and not its actual position. Considering the projected points as regions with potential to match the location of a player, we can replace each point by a function (e.g., Gaussian 2D). In our work, we represent the plane of the court with a single multimodal function, consisting of a mixture of Gaussians corresponding to the projected points by multiple cameras. Each Gaussian represents the uncertainty in projecting the corresponding point and has mean (mean error of the projection) and covariance that vary with each camera.

We can calculate the parameters of the Gaussian from a small set of training videos where the player positions are manually annotated in the court plane coordinates. By knowing the actual positions of each player in training videos, we can compare each detection with its correspondent in the annotated set. The correspondence is given by the shortest distance. However, there are cases where multiple players are close, leading to confusion. We consider a correspondence valid for the purpose of training only the ones whose nearest annotated point is less than two meters (L_1) and the second closest is more than three meters (L_2). From the correspondences, we can calculate the mean error of projection in x and y directions and covariance of error in each camera. In this way, we replace each point by a Gaussian whose mean and covariance are calculated for the camera that generated the point.

Each Gaussian $\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$ is considered one component of a mixture to form a complex function representing the court plane. A linear combination of Gaussians can give rise to very complex densities [17]. In our case, we have a set of detections with the same importance to find players on the court plane. Therefore, initially, we consider a Gaussian mixture model with equal combination coefficients π_k respecting the condition $\sum p_k = 1$ and $\pi_k = \frac{1}{K}$ considering a superposition of K Gaussians. The model is given by

$$P(\mathbf{x}) = \sum_{k=1}^{K} \frac{1}{K} \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k). \tag{3}$$

Finally, we can represent all the projected points by a multimodal function, consisting of a mixture of Gaussians whose parameters are calculated as described above. The function is a representation of the plane of the court in which we can measure the probability of finding a player at a given point and can be used directly in the observation step of the particle filter. Figure 3 depicts a representation of the system observations for an analyzed game.

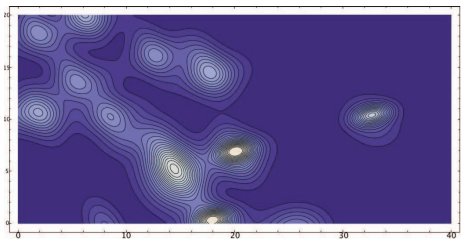

Fig. 3. Observation of the tracking system in one of the cycles of the particle filter. The multi-camera projections are replaced by Gaussian functions with previously trained parameters. As a result, there is a 2D function that returns the probability of finding a player in a given region.

*3) **Stage** $\#1$ — **Player Detection In Image Coordinates**:* The first step is finding the players in image coordinates independently in each camera. In this paper, we use the standard Viola and Jones [5] model, trained to detect indoor soccer players. Other detectors as [6] could be used as well.

C. Appearance Model to Strengthen the Multimodal Function

The multimodal function discussed in the previous section gives us the potential of finding a player in a real-world coordinate, but not distinguishing the players because all Gaussians have the same weight in the mixture.

Each Gaussian is a result of one detection in camera coordinates and corresponds to one rectangle detected in a camera point-of-view. This rectangle can be used to compute the appearance of a player detected and then is compared with an appearance model in that camera. The comparison

results in a similarity value between the detect appearance and the expected appearance given by a known model for the correspondent camera. The similarity value can be used as the coefficient π_k in Equation 3 for the corresponding Gaussian. Therefore, the most similar detection with respect to the appearance model in the correspondent camera yields a Gaussian with higher importance in the mixture.

To calculate the appearance of a detected player, we use color information for distinguishing players of different teams and gradient information to represent details of specific players. We know that the shape of a player does not have large changes between consecutive frames in a video. Hence, histograms of gradients will be very similar in this situation and can be a good representation of line distribution in the rectangle detection.

The histogram of gradients are computed by four steps:

1) We apply the Sobel algorithm on the sub-image of interest to find derivatives in x and y directions, dx and dy. This requires filtering the intensity sub-image using the kernels $[-1, 0, 1]$ and $[-1, 0, 1]^T$.
2) Use dx and dy to calculate one gradient magnitude image Gr and one gradient orientation image A.
3) Summarize the sub-region information into two 15-bin histograms: one for Gr and another for A. Additionally, compute one 255-bin histogram for each color channel of the sub-image considered. Finally we normalize all histograms.

Figure 4 depicts an appearance of a player considering a set of histograms. We divide the sub-image that contains the player in 3 regions: top, middle and bottom. Each region results in five histograms: one for the Gradient's magnitude, one for the Gradient's orientations and one to each color channel of the image considering the HSV color-space.

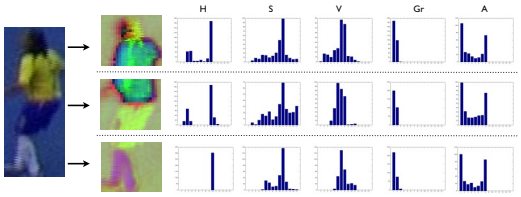

Fig. 4. Appearance model for one observed player. Each image sub-region is converted to HSV color-space and divided into three regions: top, middle and bottom. Each region is characterized with five histograms: one histogram per color channel (H, S and V), one for gradient's magnitudes (Gr) and other for gradient orientations (A).

We use histogram correlation to compare two appearance models $S(AP_1, AP_2)$. Considering one appearance model as a list of 15 histograms (five per region, with three regions) and AP^i as i-th histogram on the list, we have:

$$S\left(AP_1, AP_2\right) = \prod_{i}^{15} CORR(AP_1^i, AP_2^i). \tag{4}$$

According to this model, each detected player has an appearance model that needs to be compared to others in the corresponding camera view to obtain the π_k coefficient. This coefficient will be the importance of such Gaussian to represent the position of the corresponding detected player. In this work, the appearance model in the camera view is the set of good appearances previously found. Therefore, π_k can be the maximal similarity comparing the detected appearance (AP_d) in the current frame with all appearances on the model (AP_m) as in the Equation 5:

$$\pi_k = \max\left\{S(AP_d, AP_m)\right\}. \tag{5}$$

With this, re-writing Equation 3, we have:

$$P(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k). \tag{6}$$

This model is constantly updated to maintain a good representation of the object. This is done replacing the appearance of a candidate to get out (CO) by a new good appearance detected, considered a candidate to get in (CI). In each comparison between the model and any other appearance, the model updates a history of utilization to signal the last used appearance. This appearance is considered the CO. One observed appearance is considered a good CI when it generates the best similarity with the current model. As long as the model maintains information about the appearances use, the system knows what region of interest results the highest weight during the observation step and uses the corresponding appearance as a candidate to get in. In case of one CI appearance receives similarity greater than a threshold, this appearance CI will replace the appearance CO in that model. This information is maintained by the particle filter observation step.

At each new iteration of the filter, a set of particles must be weighted. For this reason, the Gaussian mixture corresponding to the function is weighted using the appearance model of their corresponding cutouts. The appearance of the sub-image of each Gaussian is compared to the appearance model of the corresponding camera, maintained by the tracker. Thus, the observation model used by the filter has N appearance models for the observed object, one for each camera. The result is a multimodal function with the highest peaks in the regions corresponding to the object tracked and therefore able to identify the proximity of players, but with emphasis on the correct object.

V. Experiments and Validation

A. Data

For validation, we use videos of indoor soccer games recorded by a set of four stationary cameras positioned around the court (See Figs. 5 and 6).

We have a collection of seven games in Full-HD recorded using four cameras, each at 30 frames per second. Each game has two periods. For simplicity, we consider each period as a game, with a total of 14 games recorded by four cameras. This collection was recorded during the 2009 Female South American Indoor Soccer Championship.

Fig. 6. Examples of corresponding frames in a game recorded by different cameras.

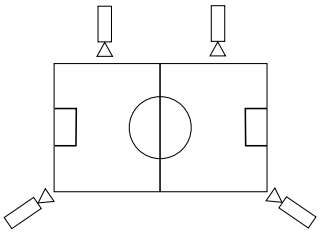

Fig. 5. Positioning of cameras around the court. Four cameras were positioned around the court with overlapping fields of view. Each camera observes one of the halves of the court so that a player is observed by at least two cameras.

B. Training and Calibration of Parameters

As we already discussed, the particle filter observation method requires two training stages: one for the player detector in image coordinates and the other to calibrate the parameters needed for the construction of the multimodal function that represents the plane of the court. For this, we used one period of the games (one period of a game with four camera footage). The remaining 13 games are used for testing.

To train the player detector in image coordinates, we use a set of 16,000 positive samples and another with approximately 18,000 negative samples. A positive sample is a small rectangle containing one player, while a negative sample contains the background. In the parameter training for the multimodal function, we compared the detections of the plane of images projected onto the plane of the court with the real markings (manual annotation). Then, we calculate the average projection errors and their covariances for each camera, in order to calibrate the parameters used in the multimodal function representing the plane of court, as discussed in Section IV-B2.

We have performed the two training steps using an entire annotated video. For testing, we consider only the first two minutes of each game. This limitation is due to the difficulty of annotating by hand the positions of players on the plane of court coordinates for ground-truth.

For the correct operation of the particle filter, we need to start the set of particles from the initial positions of the players who we want to track. For this, we need to mark the position of the players of interest in the first frame of each video. In our case, we are interested in the 10 players on the court and also the two referees. For each object of interest, we initiate a filter with a set of 500 particles.

C. Experiments of tracking without the appearance model

During the tests, most of the obtained trajectories were successful. However, there are cases in which there are confusions between players who have approached each other during tracking. In general, the referees are positioned on the side of the court and have simplified motion, leading to successful tracking trajectories.

Figure 7 shows the mean errors found in some successful trajectories and Figure 9 shows their trajectories found by the tracking approach, plotted with the manual markings. The measurement error represents the difference between position estimates and the manually annotated positions.

We note, by observing the successful cases, that the average accumulated error in each trajectory is below one meter in most cases and 0.4 meters in the best case. When we analyze all trajectories together, we find a global average error of 0.73 meters. This is an interesting value, considering the dimensions of the court ($20m \times 40m$). This result is encouraging, since it is obtained automatically after a simple training from a single video. Table I shows errors and standard deviations found by analyzing the collection of successful trajectories.

Situations in which the trajectories of two or more players cross may lead to confusions and must be addressed. Such confusions are made by the proximity of the trajectories and are solved by the prediction filter, based on the object motion dynamics. However, when the trajectories of two or more players tend to be coincident for a moment, we have a total confusion situation. This can occur, for example, in a ball dispute or celebration. In such cases, the players are very close to each other and move in the same direction. This configuration makes the filter rely on the observation stage for correction. Sometimes, the observation does not distinguish between detected players. This is due to the particle measurement that observes only the likelihood given by the observation function. Nearby peaks with similar motion cause the division in the set of particles and, in this case, the prediction system cannot deal with the confusion.

This problem can be solved by strengthening the observation model with adaptive appearance models, coupled to the particles. With this approach, the particle find the probability of a player being in a given position, and can check the probability of a player given its appearance model. To maintain the representativeness, the model needs to be updated throughout the whole process, since there are variations of light, rotation and scale of the observed objects. This improvement is what we show in the next section.

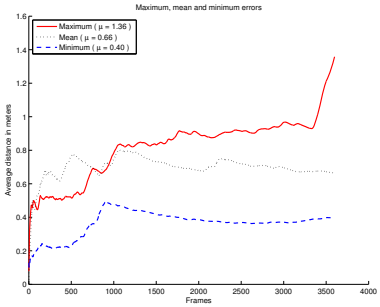

Fig. 7. Mean of accumulated errors found in successful trajectories considering the tracking **without the appearance model**. The figure shows the curves found for the trajectories of maximum, minimum, and mean error.

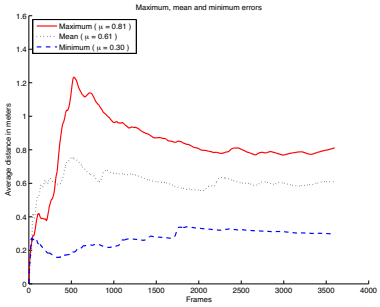

Fig. 8. Mean of accumulated errors found in successful trajectories considering the tracking **with the appearance model**. The figure shows the curves found for the trajectories of maximum, minimum, and mean error.

TABLE I
ERRORS AND STANDARD DEVIATIONS FOR THE COMPLETE TRAJECTORIES FOUND BY THE TRACKER WITHOUT THE APPEARANCE MODEL.

Track	μ	σ
BoliviaxColombia-T1-player-10.trj	0.60	0.11
BoliviaxColombia-T2-player-10.trj	0.82	0.27
BrazilxColombia-T2-player-11.trj	0.54	0.12
BrazilxColombia-T2-player-12.trj	0.73	0.36
BrazilxPeru-T1-player-09.trj	0.69	0.17
BrazilxVenezuela-T1-player-11.trj	0.59	0.17
ColombiaxUruguay-T1-player-09.trj	0.40	0.08
ColombiaxUruguay-T1-player-11.trj	1.36	2.79
ColombiaxUruguay-T2-player-10.trj	0.53	0.06
PeruxBolivia-T1-player-10.trj	0.66	0.22
PeruxBolivia-T1-player-11.trj	0.61	0.35
PeruxBolivia-T2-player-10.trj	1.20	0.24
Global	**0.73**	**0.48**

TABLE II
ERRORS AND STANDARD DEVIATIONS FOR THE COMPLETE TRAJECTORIES FOUND BY THE TRACKER USING APPEARANCE MODEL.

Track	μ	σ
BoliviaxColombia-T1-player-10.trj	0.57	0.11
BoliviaxColombia-T1-player-11.trj	0.68	0.16
BoliviaxColombia-T2-player-10.trj	0.80	0.29
BrazilxArgentina-T1-player-09.trj	0.58	0.07
BrazilxColombia-T2-player-11.trj	0.52	0.09
BrazilxColombia-T2-player-12.trj	0.72	0.33
BrazilxPeru-T1-player-09.trj	0.64	0.21
BrazilxVenezuela-T1-player-10.trj	0.30	0.11
BrazilxVenezuela-T1-player-11.trj	0.61	0.21
ColombiaxUruguay-T1-player-09.trj	0.36	0.04
ColombiaxUruguay-T2-player-10.trj	0.63	0.07
PeruxBolivia-T1-player-09.trj	0.81	0.29
PeruxBolivia-T1-player-10.trj	0.61	0.21
PeruxBolivia-T1-player-12.trj	0.58	0.18
PeruxBolivia-T2-player-09.trj	0.61	0.23
Global	**0.60**	**0.19**

D. Experiments Using the Appearance Model

When the tracker works just with de observation model presented in Section IV-B without the appearance model presented in Section IV-C, we can observe problems in some cases of confusions. Using the appearance model, we can find trajectories with errors and covariances less than the ones found without the appearance model. Figure 10 shows the best trajectory found using the system.

The results show us the improvement provided by the use of the appearance model that can be observed comparing Figures 7 and 8. This improvement reflects on the global values of error and covariance as the Table II shows. In the first test case (Table I) without the appearance model, we found 0.73 meters of accumulated average error while with an appearance model (Table II) we found 0.60 meters with a small covariance.

Using an appearance model, the tracker can solve some of problems discussed in the previous section. For instance, when two players are close in a dispute for the ball, their motion is similar and their positions are very close, but their appearance models are different. In this case, the observation part of the particle filter will analyze two different objects and can get the correct probabilities for the particles. Another situation happens when one player passes close to another. Without the appearance model, the tracker will be confused and will perceive just one player in this location. If the players are moving slowly, the observation model without the appearance model can merge trajectories. The appearance model helps the observation model to distinguish these players generating different values to different players. With this, the system can continue tracking the correct player finding good trajectories.

However, the use of appearance models does not solve all problems. In cases of celebration, the players are found very close for a long time and the detector on the image plane cannot detect all players involved in the celebration. In these cases, the tracker loses the specific information about the player tracked and fails. Basically, there are still two confusion situations: the first one occurs when two or more players appear together on image plane. During the game, it is common to find players of the same team together celebrating a goal. In this case, two or more players are merged into just one on the multimodal function, their appearances are very close and their dynamics are the same. The result of this situation is the fusion of the trajectories. The second confusion type occurs when the detector loses a player that is close to another one with the same appearance and the same motion dynamics. In these cases, the multimodal function loses one peak and causes the fusion of the corresponding trajectories.

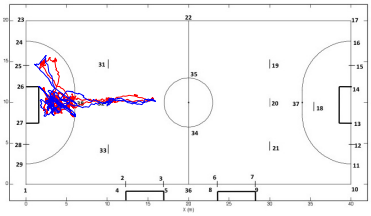

Fig. 9. Trajectory calculated by the tracker without appearance model (red) and the corresponding manual ground truth (blue). The trajectory in red correspond to the mean error in Figure 7.

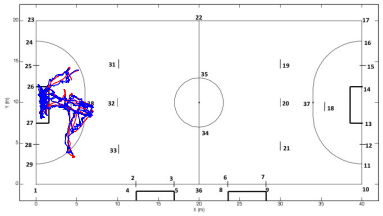

Fig. 10. Trajectory calculated by the tracker using the appearance model (red) and the corresponding manual ground truth (blue). The appearance in red corresponds to the minimum error in the Figure 8.

VI. Conclusion

This paper discussed a method to retrieve trajectories of indoor soccer players during the game, using a predictive filter to model the motion dynamics of objects of interest and a multi-camera observation system. The observation model consists of a multimodal function composed of a mixture of Gaussians and an appearance model used to weight Gaussians and strengthen the observation. The system uses a set of stationary cameras strategically placed around the court. Detections found in the images are combined, resulting in one multimodal probability function to represent the observation at the plane of court.

The main advantage of the proposed method is its automation, saving skilled labor during all the process, while most approaches are partially automated or assisted. Our method requires intervention only in the training phase, initialization and the homography matrices calculation, which are performed only once and from a single video sequence. On the other hand, it still does not represent a final solution towards indoor soccer player tracking given that there are still some situations in which the observation model is not able to deal with long and close coincident trajectories.

Moments of confusion occur when trajectories intersect or when they are coincident. In the first case, the proposed prediction system can successfully separate the trajectories and the tracking is not affected. In the second case, the trajectories merge into a single object, since their motion is similar, and this affects the prediction system. For this problem, we use adaptive appearance models to strengthen the observation.

In our tests, we can observe that the use of appearance models in conjunction with the multimodal function allows the particle filter to solve important confusion situations among players, mainly in cases with players of different teams. The method solves other situations when players of the same team are involved, but we still observe problems during celebrations. In these moments, some players of the same team appear together, and share the same or very similar motion dynamics leading the detector to fail and the tracker to lose the players involved.

As future work, we intend to focus on the problem of confusions to make an observation model that can deal with this problem. With a model that can be able to solve these situations we will have a completely automatic method to track indoor soccer players and then use trajectories to gather some important informations to the tactical and training staff of a team. Another possible extension is to detect situations relevant to the training staff based on ball tracking. As an example, we can mention the moments of passing and receiving the ball, right and wrong pass counting and shots on goal.

Acknowledgment

We would like to thank Microsoft Research, São Paulo Research Fundation – FAPESP(Grant 2010/05647-4) and CNPq(Grant 141054/2010-7) for the financial support.

References

[1] P. Figueroa, N. Leite, and R. B. M. L. Barros, "Tracking soccer players aiming their kinematical motion analysis," *CVIU*, vol. 101, no. 2, pp. 122–135, 2006.

[2] P. Figueroa, N. Leite, R. M. L. Barros, I. Cohen, and G. Medioni, "Tracking soccer players using the graph representation," in *ICPR*, Washington, DC, USA, 2004, pp. 787–790.

[3] S. Kasiri-Bidhendi and R. Safabakhsh, "Effective tracking of the players and ball in indoor soccer games in the presence of occlusion," in *ICC*, oct. 2009, pp. 524–529.

[4] K. Okuma, A. Taleghani, N. Freitas, J. Little, and D. Lowe, "A boosted particle filter: Multitarget detection and tracking," in *ECCV*, 2004, vol. 3021, pp. 28–39.

[5] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE CVPR*, vol. 1, 2001, pp. 511–518.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[7] S. M. Khan and M. Shah, "A multiview approach to tracking people in crowded scenes using a planar homography constraint," in *ECCV*, 2006.

[8] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE CVPR*, vol. 2, 1999, pp. 252–260.

[9] A. Alahi, Y. Boursier, L. Jacques, and P. Vandergheynst, "Sport player detection and tracking with a mixed network of planar and omnidirectional cameras," in *ICDSC*, 2009.

[10] J. Kang, I. Cohen, and G. Medioni, "Soccer player tracking across uncalibrated camera streams," in *IEEE VS-PETS*, 2003, pp. 172–179.

[11] W. B. Gevarter, *Robotics and Artificial Intelligence Applications Series: Overviews*. Business/Technology Books, 1984.

[12] D. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Prentice Hall, 2002.

[13] S. K. Goldenstein, "A gentle introduction to predictive filters," *RITA*, vol. 1, pp. 61–89, 2004.

[14] E. Trucco and A. Verri, *Introduction Technique for 3-D Computer Vision*. Prentice Hall, 1998.

[15] M. Isard and A. Blake, "Condensation – conditional density propagation forvisual tracking," *IJCV*, vol. 29, no. 1, pp. 5–28, 1998.

[16] W. Du and J. Piater, "Multi-camera people tracking by collaborative particle filters and principal axis-based integration," in *ACCV*, vol. Part I. Springer-Verlag, 2007, pp. 365–374.

[17] C. M. Bishop, *Pattern Recognition and Machine Learning*, M. Jordan, J. Kleinberg, and B. Schölkopf, Eds. Springer, 2006.