Pattern Recognition Letters 39 (2014) 21–30

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A multiple camera methodology for automatic localization and tracking of futsal players

Erikson Morais [a], Anselmo Ferreira [a], Sergio Augusto Cunha [b], Ricardo M.L. Barros [b], Anderson Rocha [a,*], Siome Goldenstein [a]

[a] Institute of Computing, University of Campinas (Unicamp), Cidade Universitária "Zeferino Vaz", Av. Albert Einstein, 1251, Campinas CEP 13083-852, SP, Brazil

[b] Department of Physical Education, University of Campinas (Unicamp), Cidade Universitária "Zeferino Vaz", Av. Érico Veríssimo, 701, Cx. Postal 6134, Campinas CEP 13083-851, SP, Brazil

ARTICLE INFO

Article history:
Available online 20 September 2013

Communicated by S. Sarkar

Keywords:
Futsal player tracking
Futsal particle filter
Data fusion
Multiple camera observation model

ABSTRACT

There is a growing scientific interest in the study of tactical and physical attributes in futsal. These analyses use the players' motion, generally obtained with visual tracking, a task still not fully automated and that requires laborious guidance. In this regard, this work introduces an automatic procedure for estimating the positions of futsal players as probability distributions using multiple cameras and particle filters, reducing the need for human intervention during the process. In our framework, each player position is defined as a non-parametric distribution, which we track with the aid of particle filters. At every frame, we create the observation model by combining information from multiple cameras: a multimodal probability distribution function in court plane describing the likely players' positions. In order to decrease human intervention, we address the confusion between players during tracking using an appearance model to change and update the observation function. The experiments carried out reveal tracking errors below 70 cm and enforce the method's potential for aiding sports teams in different technical aspects.

1. Introduction

The dynamic nature and narrow space for futsal playing[1] have attracted the interest of researchers and practioners toward this sport. These two features combined take the tactical behavior and the physical conditioning of the players to the extreme, transforming futsal in a special target of technical and scientific interest, showing several applications ranging from tactical, to physical and to physiological investigations (Figueroa et al., 2006).

The players' trajectories allow the tactical staff to account for the positioning effectiveness in the game, while the technical workforce, which watch over the players physical conditioning, studies the players' trajectories assessing diverse data such as speed, acceleration peaks and distance covered to establish the physical activities. Finally, these analyzes can be complemented by physiological evaluations relying on the trajectories for assessing stress levels.

The use of global positioning system (GPS) has been reported for detecting positions in soccer (Mendez-Villanueva et al., 2013). However, it still cannot be used in futsal because this sport is practiced mostly in closed environments, which interfere in GPS signals. Additionally, GPS has large errors (in meters) in the position detection on outdoor pitches and consequently represents a great problem for estimating speeds and accelerations (Gray et al., 2010). Sometimes, the GPS errors (in the case of indoors applications, for instance) can be anything from 2–20 m. n automated solution capable of retrieving the trajectories of players directly from a game video-footage will surely aid sports teams and enhance all of aforementioned aspects (Figueroa et al., 2004).

Nowadays, video cameras and recording media are getting more affordable, and now acquiring multiple simultaneous recordings of a futsal game with high-resolution cameras is more feasible than ever. This allows the teams to review not only their own motion, but also important behavioral characteristics of their opponents. This is yet another of the advantages of a purely passive system, that does not require the players to wear any type of measuring device – you can also inspect your opponents.

The underlying intuition is that multiple cameras positioned around the court provide us with enough redundancies to find the players whereabouts, so that multiple observations and data

* Corresponding author. Tel.: +55 19 3521 5854; fax +55 19 3521 5838.

E-mail addresses: emorais@ic.unicamp.br (E. Morais), ra023169@ic.unicamp.br (A. Ferreira), scunha@fef.unicamp.br (S.A. Cunha), ricardo@fef.unicamp.br (R.M.L. Barros), anderson.rocha@ic.unicamp.br (A. Rocha), siome@ic.unicamp.br (S. Goldenstein).

[1] Futsal is a variant of association football, or soccer, played on a smaller, usually indoors, court. The term futsal is a portmanteau of the Portuguese *futebol de salão* (directly translated as "hall football"). Futsal is the sport official name according to the *Fédération Internationale de Football Association (FIFA)*, and is preferred over the term "indoor soccer."

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.09.007

fusion allow an overall accuracy improvement of a player's position estimation. Computer vision methods in futsal also have important scientific challenges and wide applicability: the dynamic nature of a collective game extends naturally to surveillance applications. Most methods for player localization and tracking currently available in the state of the art are not totally automated and have focused on the tracking steps (Figueroa et al., 2004, 2006; Alahi et al., 2009; Kang et al., 2003).

This paper first introduces a method for automatically estimating the probability distribution function of futsal players' positions using multiple cameras, a first step to tracking. Next, we discuss a procedure for estimating the trajectories of futsal players throughout a match, with human intervention reduced to a simple marking of players of interest at the very beginning of the process.

We represent a player's location as a non-parametric distribution, using particle filters. We use a people-detection algorithm to detect players in each camera, and, with back projection, we map their position onto court coordinates. Since we measure each detector's error, we can calculate the observation to the particle filter as a 2D multimodal probability distribution function (in court plane) that represents the potential localization of players. In order to solve part of the confusions that occur during tracking, each player has its own particle filter with an observation function that takes into account an incremental and adaptive appearance model – so that each player's filter will see a different observation.

This paper extends upon two previous conference papers of our group (de Morais et al., 2012; Morais et al., 2012). The main differences are: (1) better formalization of the proposed methods; (2) detailing of the proposed method for the observation model and multiple camera fusion; (3) inclusion of an experimentation round regarding the proposed method for multiple camera fusion and (4) the improvement of the experimentation regarding the second contribution: the tracking method based on particle filters and an appearance model-based method.

2. Related work

The detection and tracking of futsal players have received special attention from many researchers in the past few years. Figueroa et al. (2004, 2006), for instance, have modeled the problem of tracking players in a game by means of a graph shortest path problem. Each node in such a graph represents detected blobs in video footages of games. The edges then connect blobs in different frames. In their approach, one blob in a frame connects to all blobs in the subsequent frame. After selecting one player to be tracked over the video sequence, the proposed method calculates the shortest path to a final node and maps it onto the coordinates of the game court.

Many methods in the literature share a common step called observation in which the tracked objects are detected. There are different approaches for implementing such observation step in practice and one common solution is to separate the background and the target object. Figueroa et al. (2004, 2006) have separated the objects and background by averaging a set of frames periodically updated. Kasiri-Bidhendi et al. (2009) have used color histograms for modeling the background. The authors have observed that since most of the pixels in a video frame refer to regions of the court, there are some dominant colors in the histogram that represent the area of interest that eliminate the other regions putting the players in evidence. Figueroa et al. (2006) have employed segmentation techniques based on the difference between image sequences and a representation of the background to put the players in evidence. Okuma et al. (2004), have focused on tracking hockey players by means of particle filters in models of players learned through an Adaboost ensemble technique. Similarly, Barros et al. (2011) have used a cascaded detector associated with morphological segmentation for tracking handball players.

We can generalize the problem of detecting futsal players in an image to the problem of finding people. In this regard, we refer to the groundbreaking work presented by Viola and Jones (2001), for instance, who have introduced a very efficient solution for detecting faces in images with straightforward extensions to many other object localization problems such the ones involving pedestrians and cars.

Felzenszwalb et al. (2010), have approached the people detection problem representing objects of high variability by means of different models of multi-scale deformable parts. The proposed method requires only a few reference set of images for training. Each part in the model captures appearance properties of an object whilst the deformable configuration links the parts to the pairs. Following a different strategy, Khan et al. (2006) have extended upon the work of Stauffer and Grimson (1999) to separate object and background, and then used homography to determine the likely positions of people in a 3D scene.

Alahi et al. (2009) have focused on the problem of detecting basketball players by means of shortest distance calculations between detections in neighboring frames. Kang et al. (2003), in turn, have aimed at tracking soccer players in the field employing stationary cameras. Similar to previous approaches in the literature, the authors have used background separation to highlight potential blobs representing players. Using homography, the authors map 2D positions in the image plane to positions in the virtual plane representing the game field and then employ a Kalman filter to perform the tracking. One substantial problem of the proposed method is the confusion generated by close partially occluded objects related to the camera perspective or simply physical proximity of the objects. In situations involving high proximity between the tracked objects, the detection might fail and consider them as one. Furthermore, the Kalman filter is not capable of dealing with multiple hypotheses resulting from the proximity of players, a common situation in indoor games (e.g., futsal).

Our work is focused on futsal players tracking and is different from the aforementioned works in the following aspects: first, we propose an observation model that combines data from different cameras installed at the game site, each one with its own coordinate system, and project them onto a virtual plane which we use to designate the futsal court; second, we use non-parametric distributions to represent the players positions and perform the automatic tracking of the observed players based on particle filters; finally, we use an appearance model solution to deal with confusion situations, in which players are near one another and the tracking cannot work properly. The basic concepts necessary to understand the proposed method are given in Section 3.

3. Basic concepts

We define visual tracking as the problem of analyzing an image sequence for describing Gevarter et al. (1984) or reasoning Forsyth et al. (2002) about the motion of one or more objects in a scene. Prior to the object tracking task, the object itself must be detected. Some methods use the whole image for finding specific objects (Figueroa et al., 2006; Kang et al., 2003). In this sense, for each frame in a video sequence blobs are detected, even if not being the target of the tracking.

Recently, some authors have employed predictive filters which, in turn, rely on a model for the object motion called *dynamics*, for estimating the state of a tracked object in an immediate frame of a video sequence. A local observation using only some input frames instead of all of the input video is responsible for adjusting the prediction. In this regard, we highlight two widely used filters which we shall explain in the next section: the Kalman and the particle filters.

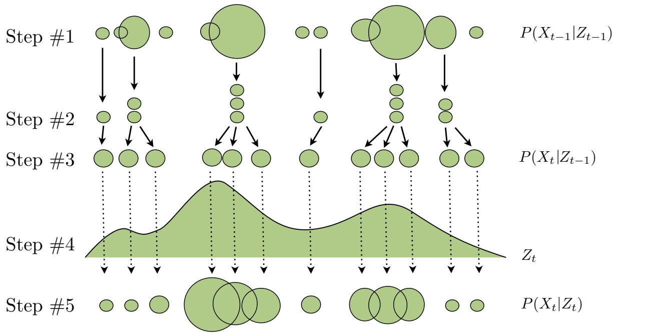


Fig. 1. Steps associated with a particle filter approach. (1) The particles associated with the *a posteriori* function $t-1$ are (2) sampled; and (3) propagated leading to the *a priori* function $P(X_t|Z_{t-1})$. Then, upon (4) observation in t, we have the (5) *a posteriori* function $P(X_t|Z_t)$ in t.

3.1. Kalman and particle filters

Kalman filters use Gaussian functions for guessing the next state of an object, also known as tracking. A Gaussian representation is a simple parametric function (Trucco et al., 1998; Goldenstein, 2004). Kalman filters estimate states that could not be explicitly detected, such as when the object of interest is partially occluded by other components of the scene. A well known limitation of Kalman filters is the unimodal nature of the Gaussian function (i.e., the filter does not deal with multiple hypotheses at the same time). A proper methodology to measure the observation step (S. Goldenstein et al., 2003) can yield exceptional results, and improve the results of a Kalman Filter as much as the use of a proper motion model (Goldenstein et al., 2004).

Particle filters (Isard and Blake, 1998; Du et al., 2007), on the other hand, represent the object's states as non-parametric distributions and rely on weighted samples called particles, which enable the modeling of multimodal and unknown functions. Each particle x_t is linked to a weight W_t denoting the value of the function at the point determined by the particle. Such particles approximate the desired distribution function, and the more particles we have, the more accurate is the representation.

More specifically, given a particle filter, a state represents the object of interest and may have the information of interest represented in a vectorial form. In the case of tracking objects, we might be interested in the position and speed of such object. Therefore, we can represent each state with a vector containing the (x,y) position of the object as well as its speed. The *a priori* probability function denotes a rough calculation of the value of what should happen for the next frame given the information (dynamics) we have about its current state.

Each particle has a step called *observation* whose objective is to check its representation in a given frame. Upon observation, the updated weights form the *a posteriori* probability function $P(x_t|Z_t)$ which represents the updated probability of a particle x_t given an observation Z_t.

For estimating the next state, we simply sample the weighted particles with replacement, preserving the size of the set of particles. After sampling, there is the *prediction* step responsible for adjusting each new particle's weight based on the motion dynamics and a random error that models the error in the process. The function represented by the set of new particles is then treated as the new *a priori* function $P(X_{t+1}|Z_t)$ for the next frame.

Fig. 1 illustrates one particle filter's cycle in which we have, in the first layer, some particles represented by ellipsis of various sizes. The size denotes the weight of a particle. The second layer illustrates the result of the sampling process which can lead to repeated particles. Upon sampling, the weights of the particles lose their meaning and new measurements are necessary as the third layer shows. In this layer, the particles' weights are adjusted according to the used dynamics and a random error applied to the sorted particles. Upon adjustments, we have an estimation for the next frame in the sequence. The fourth layer shows the function representing the updated particles. In this function, the height denotes the weight of the measurement at a given point. The fifth layers shows the *a posteriori* probability function as the result of the measurement step. In this step, the particles are properly weighted and prepared for the next iteration. We repeat this sequence of steps for every frame in a video sequence.

3.2. Viola and Jones detector

Viola and Jones (2001) introduced a method based on Haar-filters and Adaboosting algorithm to detect objects in images. Although the authors have focused on face detection, the extension of the detector to other types of objects (e.g., people) is straightforward. For that, the requirement is to obtain enough training examples representing people versus non-people images for a new training procedure.

The detector relies on simple features known as *Haar-like features* each of which is calculated using sums and differences. Such features are formed by the difference of the sum of pixel values in regions with different colors. For a two-rectangle feature, for example, the sum of pixels in a region of one color is subtracted to the sum of the pixels in a region with a different color.

The method uses sliding windows of size 24 × 24 pixels to detect faces. Within each window there are more than 180,000 possible features with different sizes and orientations. The authors propose the concept of integral images for speeding up the calculation process. For an image i, its integral image has the same dimensions and each entry represents the sum of every pixel to the left and above the current pixel position

$$ii(x,y) = \sum_{x' \leqslant x, y' \leqslant y} i(x',y'). \tag{1}$$

We can calculate the integral image (ii) with only one pass over each image pixel. With this integral image, we can calculate the summation of a given rectangle feature with only four accesses on the integral image.

The authors consider the Haar-like features as weak classifiers and use the Adaboost algorithm to select the best features among the more than 180,000 possible ones. For each weak classifier, the Adaboosting training algorithm determines the best threshold that results the lowest classification error for object and non-object classes. A weak classifier $h_j(x)$ is given by a feature f_j, a threshold θ_j and a polarization $(+/-)$ p_j

$$h_j(x) = \begin{cases} 1 & \text{if } p_j f_j(x) < p_j\theta_j, \\ 0 & \text{otherwise}. \end{cases} \tag{2}$$

where x is a 24 × 24-pixel image region. For each weak classifier selection, the most appropriate feature is selected and receives a weight associated with each training sample evaluated.

The authors combined several weak classifiers to build a strong classification procedure by means of a cascade setup. In a cascade classifier scheme, the outcome of one classifier is the input the next one giving rise to a strong classifier. Once a sub image is submitted to the cascade, each weak classifier classifies such sub image to detect whether or not it has the object of interest. If the response is negative, no further classification is performed. If a sub image passes over all the classifiers, it is tagged as having the object of interest (e.g., a player).

In Section 4, we explain how all of these concepts are used in our proposed method.

4. Localization and tracking

Here, we introduce a probabilistic method for finding and tracking futsal players using particle filters after measuring the distribution of the observation. Our method trains an object detector to look for futsal players on each of the cameras, and uses homography to map camera positions to the coordinates on a virtual plane representing the court coordinates. We assign a particle filter for each player that uses, for the observation, a multimodal and bidirectional probability density function assembled from the object detector results from all cameras and weighted by a part-based appearance model.

4.1. Motion dynamics

Each particle of a filter represents a possible state, which we model in our framework as position and velocity in court coordinates, $S(x, y, v_x, v_y)$. We use uniform motion to describe the basic object motion, with an extra Gaussian noise to represent the uncertainty of this simple model. Eq. (3) presents its standard mathematical representation,

$$S_t = S_{t-1} + V * \Delta(t) + e \tag{3}$$

and Eq. (4) in matrix form,

$$S_t = \begin{bmatrix} 1 & 0 & \frac{1}{30} & 0 \\ 0 & 1 & 0 & \frac{1}{30} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} x \\ y \\ v_x \\ v_y \end{bmatrix}_{t-1} + \begin{bmatrix} 0 \\ 0 \\ e_x \\ e_y \end{bmatrix}, \tag{4}$$

where S_t is the state at time t and V is a velocity vector containing values in x and y axes, respectively. In this motion representation, $\Delta(t)$ is the time variation or, in this case, the video frame rate (our videos have 30 frames per second), and $e(0, 0, e_x, e_y)$ represents the estimation error.

4.2. Observation function

In a particle filter, the observation function weights the particles prior to the resampling stage. We use a 3-stage methodology (see Fig. 2):

1. **Stage** #1 – Player detection in image coordinates. In this paper, we use the Viola and Jones (2001) detector.
2. **Stage** #2 – Detection projection onto court coordinates. We use homography to project the detector's results onto court coordinates.
3. **Stage** #3 – Multi-camera fusion. Each individual projection becomes a bidirectional Gaussian, with measured covariance, in court coordinates. They are all combined, weighted by a part-based appearance model, to form a multi-modal distribution. This probability distribution represents the localization of a given player in the court.

*4.2.1. Stage #1 – **Player detection in image coordinates***

As we discussed in Section 2, some relevant work in the literature use color information to detect objects of interest in images such as the work of Juang et al. (2009). Other approaches use background subtraction methods to find blobs corresponding to objects of interest such as the works of Miura et al. (2008) and Figueroa et al. (2006).

In this paper, we first use an object detector to find players in each camera, independently. Using a selected portion of the videos for training (as described in Section 5.2), we train a specialized detector to our problem instead of the traditional Viola and Jones detector. For that, we collect 15,932 positive samples (images containing examples of players) and 18,400 negative samples (background regions with no players) of size 25 × 60 pixels each. The samples consider examples in the different camera viewpoints and were drawn in three different scales, therefore extending the original formulation of Viola and Jones. The multiscale approach allowed the detector to better capture players in different viewpoints and, consequently, with different scales according to their positions in the court plane.

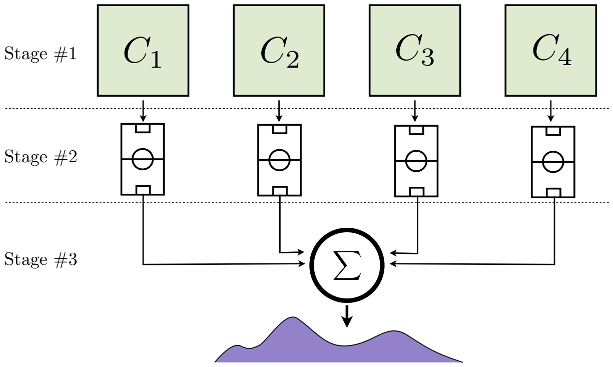


Fig. 2. Three stages for the observation function. Stage #1 detects players in each camera's image plane using an object detector. Stage #2 uses homographies to project the observations onto court coordinates. Stage #3 combines all projections to form a multimodal two-dimensional probability function – the probabilistic representation of the player's position in court coordinates.

With the two sets, we train the detector using the Adaboost approach as described in the Viola and Jones original methodology (Viola and Jones, 2001). Such approach then finds the best set of features specialized to the problem of detecting futsal players. Although we have focused on the Viola and Jones detector methodology for this part of our work, other detectors would be adequate as well. Future work includes the use of alternative detectors such as the one proposed by Felzenszwalb et al. (2010).

*4.2.2. Stage #2 – **Projection onto court coordinates***

We create a virtual court plane to represent the actual 2D plane where the players move. A homography maps image coordinates of points known to be on this virtual plane (such as where feet touch the floor) to their virtual court coordinates. The detectors find rectangles and we use the mean point of the rectangle's basis as an approximation of the projection of the center of the player's center of gravity onto the court.

We use the points with known coordinates (landmarks of the court) to estimate the homography matrix,

$$p_b = H_{ab} p_a, \tag{5}$$

considering p_a a point on image plane and p_b the correspondent point on court plane. To estimate the homography matrix H, we rewrite Eq. (5) as

$$\begin{bmatrix} x_{p_b} \\ y_{p_b} \\ z_{p_b} \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} = \begin{bmatrix} x_{p_a} \\ y_{p_a} \\ z_{p_a} \end{bmatrix} \iff p_b = H p_a. \tag{6}$$

In homogeneous coordinates, $x'_{p_a} = x_{p_a}/z_{p_a}$ e $y'_{p_a} = y_{p_a}/z_{p_a}$,

$$x'_{p_a} = \frac{h_{11}x_{p_a} + h_{12}y_{p_a} + h_{13}z_{p_a}}{h_{31}x_{p_a} + h_{32}y_{p_a} + h_{33}z_{p_a}}, , \tag{7}$$

$$y'_{p_a} = \frac{h_{21}x_{p_a} + h_{22}y_{p_a} + h_{23}z_{p_a}}{h_{31}x_{p_a} + h_{32}y_{p_a} + h_{33}z_{p_a}}. \tag{8}$$

Without loss of generality, we consider $z_{p_a} = 1$ and reorganize the equation according to

$$a_x^T\vec{h} = 0, \tag{9}$$

$$a_y^T\vec{h} = 0, \tag{10}$$

where

$$\vec{h} = (h_{11}, h_{12}, h_{13}, h_{21}, h_{22}, h_{23}, h_{31}, h_{32}, h_{33})^T,$$

$$\vec{a}_x = \left(-x_{p_a}, -y_{p_a}, -1, 0, 0, 0, x'_{p_b}x_{p_a}, x'_{p_b}y_{p_a}, x'_{p_b}\right)^T,$$

$$\vec{a}_y = \left(0, 0, 0, -x_{p_a}, -y_{p_a}, -1, y'_{p_b}x_{p_a}, y'_{p_b}y_{p_a}, y'_{p_b}\right)^T.$$

With four pairs of corresponding points between planes, we can solve the linear system

$$A\vec{h} = \vec{0}, \tag{11}$$

where

$$A = \begin{pmatrix} \vec{a}_{x_1}^T \\ \vec{a}_{y_2}^T \\ \vdots \\ \vec{a}_{x_N}^T \\ \vec{a}_{y_N}^T \end{pmatrix}. \tag{12}$$

We use a frame of the video to mark manually the correspondences between points and calculate the homography matrix of each camera.

4.2.3. *Stage #3 – Multi-camera fusion*

We replace each of the detection points from Stage #2 with an uncertainty representation of the location, a two-dimensional Gaussian. We combine these individual functions to find a mixture of Gaussians corresponding to the projected points by multiple cameras. Each of these Gaussians represents the uncertainty in the player detection stage projected in virtual court coordinates. The mean will be the detection plus a bias, and both the bias and covariance are learned for each camera.

We manually annotate a small training set of the videos with the actual player's positions, and use them to estimate the bias and covariance of the detection after projection onto court coordinates. Because we know each player's position in the training, we compare each detection with its ground-truth, estimated by the shortest distance, in the annotated set. To avoid cases where multiple players are close, leading to confusion in this ground-truth, we only consider situations where the nearest annotated point is closer than 2 m (L_1) and the second closest annotated point is farther than 3 m (L_2). We calculate the mean error of projection, the *bias*, and the covariance matrix for each camera.

Each Gaussian $\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$ is a component of a complex function representing the probability distribution function of players in the court plane. Even though we use a linear combination of Gaussians, the family of distributions is quite complex (Bishop et al., 2006). In our case, we have a set of detections with the same importance to find players on the court plane.

Initially, we consider equally probable coefficients $\pi_k = \frac{1}{K}$ for the Gaussian mixture model. As the game progresses, we update π_k with the help of the appearance model described in Section 4.3. The value of π_k is different for each player, since they all have different appearance models,

$$P(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k). \tag{13}$$

Eq. (13) measures the likelihood of finding a player at any given point of the court, and is a natural seed for the particle filter's observation step. Fig. 3 illustrate the function's behavior on a frame of a real game.

4.3. *Appearance model to reinforce the multimodal function*

Section 4.2.3 discussed the multimodal function for finding a player in a real-world coordinate. However, it is not enough to differentiate the players given that all Gaussians are weighted equally in the mixture.

Each detection in camera coordinates gives rise to a Gaussian and corresponds to one detected rectangle in the camera's vantage point. We use such rectangle for calculating the appearance for a detected player and compare it to an appearance model for that camera resulting in a similarity value. We then use such value as the weight π_k in Eq. (13) for the corresponding Gaussian. Hence, the most similar detection regarding the appearance model in the correspondent camera results in the most important Gaussian in the mixture.

The appearance of a detected player is determined by color and gradient information. The color information intends to differentiate players of different teams while the gradient information helps to capture details of specific players. Given that the shape of a player is relatively constant across consecutive frames, histograms of gradients are a natural choice for capturing the line distribution, which remains slightly constant, in the detected rectangle region.

We calculate the histogram of gradients as follows:

1. Obtain the x and y derivatives for each sub-image through Sobel algorithm using the kernels $[-1, 0, 1]$ and $[-1, 0, 1]^T$.
2. Obtain the gradient magnitude Gr and the gradient orientation A through dx and dy for each pixel.
3. Calculate a 15-bin histogram for Gr and another for A characterizing the gradient information of each sub-region.
4. Calculate a 255-bin histogram for each color channel characterizing the color information of each sub-image.
5. Normalize all the histograms.

Fig. 4 shows one player as characterized by the gradient and color histograms. We split the sub-image containing the player

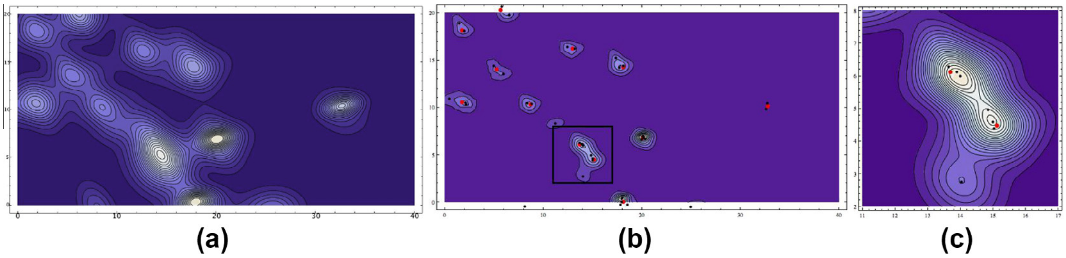


Fig. 3. One of the stages of the particle filter and the observation of the tracking. Previously trained Gaussian functions substitute the multi-camera projections. In the end, a 2D function gives the probability of detecting a player in a specific position. (a) and (b) show two different sets of parameters for the observation construction. (c) shows a zoom of (b) in which the selected region depicts the probability mass of two players. Fig. 5(b) shows the set of frames used to generate the probability function.

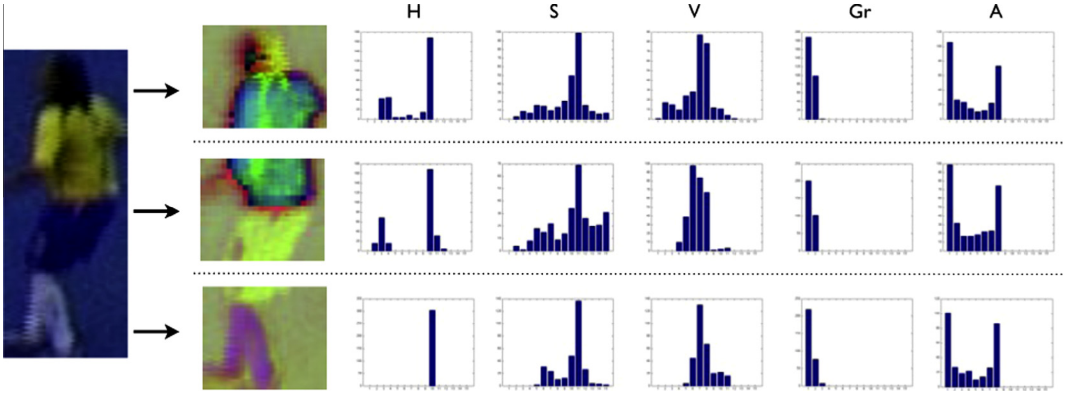


Fig. 4. One player's appearance model with the three regions of interest: top, middle and bottom. Each region contains five histograms: three color-based (one per H, S, and V color channels), and two gradient-based (gradient's magnitudes (Gr) and gradient's orientations (A)).

into three regions: top, middle, and bottom. The rationale is that these regions capture key information for differentiating players as the top region encompasses the face of a player and the middle and bottom parts comprise the shirts and the shorts, respectively. The regions for shirts and shorts play an important role for identifying the teams. In the end, each region gives rise to five histograms: two gradient-based (magnitude and orientation), and three color-based (one for each image's color channel under the HSV color-space).

For comparing two appearance models $S(AP_1, AP_2)$, we use histogram correlation. Taking one appearance model as a list of 15 histograms (three regions $\times$ five histograms per region) and AP^i as i-th histogram in this list, the correlation is given by

$$S(AP_1, AP_2) = \prod_{i}^{15} Corr(AP_1^i, AP_2^i). \tag{14}$$

Following this model, each player's appearance model needs to be compared to other models in the corresponding camera view for finding the π_k weights measuring the importance of such Gaussian for representing the positions of the corresponding detected player. In addition, previously determined good appearances define the appearance model in the camera view. Hence, π_k may be defined as the maximal similarity between the detected appearance (AP_d) in the present frame and all appearances in the model (AP_m):

$$\pi_k = \max\{S(AP_d, AP_m)\}. \tag{15}$$

Re-ordering Eq. (13), we obtain:

$$P(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k). \tag{16}$$

In order to keep a good representation of the object, we continually update this model by replacing the appearance of a Candidate to be Removed (CR) by a new good appearance detected, considered a Candidate to be Included (CI).

When comparing a model and other appearances, we update a history of use for signaling the last used appearance which will be used as the Candidate to be Removed. On the other hand, a Candidate to be Included will be the one that has the highest similarity with the current model during the observation. As the model keeps record of the information about the used appearances, the system "knows" what region leads to the highest weight during the observation step and uses the corresponding appearance as a CI. If one CI appearance has a similarity greater than a previously defined threshold, it replaces the appearance CR in that model. The observation step of the particle filter is responsible for maintaining such information.

For each new filter's iteration, we must weight the set of particles. Therefore, we use the appearance model information (and corresponding cutouts) to weight the Gaussian mixture corresponding to the multimodal function. For that, we compare the appearance corresponding to each Gaussian to the appearance model of the corresponding camera, maintained by the tracker. Therefore, the observation model used by the filter has N appearance models for the observed object, one for each camera. As a result, we have a multimodal function whose highest peaks are in the regions matching the object tracked and thus we are able to identify the proximity of players while emphasizing the correct object.

5. Experiments and validation

In this section, we present the experiments and validation protocol we used to evaluate the effectiveness of the proposed method.

5.1. Data

In the experiments, we consider a collection of seven Full-HD futsal games recorded using four stationary cameras, each at 30 frames per second, set around the court as Fig. 5 shows. Each game

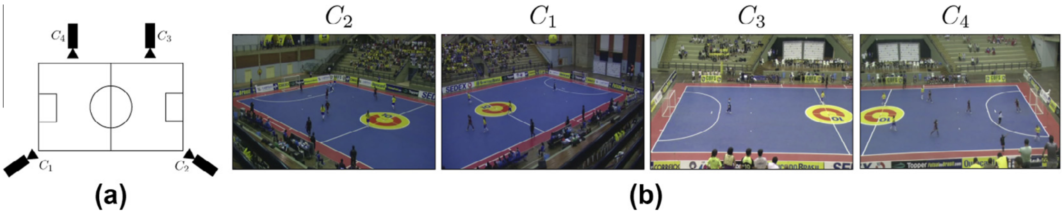


Fig. 5. (a) Camera setup in the court considering an overlap of fields of view. Each camera focuses on one half of the court in such a way each player is observed by at least two cameras. (b) Some exemplary frames according to the camera positioning.

Table 1
Average errors and standard deviations (in meters) found for multiple cameras.

Game	Period	Camera	Proposed method		Projection-only	
			μ	σ	μ	σ
Bolvia × Colmbia	T-01	1	0.58	0.13	0.73	0.16
		2	0.68	0.19	0.78	0.20
		3	0.75	0.24	0.88	0.21
		4	0.53	0.14	0.65	0.15
	T-02	1	0.65	0.21	0.77	0.21
		2	0.76	0.18	0.91	0.19
		3	0.54	0.13	0.72	0.13
		4	0.76	0.20	0.91	0.20
Brazil × Argentina	T-01	1	0.71	0.21	0.83	0.20
		2	0.74	0.27	0.79	0.27
		3	0.72	0.24	0.81	0.24
		4	0.55	0.16	0.63	0.15
	T-02	1	0.84	0.27	0.91	0.25
		2	0.63	0.19	0.68	0.19
		3	0.82	0.24	0.90	0.23
		4	0.79	0.25	0.89	0.23
Brazil × Colombia	T-01	1	0.78	0.21	0.82	0.20
		2	0.81	0.23	0.90	0.23
		3	0.78	0.25	0.90	0.22
		4	0.80	0.21	0.84	0.21
	T-02	1	0.81	0.22	0.87	0.19
		2	0.77	0.21	0.81	0.19
		3	0.88	0.27	1.00	0.26
		4	0.67	0.19	0.79	0.19
Brazil × Peru	T-01	1	1.10	0.35	1.12	0.32
		2	0.66	0.19	0.73	0.20
		3	0.64	0.20	0.74	0.19
		4	1.06	0.36	1.10	0.30
	T-02	1	0.97	0.25	1.03	0.22
		2	0.89	0.23	0.93	0.22
		3	0.98	0.23	1.01	0.21
		4	0.91	0.22	0.96	0.19

comprises two periods. For easiness, we treat each period as an individual game totaling 14 games recorded by four cameras. The games were recorded during the 2009 South American Women's Futsal Championship.

5.2. Training and calibration of parameters

As we pointed out previously, the particle filter observation method has two training stages: one for detecting players in image coordinates and another for calibrating the required parameters for creating the multimodal function representing the court's plane. Therefore, we separated one period of the games (one period of a game containing the recordings of the four considered cameras). We used the remaining 13 games for testing.

First, we train the detector in image coordinates (c.f., Section 4.2.1). For determining multimodal function parameters, we compared the image plane detections projected onto the court's plane with real markings manually annotated. Then, we calculated the average projection errors and their covariances for each camera for calibrating the parameters used in the multimodal function representing the court's plane, as Section 4.2.3 presented.

For the two training steps, we used one (out of 14) videos that was annotated manually in its full extent. For testing, we considered only the first two minutes of each game as it was too time-consuming to manually annotate the players positions on the court's plane coordinates for ground-truth.

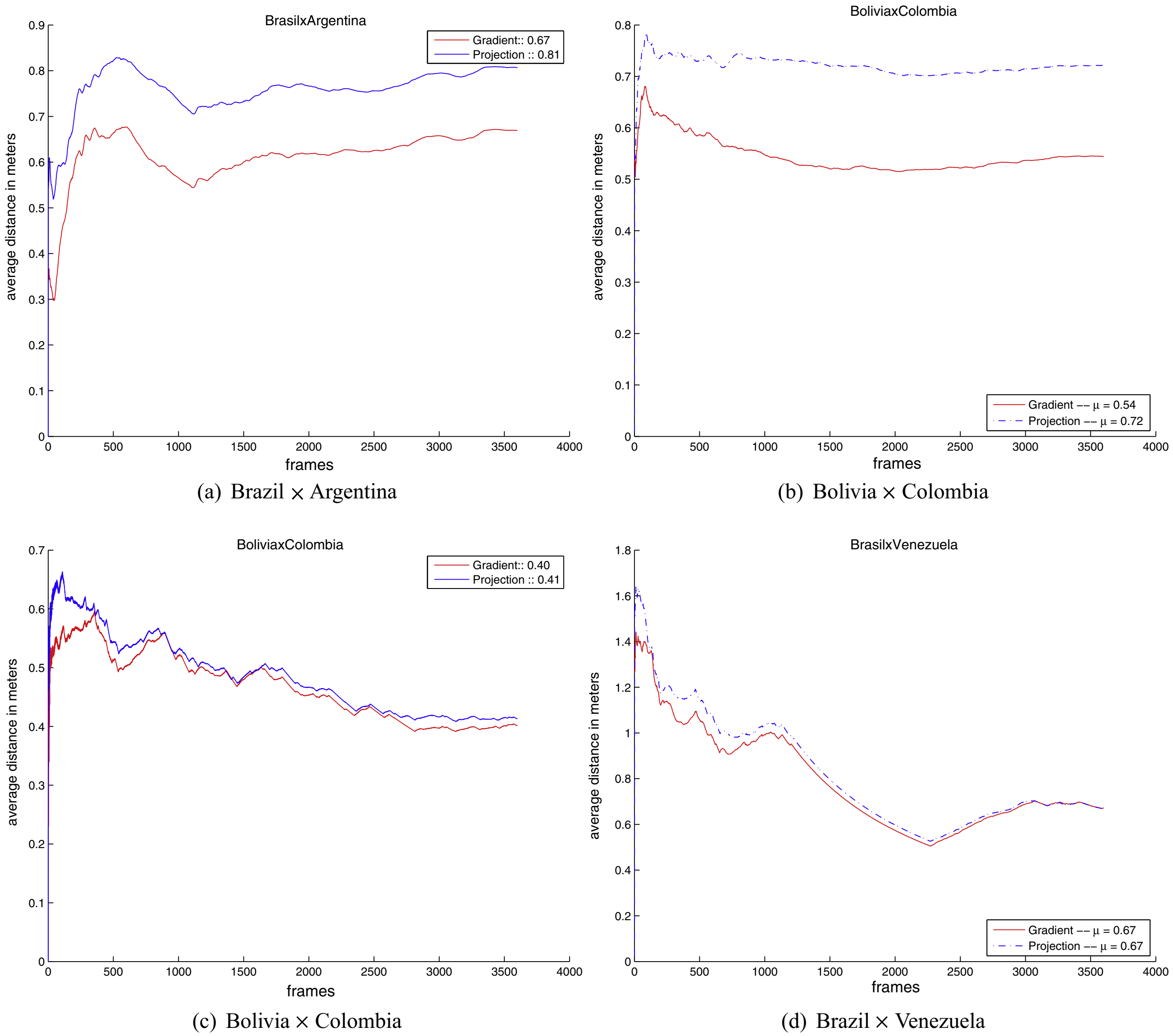
As previously mentioned, for an accurate operation of the particle filter, we need to start the particles with the initial positions of the players whom we want to track. Hence, for each player of interest, we mark its position in the first frame of each video. Here, we are interested in the 10 players on the court as well as the two referees totaling 12 objects of interest and for each of them, we start a filter with 500 particles. Thereafter, for each frame of the sequence, we need to detect the players prior to the tracking. Hence, the first round of results shows the effectiveness of the proposed method for combining the different camera views toward a more robust observation model (detection) of the players.

5.3. Experiments for localizing the players

To validate the proposed method for combining detections from multiple cameras, we compare this approach to the detection on each isolated camera using the standard Viola and Jones (2001) with no fusion. For each detection, we calculate the Euclidean distance to the annotation representing the real player's position. The smaller the distance the better the detection.

When using only the detectors separately, we project the detections in the image plane of one camera onto the world coordinates and each projection is treated independently. When using the proposed approach, we project all the detections from the four different cameras onto the world plane and the points are combined using the proposed multimodal function and gradient ascent method. In both cases, each point in the end is compared to the closest annotated point for determining the detection error.

The distance between an actual position and a detection measures the estimation error and it is represented in meters. Table 1 shows the average errors and standard deviations (in meters) found for each camera in four games while Fig. 6(a)–(d) show the two best and two worst cases when comparing both detection approaches (with and without multiple camera fusion). We also show the average error in each testing frame. We calculate the



(a) Brazil × Argentina

(b) Bolivia × Colombia

(c) Bolivia × Colombia

(d) Brazil × Venezuela

Fig. 6. Estimation average error. The curve shows the average error across the concatenated frames of different test video sequences considering one camera. (a) and (b) show cases in which our approach is significantly better than the approach with no fusion. (c) and (d) show cases in which both detection approaches are not statistically different.

error for each camera separately given that the projection errors are different for each camera as shown in Table 1.

As we can see, the obtained results show the potential of the proposed approach as it reduces the error of the detected position of players, representing an interesting start point for further tasks such as players tracking, as we discuss in the next two rounds of experiments (Sections 5.4 and 5.5).

5.4. Experimental results for tracking: no appearance model

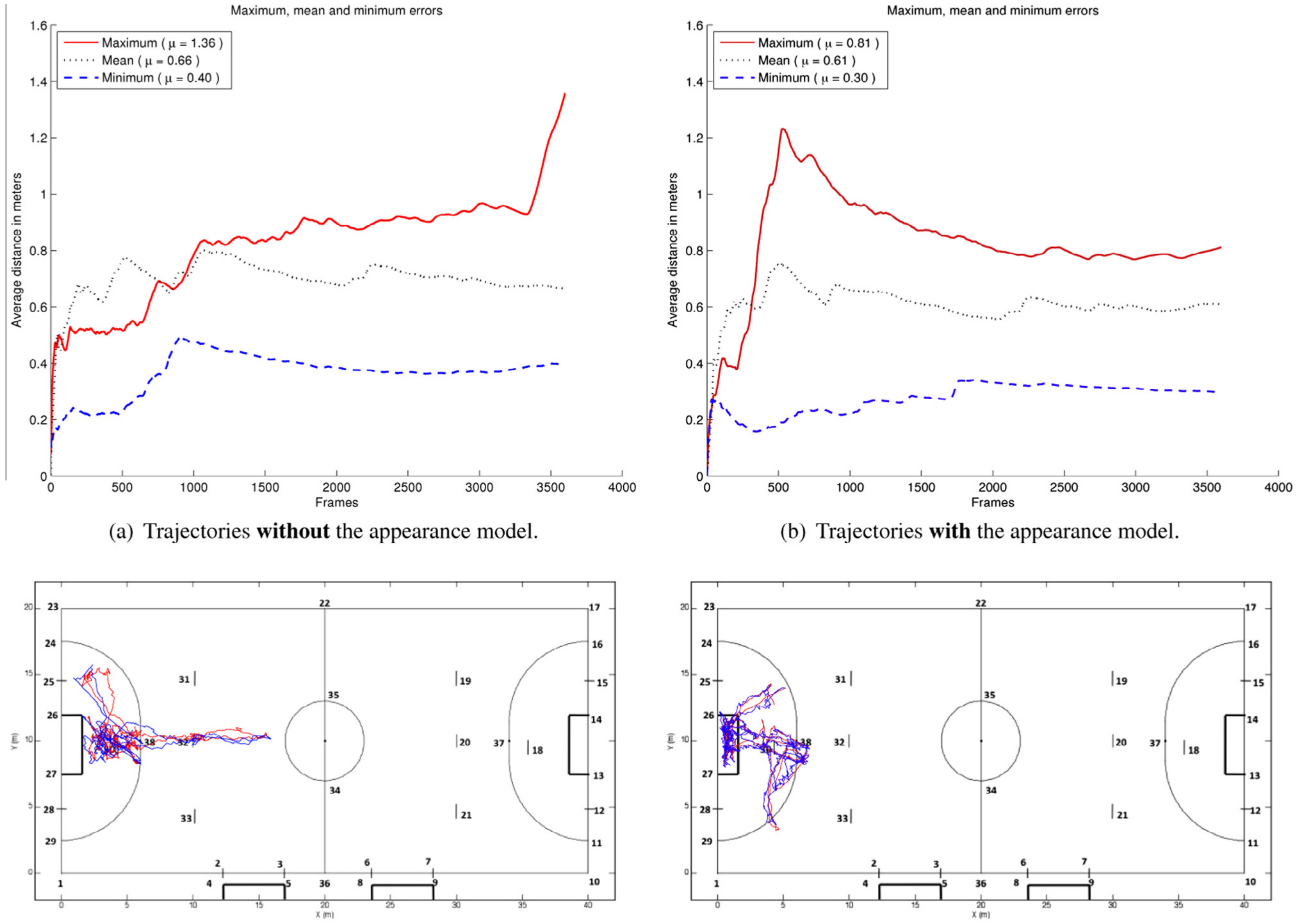
During evaluation, we obtained a successful and accurate tracking for most player's trajectories. However, in some instances there are confusions between players – when they approach each other on the court. Additionally, referees are not a problem to track as their positions are on the side of the court and have a simple motion.

Fig. 7(a) and (b) present the errors in some trajectories and Fig. 7(c) and (d) depict their trajectories along with the ground truth (from the manual annotation). The measurement error is calculated as the difference between position estimates and the ground truth.

The average accumulated error in each trajectory is below under 1 m, but at 0.4 m in the best case. We had a global average error of 0.73 m. Since the dimensions of the court are 20 m × 40 m, and the method is fully automated with an off-the-shelf object detector (OpenCV's Viola and Jones) trained on a simple set, this is a very encouraging result. In Table 2 we can see the errors and standard deviations of successful trajectories.

Confusions are the cases when we have the crossing two or more players' trajectories. Some confusions are solved using a good motion dynamics model in the predictive filter, but cases where multiple players move together along the same path for a period of time, such as during a ball possession dispute or a goal celebration, are not so easily addressed.

The way to address this issue is by strengthening the observation model, and here we use adaptive appearance models. This way the observation function of each player's particle filter will be different, taking into account the difference in the actual images to estimate the likelihood of each detection being from that player.



(a) Trajectories **without** the appearance model.

(b) Trajectories **with** the appearance model.

(c) Tracker **without** appearance model.

(d) Tracker **with** appearance model.

Fig. 7. Mean of accumulated errors. First row: maximum, minimum and mean. Second row: graphical depiction of tracking (blue) and ground truth (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Average errors and standard deviations for the full trajectories found the tracker without using the appearance model.

Tracking	μ	σ
Bolivia × Colombia, 1st period, player 10	0.60	0.11
Bolivia × Colombia, 2nd period, player 10	0.82	0.27
Brazil × Colombia, 1st period, player 11	0.54	0.12
Brazil × Colombia, 2nd period, player 12	0.73	0.36
Brazil × Peru, 1st period, player 09	0.69	0.17
Brazil × Venezuela, 1st period, player 11	0.59	0.17
Colombia × Uruguay, 1st period, player 09	0.40	0.08
Colombia × Uruguay, 1st period, player 11	1.36	2.79
Colombia × Uruguay, 2nd period, player 10	0.53	0.06
Peru × Bolivia, 1st period, player 10	0.66	0.22
Peru × Bolivia, 1st period, player 11	0.61	0.35
Peru × Bolivia, 2nd period, player 10	1.20	0.24
Global	**0.73**	**0.48**

The model is constantly updated throughout the whole tracking and adapts to issues such as illumination changes and both rigid and non-rigid motion of the players as we discuss in Section 5.5.

5.5. Experimental results for tracking: using an appearance model

When we use the appearance model to customize the observation function to each player, we simultaneously handle a larger class of confusion situations while increasing the tracking accuracy (we decrease the bias and covariance of our estimate's error). In Fig. 7(b) we see the tracking result of one player.

Table 3
Average errors and standard deviations for the full trajectories found the tracker using the appearance model.

Tracking	μ	σ
Bolivia × Colombia, 1st period, player 10	0.57	0.11
Bolivia × Colombia, 1st period, player 11	0.68	0.16
Bolivia × Colombia, 2nd period, player 10	0.80	0.29
Brazil × Argentina, 1st period, player 09	0.58	0.07
Brazil × Colombia, 2nd period, player 11	0.52	0.09
Brazil × Colombia, 2nd period, player 12	0.72	0.33
Brazil × Peru, 1st period, player 09	0.64	0.21
Brazil × Venezuela, 1st period, player 10	0.30	0.11
Brazil × Venezuela, 1st period, player 11	0.61	0.21
Colombia × Uruguay, 1st period, player 09	0.36	0.04
Colombia × Uruguay, 2nd period, player 10	0.63	0.07
Peru × Bolivia, 1st period, player 09	0.81	0.29
Peru × Bolivia, 1st period, player 10	0.61	0.21
Peru × Bolivia, 1st period, player 12	0.58	0.18
Peru × Bolivia, 2nd period, player 09	0.61	0.23
Global	**0.60**	**0.19**

Fig. 7(a) and (b) compares the results of using the appearance model, which overall improvement reflects on the mean and covariance of the global estimation error – from 0.73 m to 0.6 m (see Tables 2 and 3), a 20% accuracy improvement.

The appearance model solves some of the confusions problems we discussed earlier. For instance, when two players from different teams are in a dispute for the possession of the ball, they are close to each other with similar motion, but they have distinct appearance models, and the two player's particle filters will have quite distinct observation functions.

Unfortunately, appearance models do not solve everything and there are still two confusion situations left to be tackled in future work. The first type of confusion happens when there are two or more players close together in the image plane, such as when same team players are celebrating an achievement. This leads to multiple players merged into just one on the multimodal function. Since they all have similar appearances and motion dynamics this results in the fusion of their trajectories.

The second type of confusion happens when the detector fails for a player that is close to another one with similar appearance and motion. Again, there is a fusion of trajectories. To handle this second scenario we need to improve the detection techniques, which falls beyond the scope of this paper.

6. Conclusion

In this paper, we explored the problems of tracking futsal players. We presented an automated particle-filter tracking system that fuses the information from four cameras into a mixture of Gaussians in court space. Additionally, we incorporate an appearance model in the fusion mechanism, so that the observation is different for each player, resolving most of the confusions that happens when players get close together.

The main drawback of the existing solutions is the amount of human intervention necessary to properly track. In theory, our approach only requires human intervention for the homography calibration and for the training of the detectors, tasks done only once from a smaller video sequence. In practice, we still do not present the ultimate solution for the problem, as there are a few cases that the motion dynamics and appearance model are not enough to disambiguate players. However, as the results show, we have a promising solution which presents an error lower than we normally we expect to achieve using a GPS-based solution indoors.

Confusion during tracking can happen either due to a trajectory intersection or due to a trajectory coincidence. Our method handles the former, and the appearance model solves several, but not all, of the later.

Our experiments demonstrate that the appearance model customizes each player's observation function, enabling its particle filter to differentiate several confusion situations. When there are two players from different teams, it always works. When there are two players from the same team, the system has a good, but not perfect, performance. As expected, as the number of same-team players increases in the confusion situation, such as in a goal celebration, the system's performance degrades.

Future work will focus on the confusion problem – the main stone in the way of a fully automatic tracking solution. Once we reach this landmark, it will be easier to collect a massive data set, and to begin working on data mining and scientific visualization techniques to help the tactical analysis and understanding of the games.

Acknowledgment

This work was partially supported by São Paulo Research Foundation – FAPESP (Grant 2010/05647-4), National Counsel of Technological and Scientific Development – CNPq (Grants 307018/2010-5, 304352/2012-8, 306730/2012-0, and 477692/2012-5), Microsoft and the Coordination for the Improvement of Higher Level Education Personnel (Capes).

References

Alahi, A., Boursier, Y., Jacques, L., Vandergheynst, P. 2009. Sport player detection and tracking with a mixed network of planar and omnidirectional cameras. In: ACM/IEEE International Conference on Distributed Smart Cameras, pp. 1–8.

Barros, R.M., Menezes, R.P., Russomanno, T.G., Misuta, M.S., Brandao, B.C., Figueroa, P.J., Leite, N.J., Goldenstein, S.K., 2011. Measuring handball players trajectories using an automatically trained boosting algorithm. Computer Methods in Biomechanics and Biomedical Engineering 14 (1), 53–63.

Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer.

de Morais, E.F., Goldenstein, S., Ferreira, A., Rocha, A. 2012. Automatic tracking of indoor soccer players using videos from multiple cameras. In: 25th Conference on Graphics, Patterns and Images, Ouro Preto, Brazil, pp. 174–181.

Du, W., Piater, J., 2007. Multi-camera people tracking by collaborative particle filters and principal axis-based integration. Asian Conference on Computer Vision, vol. Part I. Springer-Verlag, pp. 365–374.

Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9), 1627–1645.

Figueroa, P., Leite, N., Barros, R.M.L., Cohen, I., Medioni, G. 2004. Tracking soccer players using the graph representation. In: International Conference on Pattern Recognition, Washington, DC, USA, pp. 787–790.

Figueroa, P., Leite, N., Barros, R.B.M.L., 2006. Tracking soccer players aiming their kinematical motion analysis. Elsevier Computer Vision and Image Understanding 101 (2), 122–135.

Figueroa, P.J., Leite, N.J., Barros, R.M.L., 2006. Background recovering in outdoor image sequences: an example of soccer players segmentation. Image and Vision Computing 24 (4), 363–374.

Forsyth, D., Ponce, J., 2002. Computer Vision: A Modern Approach. Prentice Hall.

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

Goldenstein, S., 2004. A gentle introduction to predictive filters. Journal of Theoretical and Applied Computing 1, 61–89.

Goldenstein, S., Vogler, C., Metaxas, D., 2003. Statistical cue integration in DAG deformable models. IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (7), 801–813.

Goldenstein, S., Vogler, C., Metaxas, D. 2004. 3D facial tracking from corrupted movie sequences. In: Proceedings of IEEE Computer Vision and, Pattern Recognition.

Gray, A., Jenkins, D., Andrews, M., Taaffe, D., Glover, M., 2010. Validity and reliability of GPS for measuring distance travelled in field-based team sports. Journal of Sports Sciences 28, 1319–1325.

Isard, M., Blake, A., 1998. Condensation – conditional density propagation for visual tracking. International Journal of Computer Vision 29 (1), 5–28.

Juang, C.-F., Sun, W.-K., Chen, G.-C., 2009. Object detection by color histogram-based fuzzy classifier with support vector learning. Journal of Neurocomputing 72 (10–12), 2464–2476.

Kang, J., Cohen, I., Medioni, G. 2003. Soccer player tracking across uncalibrated camera streams. In: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 172–179.

Kasiri-Bidhendi, S., Safabakhsh, R. 2009. Effective tracking of the players and ball in indoor soccer games in the presence of occlusion, in: International Computer Conference, pp. 524–529.

Khan, S.M., Shah, M. 2006. A multiview approach to tracking people in crowded scenes using a planar homography constraint. In: European Conference on Computer Vision, pp. 133–146.

Mendez-Villanueva, A., Buchheit, M., Simpson, B., Bourdon, P.C., 2013. Match play intensity distribution in youth soccer. International Journal of Sports Medicine 02 (34), 101–110.

Miura, J., Kubo, H. 2008. Tracking players in highly complex scenes in broadcast soccer video using a constraint satisfaction approach. In: International Conference on Content-Based Image and Video Retrieval, pp. 505–514.

Morais, E., Goldenstein, S., Rocha, A. 2012. Automatic localization of indoor soccer players from multiple cameras. In: International Conference on Computer Vision Theory and Applications, pp. 205–212.

Okuma, K., Taleghani, A., Freitas, N., Little, J., Lowe, D. 2004. A boosted particle filter: multitarget detection and tracking. In: European Conference on Computer Vision, vol. 3021, pp. 28–39.

Stauffer, C., Grimson, W. 1999. Adaptive background mixture models for real-time tracking. In: IEEE International Conference on Computer Visiona and Pattern Recognition, vol. 2, pp. 252–260.

Trucco, E., Verri, A., 1998. Introduction Technique for 3-D Computer Vision. Prentice Hall.

Viola, P., Jones, M. 2001. Rapid object detection using a boosted cascade of simple features. In: IEEE International Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 511–518.