

# Tracking multiple sports players through occlusion, congestion and scale

Chris J. Needham and Roger D. Boyle

School of Computing

The University of Leeds

Leeds, LS2 9JT, UK

[chrisn@comp.leeds.ac.uk](mailto:chrisn@comp.leeds.ac.uk)

<http://www.comp.leeds.ac.uk/vision>

## Abstract

Tracking sports players over a large playing area is a challenging problem. The players move quickly, and have large variations in their silhouettes.

This paper presents a framework for multi-object tracking, using a CONDENSATION based approach. Each player being tracked is independently fitted to a model, and the sampling probability for the group of samples is calculated as a function of the fitness score of each player. This function rewards consistently good scores, but punishes a group of some very good and some very bad fitness scores. Ground plane information is used throughout, and the predictive stage of the algorithm is improved to incorporate estimates of position from Kalman filters. This helps group the estimated positions of each player, and to aid in tracking through occlusions.

## 1 Introduction

This work aims to track the movements of sports (specifically football) players, from a single fixed camera, on an indoor court. This is to allow work on the behaviour modelling and positional analysis of the sports players to be undertaken.

There are two primary motivations for this. Firstly, the sports science industry is very interested in being able to know how much ground athletes have covered, and how quickly they have moved, during the course of a game. This information would allow more specific training to be designed to suit individual players. More interestingly, team games are complicated activities, which involve a lot of interaction between the players. This multi-player activity allows us to explore the relations and interactions, both between players, and the teams as a whole.

This domain exhibits several challenging aspects: the size of the pitch means that the resolution of an image of the game varies greatly between the nearest and the furthest parts of the pitch; sports games are busy areas; sports players' shapes vary significantly, often in short periods; and the players move at variable speeds, often suddenly changing direction, which makes their movements hard to predict.



Figure 1: *Example footage of soccer players.*

## 2 Background

Many different trackers with various purposes have been developed in recent years. One of the first of these, designed for individual pedestrian surveillance, was the Leeds People Tracker [1] which employs contour tracking, active shape models, and Kalman filtering to track multiple people from a single camera. Some systems allow a determination of the body pose, and real-time tracking of head and hands, such as Pfinder [16]. Pfinder is a ‘person-finder’ which uses a multi-class statistical model of colour and shape to create a blob representation of a tracked person. This will only work when there is a single person in the scene, and produces a more detailed model than is needed to obtain the position of sports players. Recent work by McKenna *et al.* [10] performs tracking on three levels of abstraction: regions, people and groups. Strong use is made of colour information in this system, to assist in coping with shadows and disambiguating occlusions in pedestrian scenes.

Currently there are several popular methods for tracking moving targets, which include: Active Shape Models [4] which are flexible shape models, allowing iterative refinements of estimates of the objects’ pose, scale and shape; the Kalman filter [15, 3], which has been used in many tracking applications due to its computational efficiency and its ability to estimate future states; and Isard and Blake’s newer method of CONDENSATION [11], which is a powerful technique allowing the propagation of conditional densities over time, and has been used with contour tracking to track an object through cluttered scenes. An initial drawback to this scheme is when tracking multiple targets. If several one-body trackers are employed, each with the same tracking algorithm, then two or more can coalesce onto the same target for which their model best fits. Recently, MacCormick and Blake introduced a probabilistic exclusion principle [8] to couple CONDENSATION trackers, in a bid to tackle this problem.

Football related work has been inspired by several different ambitions, including annotation, action recognition, game reconstruction, and evaluating play. Intille and Bobick used closed worlds [6] for video-annotation of American football footage, which has been followed by work aiming to identify actions from visual evidence, namely American football plays [7]. SoccerMan [2] is a (soccer) game reconstruction system: various

techniques are used in the tracking of players, and then a virtual 3D world with play-ground texture and textured player shapes can be formed, which can be viewed from any virtual viewpoint. Taki *et al.* [13] have looked at evaluating teamwork in soccer games, by investigating space advantage on the pitch.

The aim of this work is to produce a tracker which will automatically track sports players, and identify their real world positions for positional behaviour analysis, as opposed to recognising whether they are running, kicking the ball, or involved in a set play.

## **3 Theoretical aspects**

### **3.1 Image perspective**

Tracking multiple objects through busy cluttered scenes remains a challenging problem. Figure 1 shows a typical scene from an indoor 5-a-side soccer game. All the action is constrained to the pitch, however the perspective of the image highlights several issues. The size of the pitch (18x32 metres) means that the resolution of an image of the game varies greatly between the nearest and the furthest parts of the pitch. Analysis reveals that in a typical image (e.g., Figure 1 which is 320x240 pixels in size), if two vertically adjacent pixels on the image plane are projected onto the ground plane, then pixels in the nearest part of the image are 3cm apart, whereas those in the far goal mouth are over 45cm apart. In the area of the image representing the nearest part of the pitch, 3 metres of ground plane covers 72 pixels, compared with only 8 pixels at the far end of the pitch.

This emphasises the importance of considering the depth information in the image. It becomes important for the tracking to be performed using the ground plane coordinates, which take into account the amount of ground that a player is physically able to cover over time. The corresponding distance in image pixels varies greatly over the image. Knowledge of the players' position on the ground plane aids with resolving occlusions, particularly in scenes from such a perspective view, since often one player occludes part of another player when they are more than a metre away from each other.

The main feature of interest of a player is the position of the feet, which is why this is used instead of the players' centroid, and it is this position which we wish to determine with the greatest accuracy, for use when modelling the players' behaviour in future work. It is assumed that the players' feet are in contact with the floor when calculating their world position from the image.

### **3.2 Image segmentation**

Many attempts have been made at image segmentation from video, using background subtraction, adaptive background subtraction [10], and colour space models [14].

Maintaining a temporal background model and performing background subtraction has been shown to be a fast and efficient method of extracting moving objects from a scene [1]. This performs best in relatively empty scenes through which objects are moving, but sporting activities do not fit into this category. Sports players are always on the pitch, and often (particularly in sports like netball) have a tactical position in which they stand for short lengths of time, and can become incorporated into the background model through dynamic background maintenance. If a static background model is used to combat this, it may not allow for changes in the lighting conditions or small camera movements.

Good, fast segmentation can be achieved by creating prior colour space models [14] for foreground and background, in this case players and non-players. This method is robust to small camera jitters, and stationary objects.

In this work, a foreground model in HSI space is built offline from a sample of pixels from regions of the images identified as being foreground before tracking begins. HSI space is used because the separation between the foreground and background clusters is greater in this space than in other possible spaces such as RGB, or chromicity values. The same is done for the background. For each pixel in an image, a probability of being foreground can be assigned

$$p(\text{fore}) = d_b / (d_f + d_b) \quad (1)$$

where  $d_f$  and  $d_b$  are the Mahalanobis distance of the pixel from the respective cluster centres. This creates a noisy image, with player regions being fragmented, particularly the players' legs. Segmentation is improved by performing probabilistic relaxation on the image using

$$\begin{aligned} p(\text{fore}) &= p(\text{fore}) + \delta \quad \text{if median value of neighbouring pixels} > 0.5 \\ p(\text{fore}) &= p(\text{fore}) - \delta \quad \text{otherwise} \end{aligned} \quad (2)$$

and choosing  $\delta$  such that a pixel is able to change from foreground to background (or vice-versa) after a suitable number of applications of the relaxation. Using 3 applications of the probabilistic relaxation with  $\delta = 0.2$  produces a well segmented foreground.



Figure 2: *The variation in shape of soccer players*

### 3.3 Shape models

Generally in tracking applications the objects to be identified are similar in nature. For example industrial inspection of resistors [4], pedestrians in car parks [1], and chickens in boiler houses [12]. Figure 2 shows the variation in the outlines of extracted soccer players, which raise questions about the most suitable way of tracking these shapes. Contour models such as PDMs [5] or spline models rely on a set of points on the extracted outline of the shape. For similar shapes these cluster well, and PCA can be used to reduce the dimensionality, thus identifying the major features or characteristics of the shapes. A single shape model looks unsuitable for modelling the silhouette of a soccer player. Magee [9] has used three shape models to represent different configurations of cows' legs during tracking. A similar approach could be applied here, with a number of models being used to represent when players; stand with their legs closed; stand with their legs open; are

running, creating a diagonal shape; or have their arms out. This level of complexity may be too great for this application.

The approach taken here is just to fit a bounding box to each silhouette, and evaluate how well this fits the image data. It would be worth using a more complex model if information about the players' pose and orientation were to be our goal. The purpose of tracking the sports players is for analysis of their movements and positional behaviours, thus the most important feature is their feet.

## 4 Multiple object tracking

### 4.1 Structure

A multiple object CONDENSATION based approach is employed, as opposed to employing multiple single object trackers. This multi-object tracking adds an extra level to the structure of the algorithms. Here, a *sample* represents an instance of a player, a *sampleset* represents a collection of samples (players being tracked) at an instance, and a *supersampleset* represents a collection of samplesets.

It is the contact point of the players' feet with the floor that we wish to identify with the greatest accuracy. Image coordinates  $(u, v)$  can be used to represent the position of the players' contact point with the floor; calibrating the image plane allows image points to be projected to ground plane (world) coordinates. In this work, ground plane coordinates  $(x, y)$  are used throughout the computations, with image positions calculated from these.

To calculate a bounding box for a player, first the single world point is projected onto the image plane to a point  $(u, v)$ ; next a bounding box of width  $w$  and height  $h$  is constructed, assuming that this point is the midpoint of the base of the bounding box. On creation, an identification number,  $id$ , is included for use in determining the player to which trajectories belong. Thus each player can be represented as

$$\mathbf{x} = (x, y, h, w, id) \quad (4)$$

Let  $\mathbf{x}_t^i$  be an instance of a sample at time  $t$ . A sampleset  $\mathbf{s}_t^j$  can be formed, which consists of an instance of each different object being tracked, along with a corresponding sampling probability  $\pi_t^j$ .

$$\mathbf{s}_t^j = (\mathbf{x}_t^1, \mathbf{x}_t^2, \dots, \mathbf{x}_t^{n_j}, \pi_t^j) \quad (5)$$

where  $n_j$  is the number of objects in the sampleset  $\mathbf{s}_t^j$ .

A 'supersampleset'  $\mathbf{S}_t = (\mathbf{s}_t^1, \mathbf{s}_t^2, \dots, \mathbf{s}_t^N)$  is created to store each of these samplesets, where  $N$  is the predefined number of samples used in the CONDENSATION algorithm.

### 4.2 Propagation

The samplesets within the supersampleset are propagated in the usual way, that is  $N$  samplesets are generated from  $p(\mathbf{s}_t|\zeta_t)$  at each time step, where  $\zeta_t$  is foreground image probability data, i.e.,  $\mathbf{s}'_t$  is drawn randomly from  $p(\mathbf{s}_t|\zeta_t)$ . Then  $\mathbf{s}'_{t+1}$  is drawn randomly from  $p(\mathbf{s}_{t+1}|\mathbf{s}_t = \mathbf{s}'_t)$  and  $\pi_{t+1}$  is calculated as  $p(\zeta_{t+1}|\mathbf{s}_{t+1} = \mathbf{s}'_{t+1})$ .

The probabilities are assigned by rescaling the weights assigned to each sampleset from a fitness function which assesses how well each bounding box fits its target. If the

fitness score for each player sample within the particular sampleset is similar, then the overall score for the sampleset is increased (rewarded). If one or more of the samples is a poor fit, then the overall score is reduced (punished). This aims to aid the propagation of the samplesets with the best overall fit of the  $n_j$  objects, but not those for which one or more objects fit very well, when there are some that don't fit well at all. The sampleset with the highest sampling probability is used as the 'best' sampleset for representing the players.

### 4.3 Prediction

A simple model to use to predict each sampleset  $\mathbf{s}_t^j \in \mathbf{S}_t$  from  $\mathbf{s}_{t-1}^j \in \mathbf{S}_{t-1}$  is

$$\begin{aligned} x_t^i &= x_{t-1}^i + \varepsilon_x & h_t^i &= h_{t-1}^i + \varepsilon_h \\ y_t^i &= y_{t-1}^i + \varepsilon_y & w_t^i &= w_{t-1}^i + \varepsilon_w \end{aligned} \quad (6)$$

for  $i = \{1, \dots, n_j\}$ , and where  $\varepsilon_x$  and  $\varepsilon_y \sim \mathcal{N}(0, \sigma_1)$  with  $\sigma_1$  typically in the region of 100mm, given that the maximum distance on the ground plane that a sports player will move will be in the order of  $3\sigma_1$  (300mm per 25th of a second). This allows for the tracking of a player who moves at a speed of  $7.5 \text{ m s}^{-1}$ . Velocity information of the players could be incorporated at this stage, however the nature of the sports often involve the players making sharp, sudden changes of direction.

Due to the rapid change in shape of a player that can occur when they raise arms to attract attention, or open stride when running, the height and width of the bounding box must be allowed to react quickly to this change, thus adding Gaussian noise to the height and width with  $\varepsilon_h$  and  $\varepsilon_w \sim \mathcal{N}(0, \sigma_2)$  with  $\sigma_2 = 2$  pixels allows such a change.

Doing this has the drawback that the samples in a sampleset may no longer each correspond to different players, for example, when one sample locks to another target in close proximity, which is already being tracked.

## 5 Improving with Kalman filtering

Altering the predictive step of the CONDENSATION algorithm can prevent the samples from straying too far from the other samples representing the same target. The position of a player on the ground plane can be predicted for the next time step, given previous states. Here,  $n_j$  Kalman filters are used, one for each player. They are updated using the observed value of the position of each player (sample) in the 'best' sampleset.

Kalman filters are being used because they address the problem of estimating the position  $\mathbf{x}_t = (x, y) \in \mathbb{R}^2$  of the player at the next discrete time step. A simple linear stochastic difference equation governs this process

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{w}_{t-1} \quad (7)$$

with a measurement  $\mathbf{z} \in \mathbb{R}^2$  which directly relates to  $\mathbf{x}$  that is

$$\mathbf{z}_t = \mathbf{x}_k + \mathbf{v}_{t-1} \quad (8)$$

The independent random variables  $\mathbf{w}_t$  and  $\mathbf{v}_t$  represent the process and measurement noise, and have normal probability distributions

$$p(\mathbf{w}) \sim \mathcal{N}(0, \mathcal{Q}) \qquad p(\mathbf{v}) \sim \mathcal{N}(0, \mathcal{R}) \qquad (9)$$

Currently constant  $\mathcal{Q}$  (process noise covariance) and  $\mathcal{R}$  (measurement noise covariance) are used. However, in the future these may be used to assess the certainty of the estimates being made, which will improve the ‘trust’ in the estimate of the players position from the Kalman filter, compared to the observation  $\mathbf{z}$  from the image, when resolving occlusions.

At each time step, a Kalman estimate  $\hat{\mathbf{x}}_t = (\hat{x}_t, \hat{y}_t)$  of the position of each player is calculated, and each sampleset  $\mathbf{s}^j_t \in \mathbf{S}_t$  is predicted from  $\mathbf{s}^j_{t-1} \in \mathbf{S}_{t-1}$  using

$$\begin{aligned} x_t^i &= (\hat{x}_t + x_{t-1}^i)/2 + \varepsilon_x & h_t^i &= h_{t-1}^i + \varepsilon_h \\ y_t^i &= (\hat{y}_t + y_{t-1}^i)/2 + \varepsilon_y & w_t^i &= w_{t-1}^i + \varepsilon_w \end{aligned} \qquad (10)$$

for  $i = \{1, \dots, n_j\}$ , and the observed player positions  $\mathbf{z}_t$  from the ‘best’ sampleset are used to update each discrete Kalman filter. This has the effect of grouping the samples corresponding to each player within the CONDENSATION algorithm, since each sample is drawn towards the predicted  $\hat{\mathbf{x}}_t$  for that player. This prevents the samples for a player splitting up into two or more groups, which might have allowed the ‘best’ sample for a player to jump between the groups, or lock onto a different player instead.

## 6 Evaluation and results

Considering the implications of section 3.1, there is probably a limited area of the ground plane for which the comparison of differences in ground plane positions is valid, since on parts of the image neighbouring pixels are almost half a metre apart. Also using the assumption that the players’ position is at the midpoint of the base of the bounding box may not be valid when considering asymmetric shapes, for example when a player leans to one side. However, here it is assumed these are usable enough to be valid.

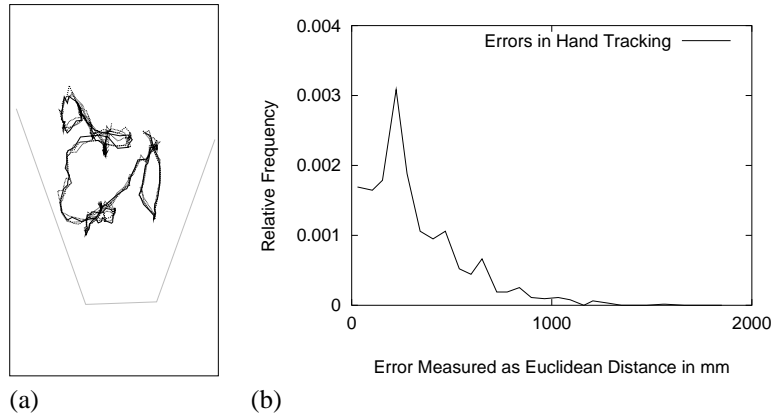


Figure 3: (a) Four independent hand-tracked trajectories of the same soccer player over 835 frames, every fifth frame. (b) Euclidean difference between player position, in the six pairwise permutations of the four hand-tracked trajectories.

In order to evaluate the tracking, the true ground plane positions of the players need to be determined. A sequence was independently hand marked-up 4 times, and analysis of each resulting trajectory with each of the other trajectories has been performed. Figure 3(a) shows the trajectories of a single footballer over 835 frames. Figure 3(b) shows the distribution of the Euclidean differences between each trajectory, which were calculated as the distance between two players' positions at each time step. Analysis of the six pairwise permutations of the four hand-tracked trajectories shows a mean difference of 312.2 mm between positions, a standard deviation of 239.7 mm, and a mode between 200 and 300 mm. Thus, it is reasonable to take the mean of four hand marked-up trajectories as the 'true' trajectory of the player, to which automatically tracked trajectories are compared.

For the modelling of sports players behaviour, zero error in players' positions would be ideal, although given the variability in human performance, an error of up to 0.5 m might be regarded as acceptable in hand-tracked data. It is expected that data will be usable for behaviour analysis if it is within 1 m of true position, so trajectories will be regarded as acceptable if they are within 1 m of the mean hand-tracked position.

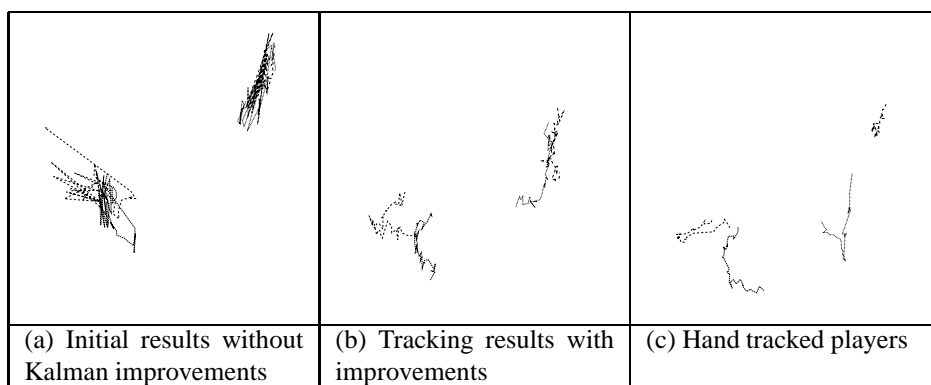


Figure 4: Comparison of trajectories of soccer players over 40 frames.

Tracking has been performed on a short sequence of indoor 5-a-side football; four players were tracked in the sequence. Firstly, the tracking was performed using the multi-target CONDENSATION described in Section 4, using  $N = 1000$  samples, which resulted in the position of the player being allowed to jump around as multiple hypotheses of each player were propagated. Samples were observed switching targets from frame to frame, and not consistently locking onto a specific target. Comparing the trajectories to hand marked-up trajectories revealed a mean error of 2.5 metres for this imperfect system, shown in Figure 5(a).

The tracking was performed again after the improvements detailed in Section 5. This time, the samples locked onto the four players much better, without swapping players, or having multiple samples tracking a single player. This reduced the mean error in the positions to 1.16 metres, and the modal value to below 400 mm. Figure 5(a) shows the error distances, and highlights the improvement in the new tracking system. Figure 5(b) shows close up how noisy the trajectories are, and that better results could be produced with more filtering to smooth the trajectories after tracking is complete. The bounding boxes marking the tracked players are illustrated in figure 6.

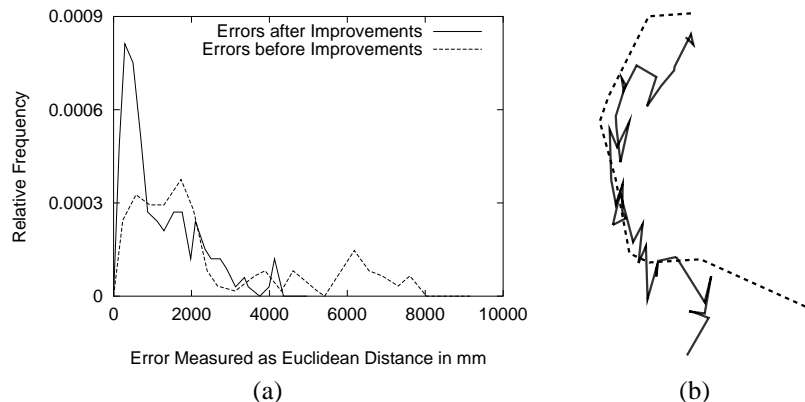


Figure 5: (a) Reduction in the size of the Euclidean error of the tracked footballers on the ground plane after the improvements over 40 frames, compared to the same hand-tracked sequence. (b) Comparison of trajectories of soccer players. The solid line represents the automatically tracked player. The dashed line represents the hand tracked player.

## 7 Conclusions

This work has presented a novel framework for multi-object tracking. The initial scheme gave 28% of the tracking as usable. With the improvements, 56% of the trajectories are within a metre of the hand marked-up trajectory, and hence usable for behaviour modelling. The errors in the tracker are characterised by the bounding box not fitting to the feet well enough, whilst player tracking is maintained.

Future work will see the introduction of a more complex model for the shape, and the positional behaviour analysis of sports players.

## Acknowledgements

The authors would like to thank the University Sports Science department for their support in obtaining video footage, and the financial support of an EPSRC PhD studentship award.

## References

- [1] A. M. Baumberg. *Learning Deformable Models for Tracking Human Motion*. PhD thesis, School of Computer Studies, University of Leeds, 1995.
- [2] T. Bebie and H. Bieri. SoccerMan - reconstructing soccer games from video sequences. In *Proc. of the Int. Conf. on Image Processing*, pages 898–902, 1998.
- [3] C. K. Chui and G. Chen. *Kalman Filtering with Real-Time Applications*. Springer, 1999.
- [4] T. F. Cootes and C. J. Taylor. Active shape models - ‘smart snakes’. In *Proc. British Machine Vision Conference*, 1992.
- [5] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Training models of shape from sets of examples. In *Proc. British Machine Vision Conference*, pages 9–18, 1992.

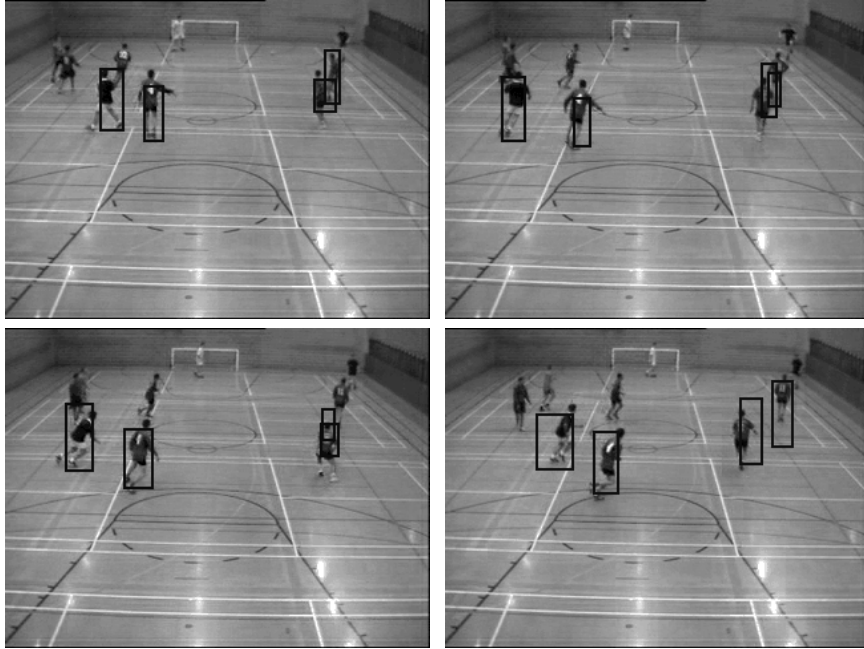


Figure 6: Tracking soccer players, frames 30,40,50,60.

- [6] S. S. Intille and A. F. Bobick. Visual tracking using closed-worlds. In *Proc. Int. Conf. on Computer Vision*, 1995.
- [7] S. S. Intille and A. F. Bobick. A framework for recognizing multi-agent action from visual evidence. In *Proc. of the Nat. Conf. on A.I.*, pages 518–525, 1999.
- [8] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. In *Proc. Int. Conf. on Computer Vision*, pages 572–578, 1999.
- [9] D. R. Magee and R. D. Boyle. Building class sensitive models for tracking application. In *Proc. British Machine Vision Conference*, pages 594–603, 1999.
- [10] S. J. McKenna, S. Jabri, Z. Duric, and H. Wechsler. Tracking interacting people. In *Proc. Fourth IEEE Int. Conf. Automatic Face and Gesture Recognition*, pages 348–353, 2000.
- [11] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *Proc. European Conf. Computer Vision*, pages 343–356, 1996.
- [12] D. M. Sergeant, R. D. Boyle, and J. M. Forbes. Computer visual tracking of poultry. *Computers and Electronics in Agriculture*, 21(1):1–18, 1998.
- [13] T. Taki, J. Hasegawa, and T. Fukumura. Development of motion analysis system for quantitative evaluation of teamwork in soccer games. In *Proc. Int. Conf. Image Processing*, 1996.
- [14] N. Vandenbroucke, L. Macaire, and J. G. Postaire. Color pixels classification in an hybrid color space. In *Int. Conf. on Image Processing*, pages 176–180, 1998.
- [15] G. Welch and G. Bishop. An introduction to the Kalman filter. Technical Report TR 95-041, University of North Carolina at Chapel Hill, 1995.
- [16] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfunder: real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.