# TRACKING FOOTBALL PLAYERS WITH MULTIPLE CAMERAS

*Ming Xu* *James Orwell* *Graeme Jones*

Digital Imaging Research Centre
Kingston University
Kingston upon Thames, KT1 2EE, UK
{m.xu,j.orwell,g.jones}@kingston.ac.uk

## ABSTRACT

A system is described for tracking the positions of football players during a match. The input is eight video streams from static cameras, each processed to generate measurements of the players for a multi view tracker. The single view processing includes foreground detection and a bounding box tracker designed to split measurements of merged players. In the multi-view process, the measurements, each from a different camera and matched to the same player, are fused into an overall measurement which is then used in the estimate updating. Results are demonstrated on real data.

## 1. INTRODUCTION

This paper describes an application that requires multiple targets to be tracked with multiple cameras. The application output is the positions of players, and ball, during a football match. This output can be used for entertainment-augmenting digital TV, or low-bandwidth match play animations for web or wireless display; and also for analysis of fitness and tactics of the teams and players.

There exist a number of research projects on tracking soccer players. Intille and Bobick [3] track players, using the concept of closed-world, in the TV broadcast of American soccer games. In [4] panoramic views and player trajectories are computed from a monocular TV sequence. The SoccerMan [2] analyses two synchronised video sequences of a soccer game and generates an animated virtual 3D view of the given scene. These projects use one (or two) pan-tilt-zoom camera to guarantee players' size in images and the correspondence between frames has to be made on the basis of matching field lines or arcs. An alternative approach to improving players' resolution is to use multiple stationary cameras. This method increases the overall field of view, minimises the effects of dynamic occlusion, provides 3D estimates of ball location, and improves the accuracy and robustness of estimation due to information fusion. There are different ways to use multi-view data, such as hand-off between best-view cameras, homography transform between the images of uncalibrated cameras, or using calibrated cameras able to determine the 3D world coordinate with the cooperation of two or more cameras.

Our system uses eight digital video cameras statically positioned as shown in Fig. 1 and calibrated to a common ground-plane coordinate system using Tsai's algorithm [6]. The first processing stage is the extraction of measurements about the players observed by each camera, which is described in Section 2. The data from each camera is input to a central tracking process, described in Section 3, to update the state estimates of the players. This includes the estimate of which of the five possible uniforms each player is wearing (two outfield teams, two goal-keepers, and the three referees. In this paper, *player* includes the referees). The output from this central tracking process is the 25 player positions per time step. The tracker indicates the category (team) of each player, and maintains the correct number of players in each category. The identification of individual players is not required, given the resolution of input data. The ball tracking is presented in an accompanying paper.

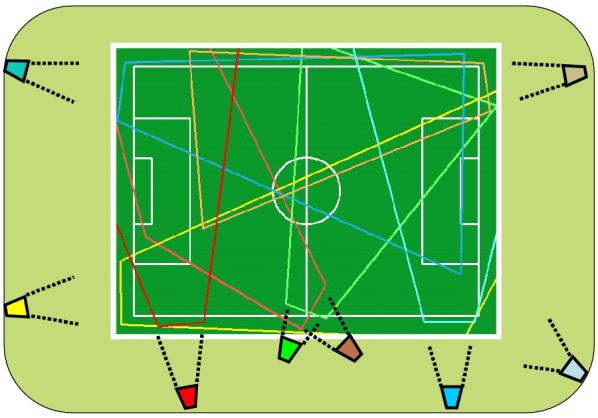

**Fig. 1**. The camera placements and fields of view.

This work is part of the INMOVE project, supported by the European Commission IST 2001-37422.

## 2. VIDEO STREAM PROCESSING

In the video processing stage, a three step approach is used to generate the measurements. Each measurement consists of a 2D ground-plane position, its spatial covariance, and a category estimate.

### 2.1. Foreground Detection

The first step is moving object detection based on image differencing and its output is connected foreground regions (Fig. 2). An initial background is modelled with a mixture of Gaussians and learned before a match without the need of an empty scene. It is first used to extract a pitch mask for saving processing time and avoiding false alarms from the crowd. This pitch mask is computed using the hue histogram and the projection of the known pitch geometry to image planes. The initial background is then forwarded to the running average algorithm for fast updating. Suppose $F_k$ is the foreground binary map at time $k$, then the background $u_k$ is updated with image $I_k$ as:

$$u_k = [\alpha I_k + (1-\alpha)u_{k-1}]\, F_k + [\rho I_k + (1-\rho)u_{k-1}]\, \bar{F}_k$$

where $0 < \alpha << \rho << 1$. The background updating is applied even in foreground regions so that any mistake in the initial background and any new litter will not be locked.

### 2.2. Single View Tracking

The second step is a local tracking process [7] to split measurements of merged people. The bounding box and centroid coordinates of each player are used as state and measurement variables in a Kalman filter:

$$\mathbf{x}_I = \left[\begin{array}{cccccccc} r_c & c_c & \dot{r_c} & \dot{c_c} & \Delta r_1 & \Delta c_1 & \Delta r_2 & \Delta c_2 \end{array}\right]^T$$
$$\mathbf{z}_I = \left[\begin{array}{cccccc} r_c & c_c & r_1 & c_1 & r_2 & c_2 \end{array}\right]^T$$

where $(r_c, c_c)$ is the centroid, $r_1, c_1, r_2, c_2$ represent the top, left, bottom and right bounding edges, respectively ( $r_1 < r_2$ and $c_1 < c_2$ ). $(\Delta r_1, \Delta r_2)$ and $(\Delta r_2, \Delta c_2)$ are the *relative* positions of the two opposite bounding box corners to the centroid. The state transition and measurement matrices are:

$$\mathbf{A}_I = \left[\begin{array}{cccc} \mathbf{I}_2 & T\mathbf{O}_2 & \mathbf{O}_2 & \mathbf{O}_2 \\ \mathbf{O}_2 & \mathbf{I}_2 & \mathbf{O}_2 & \mathbf{O}_2 \\ \mathbf{O}_2 & \mathbf{O}_2 & \mathbf{I}_2 & \mathbf{O}_2 \\ \mathbf{O}_2 & \mathbf{O}_2 & \mathbf{O}_2 & \mathbf{I}_2 \end{array}\right]$$

$$\mathbf{H}_I = \left[\begin{array}{cccc} \mathbf{I}_2 & \mathbf{O}_2 & \mathbf{O}_2 & \mathbf{O}_2 \\ \mathbf{I}_2 & \mathbf{O}_2 & \mathbf{I}_2 & \mathbf{O}_2 \\ \mathbf{I}_2 & \mathbf{O}_2 & \mathbf{O}_2 & \mathbf{I}_2 \end{array}\right]$$

where $\mathbf{I}_2$ and $\mathbf{O}_2$ are $2 \times 2$ identity and zero matrices, respectively; $T$ is the sampling time. [7] assumes that each target has a nearly constant height and width. Once some bounding edge of a target is decided to be observable, its opposite, unobservable bounding edge could be roughly estimated (Fig. 2). Because the estimate is updated using partial measurements whenever available, it is more accurate and robust than using prediction only.

For an isolated player, the image measurement comes from the bottom of foreground region directly. The measurement covariance in an image plane is assumed to be a constant and diagonal matrix $\mathbf{\Lambda}$, because foreground detection in an image is a pixelwise operation. The corresponding measurement and covariance projected on the ground plane from $i$-th camera image are:

$$\begin{aligned} \mathbf{z}^{(i)} &= E^{(i)}(r_2, c_c) \\ \mathbf{R}^{(i)} &= \mathbf{J}_E^{(i)}(r_2, c_2)\mathbf{\Lambda}\mathbf{J}_E^{(i)}(r_2, c_2)^T \end{aligned}$$

where $E$ and $\mathbf{J}_E$ is the coordinate transformation and its Jacobian from the image plane to the ground plane (see Fig. 3). For a grouped player, the measurement is calculated from the estimate and the covariance increases ($\lambda > 1$):

$$\begin{aligned} \mathbf{z}^{(i)} &= E^{(i)}(\hat{r}_2, \hat{c}_c) \\ \mathbf{R}^{(i)} &= \mathbf{J}_E^{(i)}(\hat{r}_2, \hat{c}_2)\lambda\mathbf{\Lambda}\mathbf{J}_E^{(i)}(\hat{r}_2, \hat{c}_2)^T \end{aligned}$$

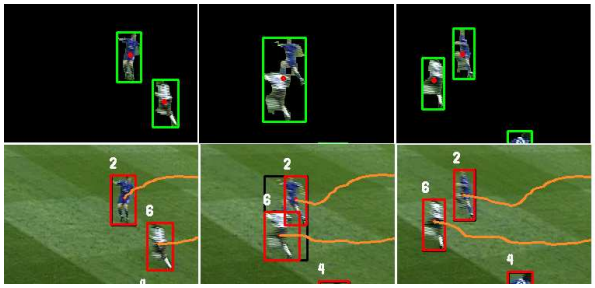

**Fig. 2**. Player detection and tracking from a single camera.

### 2.3. Category Measurement

The final step adds to each measurement an estimate of the category (or player's uniform). This is implemented using a histogram-intersection method [5]. The result for each player is a five-element vector $\mathbf{c}^{(i)}$, indicating the probability that the observed player is wearing one of the five categories of uniform.

## 3. MULTI VIEW TRACKING

For the multi-view tracking process, a three-step procedure is applied. The first step is to associate measurements to established tracks and update these tracks. The second step is to initialize tracks for the measurements unmatched to any existing track. Finally, the fixed population constraint for each category of players (ten outfield players and one goalkeeper per team, three referees) is used to recognize the members in each category.

### 3.1. Track Update

Each player is modelled as a track $x_t$ and has its state estimate updated, if possible, by an overall measurement $m_t$ fused from at least one camera. The state and measurement variables in a ground-plane Kalman filter are $\mathbf{x} = \begin{bmatrix} x & y & \dot{x} & \dot{y} \end{bmatrix}^T$, and $\mathbf{z} = \begin{bmatrix} x & y \end{bmatrix}^T$. The state transition and measurement matrices are:

$$\mathbf{A}_w = \begin{bmatrix} \mathbf{I}_2 & T\mathbf{I}_2 \\ \mathbf{O}_2 & \mathbf{I}_2 \end{bmatrix} \quad \mathbf{H}_w = \begin{bmatrix} \mathbf{I}_2 & \mathbf{O}_2 \end{bmatrix}$$

The established tracks $\{x_t\}$ are associated with the measurements from $i$-th camera, the result of which is expressed as an association matrix $\beta^{(i)}$ for that camera. Each element $\beta_{jt}^{(i)}$ is 1 for association between the *t*-th track and *j*-th measurement or 0 otherwise. The association matrix is decided according to the Mahalanobis distance between the measurement and the track prediction. For a possible association this distance must be within a validation gate. Then the nearest neighbour algorithm is applied and each track can be associated with at most one measurement from a camera, i.e. $\sum_j \beta_{jt}^{(i)} \leq 1$ . The individual camera measurements assigned to each track are weighted by measurement uncertainties and integrated into an overall measurement $m_t$ as follows (Fig. 3):

$$\begin{aligned} \mathbf{R}_t &= \left[ \sum_i \sum_j \beta_{jt}^{(i)} \left( \mathbf{R}_j^{(i)} \right)^{-1} \right]^{-1} \\ \mathbf{z}_t &= \mathbf{R}_t \left[ \sum_i \sum_j \beta_{jt}^{(i)} \left( \mathbf{R}_j^{(i)} \right)^{-1} \mathbf{z}_j^{(i)} \right] \\ \mathbf{c}_t &= \sum_i w_t^{(i)} \sum_j \beta_{jt}^{(i)} \mathbf{c}_j^{(i)} \\ w_t^{(i)} &= \frac{\sum_j \beta_{jt}^{(i)} / tr \left( \mathbf{R}_j^{(i)} \right)}{\sum_i \sum_j \beta_{jt}^{(i)} / tr \left( \mathbf{R}_j^{(i)} \right)} \end{aligned}$$

Each track with measurements is then updated using the integrated measurement:

$$\begin{aligned} \mathbf{K}_t &= \mathbf{P}_t^- \mathbf{H}_w^T \left[ \mathbf{H}_w \mathbf{P}_t^- \mathbf{H}_w^T + \mathbf{R}_t \right]^{-1} \\ \hat{\mathbf{x}}_t^+ &= \hat{\mathbf{x}}_t^- + \mathbf{K}_t \left[ \mathbf{z}_t - \mathbf{H}_w \hat{\mathbf{x}}_t^- \right] \\ \mathbf{P}_t^+ &= \left[ \mathbf{I} - \mathbf{K}_t \mathbf{H}_w \right] \mathbf{P}_t^- \\ \hat{\mathbf{e}}_t(k) &= (1-\eta) \hat{\mathbf{e}}_t(k-1) + \eta \mathbf{c}_t(k) \end{aligned}$$

where $0 < \eta < 1$. If no measurement is available for an existing track, then the state estimate is updated using its prior estimate with a linearly increasing error covariance. Once the error covariance becomes too large after a certain frames, this track will be terminated.

### 3.2. Track Initialization

After checking measurements against existing tracks, there may be some measurements unmatched. Then those measurements, each from a different camera, are checked against each other to find potential new tracks. If there exists an unmatched measurement $\mathbf{z}_{j_1}^{(i_1)}$ from $i$-th camera, then a new track $x_n$ will be established. All the association matrices $\beta^{(i)}$ are extended by one column, each element of which indicates the correspondence between the new track and each measurement from *i*-th camera. For the $i_1$-th camera, the measurement $\mathbf{z}_{j_1}^{(i_1)}$ is automatically associated with the new track:

$$\beta_{jn}^{i_1} = \begin{cases} 1 & \text{if } j = j_1 \\ 0 & \text{otherwise} \end{cases}$$

For each unmatched measurement from the other cameras, $\mathbf{z}_j^{(i)}$, it is checked against the measurement $\mathbf{z}_{j_1}^{(i_1)}$ , and thought to be associated with the new track if the Mahalanobis distance:

$$d^2 = \left[ \mathbf{z}_j^{(i)} - \mathbf{z}_{j_1}^{(i_1)} \right]^T \left[ \mathbf{R}_j^{(i)} + \mathbf{R}_{j_1}^{(i_1)} \right]^{-1} \left[ \mathbf{z}_j^{(i)} - \mathbf{z}_{j_1}^{(i_1)} \right]$$

is within a validation gate and smallest in all the unmatched measurements from the *i*-th camera. Therefore, the new track can only be associated with at most one measurement from each camera. All the individual camera measurements assigned to the new track are then integrated into an overall measurement, $\mathbf{z}_n$, $\mathbf{R}_n$ and $\mathbf{c}_n$, as in Section 3.1. The new track is then initialised with the integrated measurement.

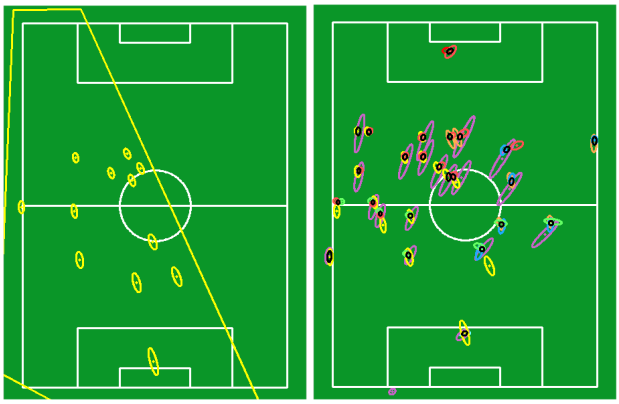

**Fig. 3**. Measurement covariance (left) from a single camera and (right) from multiple cameras with data integration (black ellipes).

### 3.3. Track Selection

This is a procedure of tracking aided recognition for the 25 most likely players. Due to false alarms and tracking errors, there normally exist more than 25 tracks for the players. A player likelihood measure is calculated for each target on the basis of confidence of category estimate, number of support cameras, domain knowledge in positions (for goalkeepers and linesmen), frames of being tracked or missing, as well as the fixed population constraint. A fast sub-optimal search method gives reasonable results.

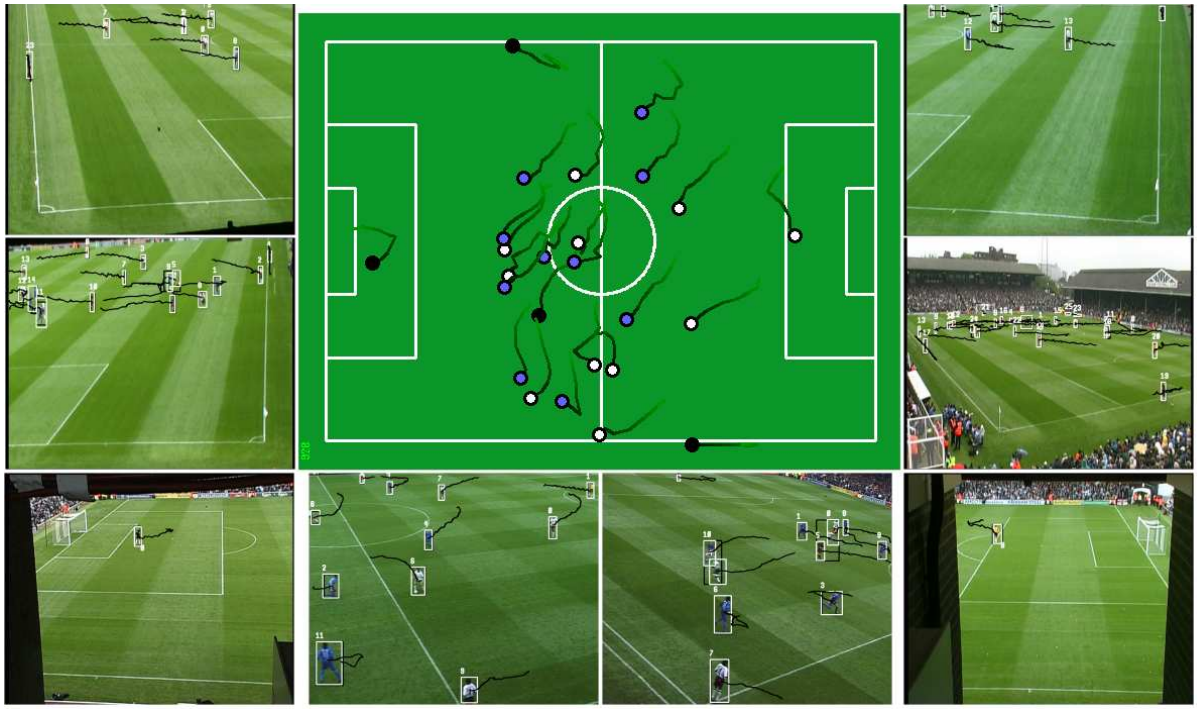

**Fig. 4**. The tracking results from single view trackers and multi-view tracker.

## 4. RESULTS AND DISCUSSION

The two-stage method outlined in this paper can be successfully demonstrated in several recorded matches. In Fig. 4, the central overhead plan shows actual working system output; the surrounding images show single view tracker output, at that frame, for cameras shown in Fig. 1.

There are several difficulties and opportunities to improve the reliability and accuracy of the system. Situations include many tightly packed players leading to tracking errors as the players re-disperse. In the limiting case, these situations are insoluble. In general several system components critically affect the performance of the system, and so merit attention for improvement. Firstly, there are sparse landmarks in a large area within the pitch and thus an accurate calibration and data association may suffer. Synthetic landmarks would be a remedy. Secondly, the single view tracker is designed to split the measurements for merged players. When more than two players grouped in the same foreground region, the uncertainty in estimation is large and the feedback from the multi-view tracker would be an advantage. Finally, probabilistic data association schemes, such as the Multiple Hypothesis Tracking and JPDAF [1], can be used to improve the tracking. It is worth noting that the aim of our system is not to track individual players but to recognise the players (teams) with the assistance of tracking. Therefore, the system can recover from a tracking failure by re-initialising and terminating tracks.

## 5. CONCLUSION

An application of people tracking has been presented that demonstrates successful modelling of football players using object detection and tracking, first in the image plane with single camera, and then in the ground plane using multiple cameras.

## 6. REFERENCES


[1] Y. Bar-Shalom and X.R. Li, Multitarget-Multisensor Tracking: Principles and Techniques, *YBS*, 1995.

[2] T. Bebie and H. Bieri, SoccerMan: reconstructing soccer games from video sequences, *Proc. ICPR*, pp. 898-902, 1998.

[3] S. S. Intille and A. F. Bobick, Closed-world tracking, *Proc. ICCV*, pp. 672-678, 1995.

[4] Y. Seo, S. Choi, H. Kim and K. S. Hong, Where are the ball and players?: Soccer game analysis with color-based tracking and image mosaic, *Proc. ICIAP*, pp. 196-203, 1997.

[5] M. J. Swain and D. H. Ballard, Colour indexing, *Int. J. Computer Vision*, 7(1):pp. 1132, 1991.

[6] R. Tsai, An efficient and accurate camera calibration technique for 3D machine vision, *Proc. CVPR*, pp. 323-344, 1986.

[7] M. Xu and T. Ellis, Partial observation vs. blind tracking through occlusion, *Proc. BMVC*, pp. 777-786, 2002.