Tracking when the camera looks away

Khurram Soomro, Salman Khokhar, Mubarak Shah
Center for Research in Computer Vision (CRCV), University of Central Florida (UCF)
{ksoomro, skhokhar, shah}@eecs.ucf.edu

Abstract

Tracking players in sports videos presents numerous challenges due to weak distinguishing features and unpredictable motion. Considerable work has been done to track players in such videos using a combination of appearance and motion modeling, mostly in continuous streams of video. However, in a broadcast sports video, having advertisements, replays and intermittent change of camera view, it becomes a challenging task to keep track of players over an entire game. In this work, we solve a novel problem of tracking over a sequence of temporally disjoint soccer videos without the use of appearance cue, using a Graph based optimization approach. Each team is represented by a graph, in which the nodes correspond to player positions and the edge weights depend on spatial inter-player distance. We use team formation to associate tracks between clips and provide an end-to-end system that is able to perform statistical and tactical analysis of the game. We also introduce a new challenging dataset of an international soccer game.

1. Introduction

In this paper we seek to use archived footage of past sporting events shown on television; and provide an end-to-end framework that is able to perform statistical and tactical analysis. In order to do that we need to be able to detect and track each individual over the course of the game. This however leads to a lot of challenges that arise due to the nature of broadcast soccer videos. As we can see in Fig. 1, the camera keeps shifting its focus from the soccer field to other things such as: the audience, zooming-in of players, goalpost view, etc. This makes it impossible to be able to have a consistent view of the players and track them individually. Hence, we are left with fragmented sequences of videos that give us a panoramic view of the field as shown in Fig. 1 (outlined in yellow). Since temporal video segmentation is beyond the scope of this work and has been extensively dealt with in computer vision literature, we assume segmention has been performed to exclude replays, advertisements or frames from camera positions close to the ground plane. The number of missing frames between useful continuous clips therefore could be large, hence the problem poses significant challenges in player tracking across clips, estimating player activity when outside field of view and in analyzing the combined strategy and actions of the entire group.

The problem of tracking across temporally disjoint clips is similar to that of person re-identification, where in order to maintain the track of an individual we need to be able to associate tracks in different clips. We make use of the fact that players in almost all team sports tend to arrange themselves in distinct formations and try to maintain these formations during short intervals or even for the entire length of the game. We model this group structure in a graph based framework and use it to estimate a best fit solution for global player identity assignment on a frame-by-frame basis, and use this information to assign identities to long-term agent activities.

Our novel contributions include: 1) introduction of a new problem of player role identification in temporally disjoint sports broadcast videos, 2) the use of team formation to perform player role identification, and 3) introduction of a new and challenging dataset for tracking and player role identification problems. Instead of using a single player motion model or extrapolation of tracks based on scene model, both of which do not work in the case of sports, we use a graph based model for tactical analysis, which has not been done before to the best of our knowledge. We seek to use singular and global motion information. Fig. 2. shows the track positions visible at the end of one clip and those visible at the beginning of the adjacent clip. This visually illustrates the difficulty and complexity of the problem, that with the presence of temporal gap between clips, the spatial location and arrangement of players varies to a huge ex-

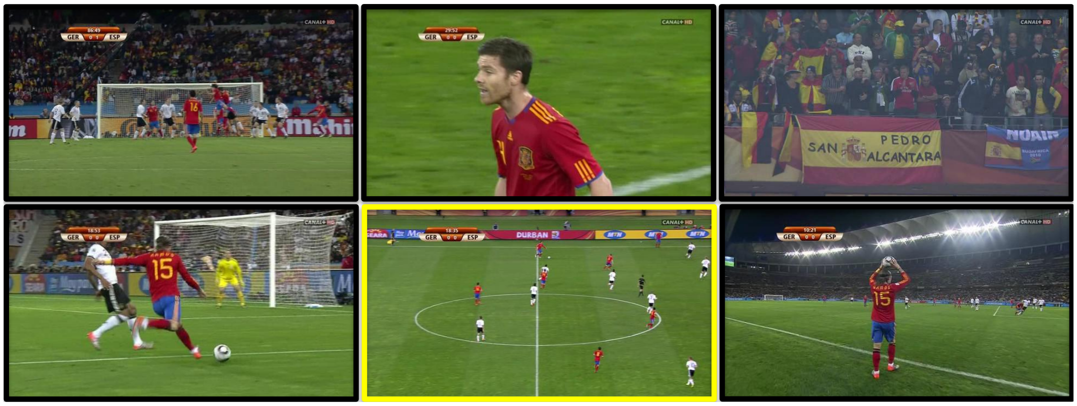


Figure 1. The figure shows a number of camera angles present in the broadcast video that we have used for our dataset. As we can see, tracking cannot be performed in most of the camera views except the one outlined in yellow.

tent. As is observed without contextual knowledge of the visible players in the entire formation it is not very easy for humans to judge player roles. In the next section we summarize the existing literature in sports and group video analysis and tracking. We then describe our tracking system in detail followed by graph based modeling and analysis, and finally we present the results. The final output is a tracking based tactical analysis of the entire game involving all players, that takes into account missing frames and unreliable tracking in the difficult scenario of team sports. We have collected our own dataset from a publicly available soccer game. The dataset consists of manually segmented clips for which ground truth tracks are available. Warping homographies from broadcast camera view to an orthogonal view of our soccer field model are also available. We plan to publicly release the dataset.

2. Related Work

Sports video analysis typically focuses on extracting highlights from sports videos. These systems often use additional cues for this task such as text and audio metadata [26, 22, 14], replays [23, 28], graphic overlays [34] and social media content [27].

The information that is most useful to soccer coaches and players is on team strategy and player performance. Recent years have seen a lot of work in this direction. Lucey *et al*. [19] analyze offensive and defensive formations of teams in basketball videos and the spatio-temporal changes in a team's formation. Tracking data from a large number of games has been used [21, 17] to build models for team behaviors during home/away games. The authors in [33, 30] build predictive models for near-future events or plays in a game. Lucey *et al*. [20] look at predicting scoring chances using short time intervals. Bialkowski *et al*. [7] use game stats, occupancy maps and formation estimates to get team identities. Wei *et al*. [31] estimate formations using spatio-temporal formation analysis, whereas Bialkowski *et al*. [6] estimate team activities using occupancy maps. Gyarmati and Anguera [11] extract knowledge about ball passing. Beetz and Gedikli [4] use classification tree to classify ball actions. Lucey *et al*. [18] perform a very similar task to ours, in that they estimate a player's role within a team in each frame. However in their setting, video input is from a set of eight cameras covering the entire playing area. Hence all players are visible at all times and cross clip track association is not needed.

While a lot of work has been done on tracking in general, we cover tracking in sports settings only. Tracking in sports presents unique challenges due to camera motion, person-person occlusion and complicated motion models. The authors in [16, 8] perform tracking using respectively a random forest of motion models for sports and a reversible jump MCMC framework. Xu *et al*. [32] present a framework for tracking specifically in multi-camera setting, while Khatoonabadi and Rahmati [15] use field lines to assist in video stabilization and tracking.

Person re-identification has been frequently researched in recent computer vision literature. Learning of discriminative appearance models in a single camera [25, 3, 9, 2] or multi-camera settings [13, 12] has been widely studied. Javed and Shah [12] have also looked at conformity to known paths or patterns to associate people. To the best of our knowledge, there is no work that performs re-identification in a moving camera video (such as for sports or high altitude surveillance) with large temporal gaps between clips and no guarantee of spatial overlap.

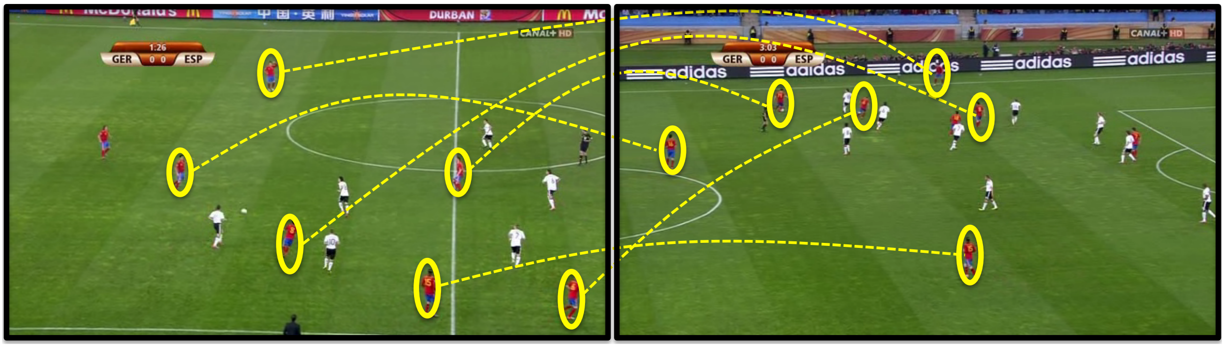


Figure 2. A figure illustrating the nature of the tracking problem we tackle in this paper. Appearance is not useful and motion cues are not reliable. We must use team strategy information to track in temporally disjoint clips.

3. Player Tracking and Role Identification

3.1. Problem description

Given a soccer broadcast game $\mathbb{M}$, we temporally segment it to obtain video clips with a panoramic view of the field. Let the index of broadcast video clips be $v = 1, \ldots, V$. The proposed approach begins by detecting players in each frame $f = 1, \ldots, \mathbf{F}_v$ of a video clip v. These detections are then associated across frames using a Greedy Bipartite (GB) [24] matching algorithm to obtain tracklets $\mathbf{t}_v = \{{}_n t_v\}_{n \in (1,\ldots,\mathbf{T}_v)}$. Since, it is challenging to track players in broadcast sports videos, due to moving camera, similar appearance and heavy occlusions, tracking produces several fragmented tracklets. These tracklets are merged together by a second application of GB algorithm to obtain merged tracklets $\mathbf{p}_v = \{{}_m p_v\}_{m \in (1,\ldots,\boldsymbol{\tau}_v)}$, where $\boldsymbol{\tau}_v \leq \mathbf{T}_v$.

As tracking is performed by associating player detections in consecutive frames in a panoramic-view video, it becomes impractical to associate tracks frame-wise when the camera looks away. Therefore, our contribution lies in maintaining player identity throughout the course of the broadcast game, despite temporal discontinuities in the panoramic-view of the game. Our proposed approach for player role identification uses team formation $\boldsymbol{\Phi} \in \boldsymbol{\pi}$, where $\boldsymbol{\pi}$ is a finite set of team formations, to associate player positions from each frame of the video.

In the following sub-sections we describe how players are detected and tracked in each video clip. Then we propose our approach for player role identification in temporally disjoint videos.

3.2. Multi-Player Detection and Tracking

3.2.1 Player Detection

Detecting players in soccer broadcast videos using standard human detectors [10] doesn't work very well due to high articulation of human body with varying poses and low player resolution. Players are detected in each frame by modeling the background using Gaussian Mixture Model (GMM). Firstly, the area within the frame corresponding to the playing field is identified to exclude image regions that correspond to crowd or stadia. Training samples of the field are used to determine the mean and variance for each color channel of the field. Using the background model, player blobs are detected as foreground. Each team's players are modeled using a color histogram. This histogram is used to classify blobs into one of the two participating team's players, so that each team's players can be tracked separately. The background subtraction method is noisy and does give blobs of various sizes and false detections, which are filtered based on size.

3.2.2 Player Tracking

In multi-target tracking players are associated between frames of a video using appearance, motion and player positions. Appearance plays an important role in distinguishing a player from the rest. However, in broadcast sports, appearance doesn't help in differentiating members of the same team. Therefore, we classify the set of detections in each frame into two groups: 1) Team 1 (*e.g. Spain*) and 2) Team 2 (*e.g. Germany*), based on their appearance features. Tracking is performed individually for each group, where players belonging to the same group are associated across frames. This helps avoid any inter-team occlusions and identity (ID) switches.

Greedy Bipartite (GB) matching algorithm is applied in a two-step approach, where initially it is used to associate detections (within a group) across consecutive frames to form short tracklets. These tracklets are then merged using GB (see Section 3.2.3) matching to get tracks for every player within a video clip.

GB algorithm uses distance matrix, computed using euclidean distance between player positions in each frame, to find tracklet associations. Before associating detections in each frame, we compensate for camera motion using COCOA [1]. The detections in current frame, which aren't associated with any tracklet are used to initialize a new tracklet; for those detections which get associated with current tracks, their state is updated accordingly. In case of a missing detection, a linear motion model is used to project the track in next frames. This generates a set of tracklets $\mathbf{t}_v = \{{}_n t_v\}_{n \in (1,\ldots,\mathbf{T}_v)}$, where ${}_n t_v$ = $[{}_n t_v^b, \ldots, {}_n t_v^e]$, $b{,}e \in \mathbf{F}_v$ is a set of player locations ${}_n t_v^f$ = $({}_n x_v^f, {}_n y_v^f)$ in each frame f of a video v; b and e represent beginning frame and ending frame of nth tracklet ${}_n t_v$, respectively.

3.2.3 Tracklet Merging

In a video sequence with irregular and non-linear player movements, fast camera motion, frequent occlusions and indistinct appearances, it becomes difficult to keep track of players and often gives us fragmented tracks. This problem can be solved by merging these tracklets to get one whole trajectory for each player over the course of an entire clip. In our case we have used GB matching algorithm to merge these tracklets.

To start with, we classify each tracklet ${}_n t_v$ as belonging to one of three possible cases: 1) $\mathbf{S_S}$ (Starting Tracklet); tracklet either started in the first frame of a video v or just entered the field of view, hence starting location of this tracklet would appear close to the frame edge, 2) $\mathbf{S_I}$ (Intermediate Tracklet); tracklet's starting and ending location isn't close to the frame edge, and 3) $\mathbf{S_T}$ (Terminating Tracklet); tracklet ended at the last frame $\mathbf{F}_v$ of a video clip v or is about to leave the field of view. To merge tracklets we start from a terminating tracklet and join it with either an intermediary or a starting tracklet, which satisfies spatio-temporal proximity constraints. We keep recurring this procedure till all tracks in $\mathbf{S_T}$ are merged with tracks in either $\mathbf{S_S}$ or $\mathbf{S_I}$, hence making them complete. This generates merged tracklets $\mathbf{p}_v = \{{}_m p_v\}_{m \in (1,\ldots,\boldsymbol{\tau}_v)}$, where ${}_m p_v$ = $[{}_m p_v^b, \ldots, {}_m p_v^e]$, $b{,}e \in \mathbf{F}_v$ is a set of player locations ${}_m p_v^f$ = $({}_m x_v^f, {}_m y_v^f)$.

The following section explains how player role is re-identified, after a temporal discontinuity in play. The tracks obtained from each individual clip v are connected across clips by associating track positions to player formation identities.

3.3. Player Role Identification in Temporally Disjoint Videos

Once we obtain the set of tracks $\mathbf{p}_v$ in each of the clips v used in our dataset, our next goal is to identify player role in each of the sets $\mathbf{p}_v$, $v \in \{1, 2, .., V\}$ to form a set of player tracks $\mathbf{r}_{\mathbb{M}}$ that spans over the entire game $\mathbb{M}$ and contains total number of tracks equaling twice the number of players in each team, $|\mathbf{\Phi}|$ (*11 in the case of soccer*). Fig. 2 visually illustrates the difficulty and complexity of the problem. As is observed without contextual knowledge of the visible players in the entire formation it is not very easy for humans to judge player roles.

In any given clip v, the set of visible tracks $\mathbf{p}_v$ only represents a subset of formation $\mathbf{\Phi}$. To assign every player in $\mathbf{p}_v$ a role from formation $\mathbf{\Phi}$, there can exist a number of possibilities. Similarly, when matching player identity from one clip to another, the number of combinations increases even further. Given the number of players $|\,\mathbf{p}_v|$ in clip v, and $|\,\mathbf{p}_w|$ in clip w, where $v, w \ \in \ V$, the possible number of solutions is ${}^{|\mathbf{\Phi}|}C_{|\,\mathbf{p}_v|} \times {}^{|\mathbf{\Phi}|}C_{|\,\mathbf{p}_w|}$.

A naive solution to player role identification problem can be to take the average location of each track within a clip and find a bipartite solution over a matrix of distances between the average track positions and model formation locations. This solution is ineffective because of two major reasons: 1) It is very difficult to get complete tracks within a clip due to large camera motion; and 2) players show large variation in their positions on the field based on the conditions of the game (See Fig. 4).

The problem is further complicated by the fact that not all players visible in a clip may be visible in a given frame of the clip. Therefore, we must analyze each frame separately. We propose a graph matching based voting method to determine player identity within one clip with respect to the model formation. The approach accumulates votes over frames using an optimization procedure explained in Section 3.3.3. This optimization is performed for each individual clip to find player role in a formation. This helps us in establishing consistent player roles across clips, under the assumption that the players retains same role in the formation over multiple video clips.

In the following sub-sections we will explain how we model team formation and estimate camera parameters

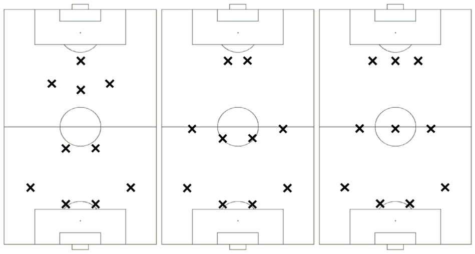


Figure 3. A figure showing three sample formations: from left to right: {4-2-3-1},{4-4-2},{4-3-3}

to project track positions onto the field model. Finally, we will provide the details of the graph matching based player role identification method.

3.3.1 Team Formation

Each team in a soccer game consists of 10 outfield players plus a goalkeeper. The manner in which players from one team, except the goalkeeper, arrange themselves relative to one another during a game is known as their formation, $\mathbf{\Phi} = \{\Phi_k\}_{k\in(1,\ldots,11)}$. Players arrange themselves in lines of defense, midfield and attack. Therefore, formations are defined by a string of numbers representing the number of players in subsequent lines from defensive to attacking players. Fig. 3 shows a visualization of three sample formations on a field model. A formation of {4-2-3-1} was used by both teams in our dataset. The formation of a team is available in pre-match team announcements.

3.3.2 Camera Parameter Estimation

To match track player positions in a frame to a model formation, we first need to project these positions to a field model by estimating camera paramters. Fig. 4 shows tracks in original frames being projected onto the field model. The extrinsic camera parameters for all camera frames were estimated using a wire frame tracking framework for field lines. To extract field lines, first field region is detected using GMM, where field color was used to remove the background, we then obtain edges within the field area and remove player detections using color information as well. Finally, we remove noisy edges using simple morphological operations. The resulting field lines are shown in Fig. 5. Each of these extracted field lines are then used to estimate camera parameters for the particular frame by searching over a range of possible parameters to get the best match over a field model. The field model was constructed according to FIFA (International Federation of Association Football) regulations, with a width to height ratio of 1.62. Fitness criteria was used as defined by Watanabe *et al.* [29].

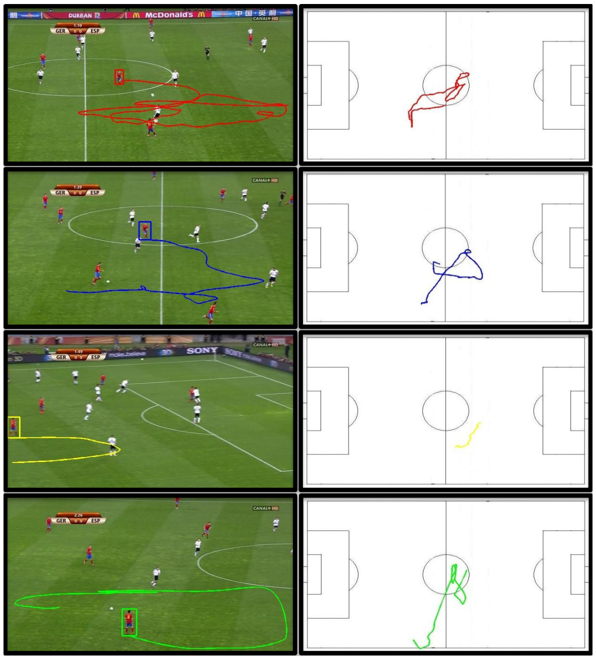


Figure 4. The figure shows the tracks from one player visible over multiple clips. Tracks in each clip are shown in a single color. We can observe large variation in position and large distances between track endpoints.

3.3.3 Single Team Player Role Identification

To establish player roles within a clip, we observe that although players tend to remain in relatively similar positions over the length of play, they may temporarily interchange or leave their positions depending on the state of play. Also, the actual positions are not fixed and the entire formation may move forwards or backwards on the pitch. Obviously determining the identity of one

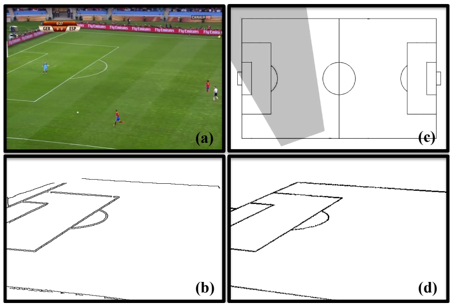


Figure 5. Steps involved in the estimation of camera parameters: (a) Original frame (b) Extracted field lines (c) Visible field region in original frame (d) Field model using camera parameters.

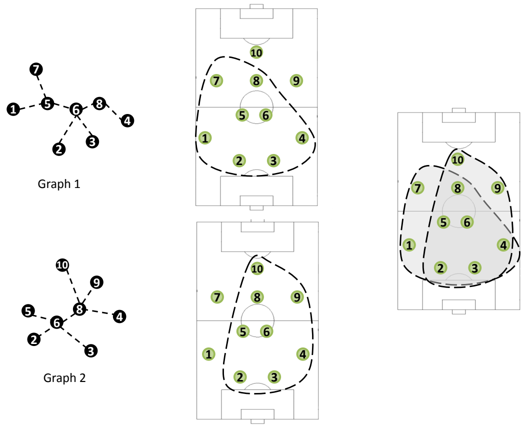


Figure 6. An outline of the track association across clips using formation cue. Graph 1 corresponds to the players visible in the first clip which are enclosed within the dotted line in the figure in the top-centre. Graph 2 and the figure in the bottom centre show the same for the adjacent clip. The figure on the right shows the players common to both clips.

player based on his instantaneous position is not possible. Our task is to identify player role in the formation model $\boldsymbol{\Phi}$. As we mentioned before a simple nearest neighbor approach will not work as the locations of the players change.

We formulate the problem of player role identification as a graph matching optimization problem. For each tracklet in $\mathbf{p}_v$ we associate it's position to a role in formation $\boldsymbol{\Phi}$. We form a graph ${}^{p}\mathbf{G}_v^f$ for every frame f in video v, in which nodes represent player positions p_v^f and edges denote inter-player distances. We match this graph with a set of graphs ${}^{\boldsymbol{\Phi}}\mathbf{G}_v^f = \{{}^{\boldsymbol{\Phi}}_{i}\mathbf{G}_v^f\}_{i \in (1,\ldots,|\boldsymbol{\Psi}_v^f|)}$, where $|\boldsymbol{\Psi}_v^f| = {}^{|\boldsymbol{\Phi}|}C_{|\,p_v^f|}$ is all possible combinations from the model of $|p_v^f|$ nodes. The nodes in this set of graphs represent relative player positions in the model formation given by $\boldsymbol{\Psi}_v^f = \{{}_{i}\boldsymbol{\Psi}_v^f\}_{i \in (1,\ldots,|\boldsymbol{\Psi}_v^f|)}$ and edges denote inter-player distances within the formation. We search over all player combinations from the model and compare them with actual player positions. For this comparision we minimize over two costs: 1) A deformation cost ${}^{\sigma}_{i}\mathbf{D}_v^f$ defined as the Bipartite cost over node-node distances for graphs ${}^{p}\mathbf{G}_v^f$ and ${}^{\boldsymbol{\Phi}}_{i}\mathbf{G}_v^f$. These distances are the summation of distances computed using original field co-ordinates and after registering both graph's means positions. This allows us to capture deformation between frame tracks and the model subgraph more accurately. 2) a spatial displacement cost defined as ${}^{\delta}_{i}\mathbf{D}_v^f = \|\,{}^{p}\boldsymbol{\mu}_v^f - {}^{\boldsymbol{\Phi}}_{i}\boldsymbol{\mu}_v^f\|$, where ${}^{p}\boldsymbol{\mu}_v^f$ and ${}^{\boldsymbol{\Phi}}_{i}\boldsymbol{\mu}_v^f$ are mean positions of ${}^{p}\mathbf{G}_v^f$ and ${}^{\boldsymbol{\Phi}}_{i}\mathbf{G}_v^f$ respectively. We generate a set of best player role candidates within the formation by optimizing over the following cost function:

$${}_{i}\boldsymbol{\Upsilon}_v^f = \underset{\mathbf{t}op-\lambda}{\arg\min}(\alpha.\,{}^{\delta}_{i}\mathbf{D}_v^f + \beta.\,{}^{\sigma}_{i}\mathbf{D}_v^f). \qquad (1)$$

We accumulate the votes for roles assigned to each tracklet and chose the role which has the maximum votes. Fig. 6 shows how two graphs from two different video clips are matched to model formation. The overlap of player roles in model formation shows how we can obtain continuous tracks of all players over the course of the game. Details of the entire algorithm can be found in Algorithm 1.

Algorithm 1 : Player Role Identification

Input: $\boldsymbol{\Psi}_v$ and $\mathbf{p}_v$
Output: Top λ elements of sorted ${}^{\boldsymbol{\Phi}}_{i}\mathbf{G}_v^f$

```
procedure PLAYERROLEID($\boldsymbol{\Psi}_v$,$\mathbf{p}_v$)
    for $f = 1$ to $\mathbf{F}_v$ do
        Generate Model Graph Combinations ${}^{\boldsymbol{\Phi}}\mathbf{G}_v^f$
        Generate Player Graph ${}^{p}\mathbf{G}_v^f$
        for all ${}^{\boldsymbol{\Phi}}_{i}\mathbf{G}_v^f$ in ${}^{\boldsymbol{\Phi}}\mathbf{G}_v^f$ do
            ${}^{p}\boldsymbol{\mu}_v^f$ = Mean of ${}^{p}\mathbf{G}_v^f$
            ${}^{\boldsymbol{\Phi}}_{i}\boldsymbol{\mu}_v^f$ = Mean of ${}^{\boldsymbol{\Phi}}_{i}\mathbf{G}_v^f$
            $\mathbf{R}$ = $\|\,{}^{p}\mathbf{G}_v^f - {}^{\boldsymbol{\Phi}}_{i}\mathbf{G}_v^f\|$
            ${}^{\delta}_{i}\mathbf{D}_v^f = \|\,{}^{p}\boldsymbol{\mu}_v^f - {}^{\boldsymbol{\Phi}}_{i}\boldsymbol{\mu}_v^f\|$
            ${}^{\sigma}_{i}\mathbf{D}_v^f$ : Bipartite cost over $\mathbf{R}$
            ${}_{i}\boldsymbol{\Upsilon}_v^f = \alpha.\,{}^{\delta}_{i}\mathbf{D}_v^f + \beta.\,{}^{\sigma}_{i}\mathbf{D}_v^f$
        end for
        Sort ${}^{\boldsymbol{\Phi}}_{i}\mathbf{G}_v^f$ by ${}_{i}\boldsymbol{\Upsilon}_v^f$
        Return top $\lambda$ elements of sorted ${}^{\boldsymbol{\Phi}}_{i}\mathbf{G}_v^f$
    end for
end procedure
```

3.3.4 Joint Player Role Identification of Competing Teams

In all cases of competitive behaviors that involve two or more agents, the strategy of one team is closely linked with the strategies adopted by the other teams. In soccer, therefore we observe a high correlation between the activitites or states of opposing teams. Correlation may also be observed between individual players of opposing teams as players are often assigned the task to counter specific individuals. Therefore, the behavior of one team can serve as a reliable cue in the estimation of the behavior of the other team. This is however a chicken and egg problem as there is no obvious order by which we can analyze the two behaviors one after the other.

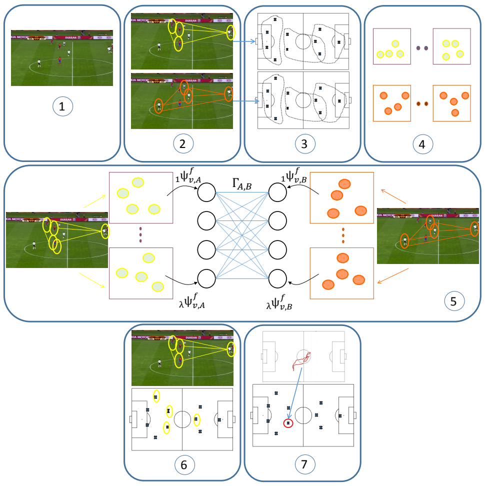


Figure 7. An explanation of the simultaneous best fit estimation for both teams' formation identities in a given frame. 1: input frame. 2: Graphs constructed from detections for both teams. 3: Graphs matched against candidates from model subsets. 4: Best candidate solutions are chosen for each team. 5: A simultaneous solution for best assignments for both teams is estimated. 6: Each track point is assigned a formation identity. 7: An entire track is assigned an identity based on vote pooling.

Therefore, we must compute a simultaneous likelihood of joint events which we do so using high probability candidate events for each team. Fig. 7 explains the interaction between the two states and the cues we use to estimate them.

A candidate solution for assigning formation identities to one team's players in a frame is denoted by ${}^{\mathbf{\Phi}}_{\mathbf{A}_i}\mathbf{G}^f_v$, where $\mathbf{A}$ denotes team identity. At the same frame, a candidate solution for a second team (denoted by $\mathbf{B}$) is ${}^{\mathbf{\Phi}}_{\mathbf{B}_i}\mathbf{G}^f_v$. We divide members of each team into three types, $\boldsymbol{\kappa} \in \{defender, midfielder, attacker\}$. We know that at any one time, typically one team is on the attack while the other is on the defense. Hence, if players from a team in one frame are mostly attacking players, it is likely that players from the other team in camera view will be defensive. Let us define sets $\mathbf{J_A}$ and $\mathbf{J_B}$ as the sets of player types in frame f of clip v for each team. The notation $\mathbf{\Gamma}_{\mathbf{A}_i,\mathbf{B}_j}$ is defined as the affinity between candidates ${}^{\mathbf{\Phi}}_{\mathbf{A}_i}\mathbf{G}^f_v$ and ${}^{\mathbf{\Phi}}_{\mathbf{B}_i}\mathbf{G}^f_v$ from the two teams respectively.

$$\mathbf{\Gamma}_{\mathbf{A}_i,\mathbf{B}_j} = |\mathbf{J_A} \cup \mathbf{J_B}| - |\mathbf{J_A} \cap \mathbf{J_B}| / |\mathbf{J_A} \cup \mathbf{J_B}| \quad (2)$$

This is the Jaccard distance of the two sets $\mathbf{J_A}$ and $\mathbf{J_B}$. Once we obtain the set of candidate solutions for both teams for a given frame we choose a certain number λ of highly likely solutions and form an affinity matrix $\mathbf{\Gamma}_{\mathbf{A}_i,\mathbf{B}_j}$ between candidate pairs ${}^{\mathbf{\Phi}}_{\mathbf{A}_i}\mathbf{G}^f_v$ and ${}^{\mathbf{\Phi}}_{\mathbf{B}_i}\mathbf{G}^f_v$ where $i, j \in 1, 2, .., \lambda$. The final solutions for ${}^{\mathbf{\Phi}}_{\mathbf{A}_i}\mathbf{G}^f_v$ and ${}^{\mathbf{\Phi}}_{\mathbf{B}_i}\mathbf{G}^f_v$ are chosen by optimizing over the equation:

Video #	1	2	3	4	5
MOTA (%)	82.07	85.01	80.02	80.17	79.43
MOTP (%)	89.24	93.13	82.59	81.23	84.18
ID switches	5	1	10	3	1
Frames	945	74	401	250	221

Table 1. Results for tracking within each clip.

Video #	1	2	3	4	5
Team 1 (Spain)	52	75	50	73	67
Team 2 (Germany)	50	78	60	-	-

Table 2. Results showing accuracy (%) for player role identity estimation of tracklets.

$$\mathbf{\Xi} = \arg\min_{i,j}({}_{\mathbf{A}_i}\mathbf{\Upsilon}^f_v.\omega_{\mathbf{A}} + {}_{\mathbf{B}_j}\mathbf{\Upsilon}^f_v.\omega_{\mathbf{B}} + \mathbf{\Gamma}_{\mathbf{A}_i,\mathbf{B}_j}.\omega_\gamma). \quad (3)$$

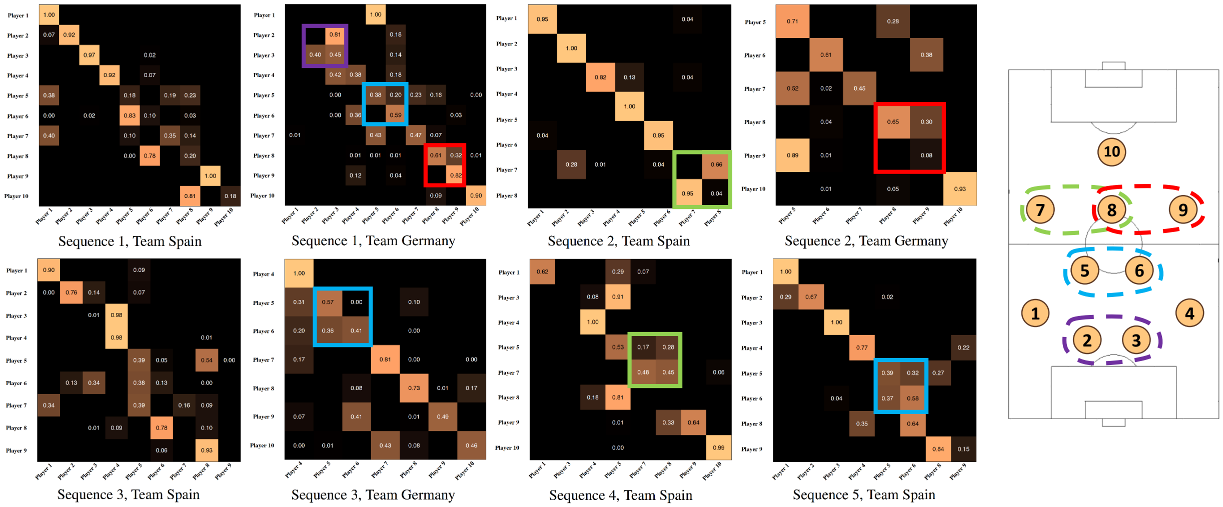
4. Experiments

4.1. Soccer Broadcast Dataset

We have collected our own dataset from the television broadcast of an international soccer game. We have manually segmented the video to extract only the segments that show the field from an almost overhead view from the sidelines. We provide $V = 5$ such clips, including track annotations covering almost 1900 frames. We intend to make the dataset publicly available. Manually computed tracks as well as homographies, that warp every frame of a video clip to the field model, are available.

4.2. Experimental Results

We perform tracking over 5 disjoint clips from a soccer game. The initial tracklets are formed using Greedy Bipartite and then all of them are merged together to get one whole trajectory using Bipartite Graph matching. The results are shown in Table 1 and are calculated using CLEAR MOT [5] metrics. MOTA measures the rate of false positives, false negatives and ID switches over all tracks in each video, where as MOTP is related to



Sequence 1, Team Spain

Sequence 1, Team Germany

Sequence 2, Team Spain

Sequence 2, Team Germany

Sequence 3, Team Spain

Sequence 3, Team Germany

Sequence 4, Team Spain

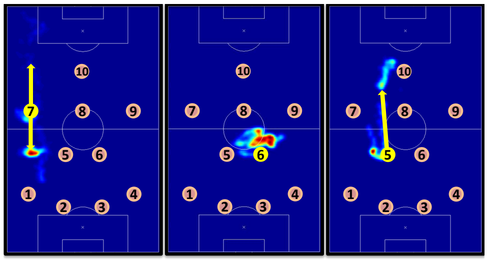
Sequence 5, Team Spain

Figure 9. Confusion matrices between players of a team in one clip. The diagonal should ideally be large. Large non-diagonal elements indicate confusion between corresponding players. Each color represents a pair of player roles.

precisely locating the object in each frame. Our results show that we achieve an average accuracy of 81.5%.

Once we have the tracks for each sequence, our target is to link all these tracks, which means we should be able to identify each player in all the videos.

To test our identity estimation method within the clips, we perform the method outlined in the algorithm given in Algorithm 1 on each clip, on the tracks generated automatically. We then compare the estimated identity with the ground truth identity for each tracklet. The results for each clip are given in Table 2. The combined accuracy is 58.4%, while the probability of randomly choosing the correct answer is $< 1\%$ (Since track labels are not independent, this is calculated as $1/{}^{11}C_7$, where 7 is the average number of players per clip). This is notably high considering that a fair evaluation needs ground truth that takes into account players switching roles for brief periods of time.



(a) Left Midfielder (b) Defensive Midfielder (c) Offensive Midfielder

Figure 8. Kernel Density Estimation over tracks of the highlighted player (yellow). Arrows indicate player movement direction.

5. Analysis and Discussion

With cross clip tracking, we can retain the identity of a track over the course of a game by associating each clip to a known formation, allowing us to compute individual player statistics over entire games. On the other hand, qualitative analysis can be done using Kernel Density Estimation of the tracks of either an individual or an entire team. This helps us in determining a player's strategy during the game and also analyze whether or not he was successful in that strategy. Fig. 8 illustrates the Kernel Density Estimation of the tracks of three midfielders. It can be observed that although all three are midfielders, they have very different strategies and participate in different kinds of plays. Fig. 9 shows confusion matrices built from votes accumulated from each of our clips. This allows us to observe that adjacent players from the model are often confused. The most confusing pairs are shown in colored boxes and highlighted on the field model on the right of the figure.

6. Conclusion

Our work shows that with the use of formations we can associate tracks between consecutive clips, which is possible due to the fact that players remain in an organized manner. We believe that this work is applicable to other forms of sports and aerial surveillance of multiple agents, where well structured group activity can be observed. Performing statistical and tactical analysis using tracks obtained over long durations wouldn't have been possible in archived broadcast sports videos, where multi-camera systems were not available.

References

[1] S. Ali and M. Shah. Cocoa: tracking in aerial imagery. In *Defense and Security Symposium*, 2006. 4

[2] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Boosted human re-identification using riemannian manifolds. *IVC*, 30(6), 2012. 2

[3] L. Bazzani, M. Cristani, and V. Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. *CVIU*, 117(2), 2013. 2

[4] M. Beetz, N. v. Hoyningen-Huene, J. Bandouch, B. Kirchlechner, S. Gedikli, and A. Maldonado. Camera-based observation of football games for analyzing multi-agent activities. In *AAMAS*, 2006. 2

[5] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *IVP*, 2008, 2008. 7

[6] A. Bialkowski, P. Lucey, P. Carr, S. Denman, I. Matthews, and S. Sridharan. Recognising team activities from noisy data. In *CVPRW*, 2013. 2

[7] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews. Identifying team style in soccer using formations learned from spatiotemporal tracking data. In *ICDMW*, 2014. 2

[8] R. T. Collins and P. Carr. Hybrid stochastic/deterministic optimization for tracking sports players and pedestrians. In *ECCV*. 2014. 2

[9] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 2

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32(9), 2010. 3

[11] L. Gyarmati and X. Anguera. Automatic extraction of the passing strategies of soccer teams. *arXiv preprint arXiv:1508.02171*, 2015. 2

[12] O. Javed, K. Shafique, Z. Rasheed, and M. Shah. Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views. *CVIU*, 109(2), 2008. 2

[13] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *CVPR*, 2005. 2

[14] Y.-L. Kang, J.-H. Lim, M. S. Kankanhalli, C.-S. Xu, and Q. Tian. Goal detection in soccer video using audio/visual keywords. In *ICIP*, 2004. 2

[15] S. H. Khatoonabadi and M. Rahmati. Automatic soccer players tracking in goal scenes by camera motion elimination. *IVC*, 27(4), 2009. 2

[16] J. Liu and P. Carr. Detecting and tracking sports players with random forests and context-conditioned motion models. In *Computer Vision in Sports*, pages 113–132. Springer, 2014. 2

[17] P. Lucey, A. Bialkowski, P. Carr, E. Foote, and I. Matthews. Characterizing multi-agent team behavior from partial team tracings: Evidence from the english premier league. In *AAAI*, 2012. 2

[18] P. Lucey, A. Bialkowski, P. Carr, S. Morgan, I. Matthews, and Y. Sheikh. Representing and discovering adversarial team behaviors using player roles. In *CVPR*, 2013. 2

[19] P. Lucey, A. Bialkowski, P. Carr, Y. Yue, and I. Matthews. how to get an open shot: Analyzing team movement in basketball using tracking data. MIT Sloan, 2014. 2

[20] P. Lucey, A. Bialkowski, M. Monfort, P. Carr, and I. Matthews. quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. MIT Sloan, 2014. 2

[21] P. Lucey, D. Oliver, P. Carr, J. Roth, and I. Matthews. Assessing team strategy using spatiotemporal data. In *KDD*, 2013. 2

[22] P. Oskouie, S. Alipour, and A. M. Eftekhari-Moghadam. Multimodal feature extraction and fusion for semantic mining of soccer video: a survey. *AIR*, 42(2), 2014. 2

[23] H. Pan, P. Van Beek, and M. I. Sezan. Detection of slow-motion replay segments in sports video for highlights generation. In *ICASSP*, 2001. 2

[24] K. Rangarajan and M. Shah. Establishing motion correspondence. In *CVPR*, 1991. 3

[25] W. R. Schwartz and L. S. Davis. Learning discriminative appearance-based models using partial least squares. In *SIBGRAPI*, 2009. 2

[26] Y. Takahashi, N. Nitta, and N. Babaguchi. Automatic video summarization of sports videos using metadata. In *PCM*. 2005. 2

[27] A. Tang and S. Boring. # epicplay: crowd-sourcing sports video highlights. In *SIGCHI*, 2012. 2

[28] L. Wang, X. Liu, S. Lin, G. Xu, and H.-Y. Shum. Generic slow-motion replay detection in sports video. In *ICIP*, 2004. 2

[29] T. Watanabe, M. Haseyama, and H. Kitajima. A soccer field tracking method with wire frame model from tv images. In *ICIP*, volume 3, 2004. 5

[30] X. Wei, P. Lucey, S. Vidas, S. Morgan, and S. Sridharan. Forecasting events using an augmented hidden conditional random field. In *ACCV*. 2014. 2

[31] X. Wei, L. Sha, P. Lucey, S. Morgan, and S. Sridharan. Large-scale analysis of formations in soccer. In *DICTA*, 2013. 2

[32] M. Xu, J. Orwell, L. Lowey, and D. Thirde. Architecture and algorithms for tracking football players with multiple cameras. *VISP*, 152(2), 2005. 2

[33] Y. Yue, P. Lucey, P. Carr, A. Bialkowski, and I. Matthews. Learning fine-grained spatial models for dynamic sports play prediction. In *ICDM*, 2014. 2

[34] H. M. Zawbaa, N. El-Bendary, A. E. Hassanien, and T. Kim. Event detection based approach for soccer video summarization using machine learning. *IJMUE*, 7(2), 2012. 2