2013 IEEE Conference on Computer Vision and Pattern Recognition

Representing and Discovering Adversarial Team Behaviors using Player Roles

Patrick Lucey[1], Alina Bialkowski[1,2], Peter Carr[1], Stuart Morgan[3], Iain Matthews[1] and Yaser Sheikh[4]
[1]Disney Research, Pittsburgh, USA, [2]Queensland University of Technology, Australia,
[3]Australian Institute of Sport, Australia, [4]Carnegie Mellon University, Pittsburgh, USA
{patrick.lucey,alina.bialkowski,peter.carr,iainm}@disneyresearch.com
stuart.morgan@ausport.gov.au, yaser@cs.cmu.edu

Abstract

*In this paper, we describe a method to represent and discover adversarial group behavior in a continuous domain. In comparison to other types of behavior, adversarial behavior is heavily structured as the location of a player (or agent) is dependent both on their teammates and adversaries, in addition to the tactics or strategies of the team. We present a method which can exploit this relationship through the use of a spatiotemporal basis model. As players constantly change roles during a match, we show that employing a "role-based" representation instead of one based on player "identity" can best exploit the playing structure. As vision-based systems currently do not provide perfect detection/tracking (*e.g. *missed or false detections), we show that our compact representation can effectively "denoise" erroneous detections as well as enabling temporal analysis, which was previously prohibitive due to the dimensionality of the signal. To evaluate our approach, we used a fully instrumented field-hockey pitch with 8 fixed high-definition (HD) cameras and evaluated our approach on approximately 200,000 frames of data from a state-of-the-art real-time player detector and compare it to manually labelled data.*

1. Introduction

When a group of individuals occupies a space, such as a crowd in a foyer or a gathering at a public square, recognizable patterns of interaction occur opportunistically (*e.g.* people moving to avoid collisions [23]) or because of structural constraints (*e.g.* divergence around lamp-posts [17]). When these individuals form competitive cliques, as seen in games on a sports field, distinct and deliberate patterns of activity emerge in the form of plays, tactics, and strategies. In the former case, each individual pursues an individual goal on their own schedule; in the latter, the teams engage in adversarial goal-seeking usually under the synchronized direction of a captain or a coach. Identifying these emergent patterns of play is critical to understanding the evolving game for fans, players, coaches, and broadcasters (including commentators, camera operators, producers, and game statisticians).

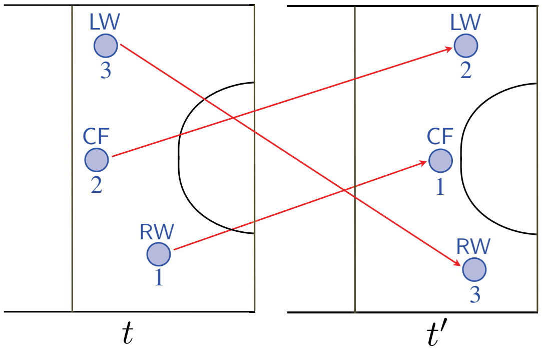

Figure 1. We can identify a player by their name or number (*e.g.* 1, 2 or 3) or via their formation role (*e.g.* left wing LW, center forward CF and right wing RW). Given two snapshots of play at time t and t', using player identity $(1, 2,$ and $3)$ the two snapshots will look different as the players have swapped positions. However, if we disregard identity and use role (LW, CF, RW), the arrangements are similar which yields a more compact representation and allows for generalization across games.

The behavior of a team may be described by how its members cooperate and contribute in a particular situation. In team sports, the overall style of a team can be characterized by a *formation*: a coarse spatial structure which the players maintain over the course of the match. Additionally, player movements are governed by physical limits, such as acceleration, which makes trajectories smooth over time. These two observations suggest significant correlation (and therefore redundancy) in the spatiotemporal signal of player movement data. A core contribution of this work is to recover a low-dimensional approximation for a time series of player locations. The compact representation is critical for understanding team behavior. First, it enables the recovery of a true underlying signal from a set of noisy detections. Second, it allows for efficient clustering and retrieval of game events.

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.349

A key insight of this work is that even perfect tracking data is not sufficient for understanding team behavior. A formation implicitly defines a set of *roles* or individual responsibilities which are then distributed amongst the players by the captain or coach. In dynamic games like soccer or field hockey, it may be opportunistic for players to swap roles (either temporarily or permanently). As a result, when analyzing the strategy of a particular game situation, players are typically identified by the role they are currently playing and not necessarily by an individualistic attribute like name (*e.g.* Figure 1).

In this paper, we have two contributions: 1) we present a *representation* based on player role which provides a more compact representation compared to player identity, and allows us to use subspace methods such as the bilinear spatiotemporal basis model [4] to "denoise" noisy detections (which is common from a vision system); and 2) we show that we can effectively *discover* team formation and plays using the role representation. Identifying formations and plays quickly from a large repository could enhance sports commentary by highlighting recurrent team strategies and long term trends in a sport. The process of post-game annotation, which coaches and technical staff spend hours performing manually, could be automated enabling more detailed data mining. Additionally, understanding plays in realtime, is a step towards a fully automated sports broadcasting system. We demonstrate our ideas on approximately $200k$ frames of data acquired from a state-of-the-art realtime player detector [10] and compare it to manually labelled data.

2. Related Work

Recent work in the computer vision community has evolved from action and activity recognition of a single person [19, 1], to include entire groups of people [14, 15, 26]. Research into group behavior can be broken into two areas: 1) crowd analysis, and 2) group analysis. Crowds consist of individuals attempting to achieve goals independent of other individuals in the group. Most of the research in this area has focussed on multi-agent tracking [5, 23] and anomaly detection [32].

Due to the host of military, surveillance and sport applications, research into recognizing group behavior has increased recently. Outside of the sport realm, Sukthankar and Sycara recognized group activities for dynamic teams [30]. Sadilek and Kautz [27] used GPS locations of multiple agents in a "capture the flag" game to recognize low-level activities. Recently, Zhang et al. [34] used a "bag of words" and SVM approach to recognize group activities in a prison setting. Sport related research mostly centers on low-level activity detection with the majority focussed on American Football. In the initial work by Intille and Bobick [14], they recognized a single football play, using a Bayesian network to model the interactions between the players trajectories. Li et al. [21] and Siddiquie et al. [28], used spatiotemporal models to classify different offensive plays. Li and Chellapa [20] used a spatio-temporal driving force model to segment the two groups/teams using their trajectories. Researchers at Oregon State University have looked at automatically detecting offensive plays from raw video and transfer this knowledge to a simulator [29]. For soccer, Kim et al. [16] used the global motion of all players in a soccer match to predict where the play will evolve in the short-term. Beetz et al. [7] proposed a system which aims to track player and ball positions via a vision system for the use of automatic analysis of soccer matches. In basketball, Perse et al. [24] used trajectories of player movement to recognize three types of team offensive patterns. Morariu and Davis [22] integrated interval-based temporal reasoning with probabilistic logical inference to recognize events in one-on-one basketball. Hervieu et al. [13] also used player trajectories to recognize low-level team activities using a hierarchical parallel semi-Markov model. In addition to these works, plenty of work has centered on analyzing broadcast footage of sports for action, activity and highlight detection [33, 12][1].

3. Adversarial Player Movements

In this work, we investigate the behaviors of several international field hockey teams. Games from an international hockey tournament of 24 games was recorded using eight stationary HD cameras mounted on the stadium lighting which collectively covered the entire $91.4\text{m} \times 55.0\text{m}$ playing surface. A state-of-the-art player detector [10] generated a series $\mathcal{O}$ of observations where each observation consisted of an (x, y) ground location, a timestamp t, and a team affiliation estimate $\tau \in \{\alpha, \beta\}$.

At any given time instant t, the set of detected player locations $\mathcal{O}_t = \{x_A, y_A, x_B, y_B, \dots\}$ is of arbitrary length. Generally, the number of detections N_t at time t is not equal to the number of players P because some players may not have been detected and/or background clutter may have been incorrectly classified as a player.

Typically, the goal is to track all $2P$ players over the duration of the match. In field hockey, that corresponds to 20 players ($P = 10$ per team ignoring goalkeepers) and two 35 minute long halves. The task of tracking all players across time is equivalent to generating a vector of ordered player locations $\mathbf{p}_t^\tau = [x_1, y_1, x_2, y_2, \dots, x_P, y_P]^\intercal$ for each team τ from the noisy detections $\mathcal{O}_t$ at each time instant. The particular ordering of players is arbitrary, but must be consistent across time. Therefore, we will refer to $\mathbf{p}_t^\tau$ as a *static labeling* of player locations. It is important to point out that

[1] These works only capture a portion of the field, making group analysis difficult as all active players are rarely present in the all frames.

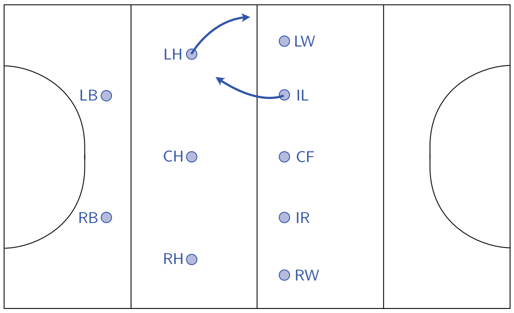

Figure 2. The dynamic nature of the game requires players to switch roles and responsibilities on occasion, for example, the left halfback LH overlaps with the inside left IL to exploit a possible opportunity.

Match ID	Game	No. Frames
1	USAvsRSA-1	3894
2	USAvsRSA-2	8839
3	USAvsJPN-1	4855
4	USAvsJPN-2	7418

Table 1. We manually labelled player location, identity and role at each frame for parts of four games from an international field-hockey tournament.

$\mathbf{p}_t^{\mathcal{T}}$ is not simply a subset of $\mathcal{O}_t$. If a player was not detected, an algorithm will somehow have to infer the (x, y) location of the unseen player based on spatiotemporal correlations.

We focus on generic team behaviors and assume any observed arrangement of players from team α could also have been observed for players from team β. As a result, there is a 180° symmetry in our data. For any given vector of player locations $\mathbf{p}_t^{\mathcal{T}}$, there is an equivalent complement $\overleftrightarrow{\mathbf{p}}_t^{\mathcal{T}}$ from rotating all (x, y) locations about the center of the field and swapping the associated team affiliations.

3.1. Formations and Roles

In the majority of team sports, the coach or captain designates an overall structure or system of play for a team. In field hockey, the structure is described as a *formation* involving *roles* or individual responsibilities (see Fig. 2). For instance, the 5:3:2 formation defines a set of roles $\mathcal{R} = \{$left back (LB), right back (RB), left halfback (LH), center halfback (CH), right halfback (RH), inside left (IL), inside right (IR), left wing (LW), center forward (CF), right wing (RW)$\}$. Each player is assigned exactly one role, and every role is assigned to only one player. Generally, roles are not fixed. During a match, players may swap roles and temporarily adopt the responsibilities of another player. Mathematically, assigning roles is equivalent to permuting the player ordering $\mathbf{p}_t^{\mathcal{T}}$. We define a $P \times P$ permutation matrix $\mathtt{x}_t^{\mathcal{T}}$ at time t which describes the players in terms of roles $\mathbf{r}_t^{\mathcal{T}}$

$$\mathbf{r}_t^{\mathcal{T}} = \mathtt{x}_t^{\mathcal{T}} \mathbf{p}_t^{\mathcal{T}} \tag{1}$$

By definition, each element $\mathtt{x}_t^{\mathcal{T}}(i, j)$ is a binary variable, and every column and row in $\mathtt{x}_t^{\mathcal{T}}$ must sum to one. If $\mathtt{x}_t^{\mathcal{T}}(i, j) = 1$ then player i is assigned role j. In contrast to $\mathbf{p}_t^{\mathcal{T}}$, we refer to $\mathbf{r}_t^{\mathcal{T}}$ as a *dynamic labeling* of player locations.

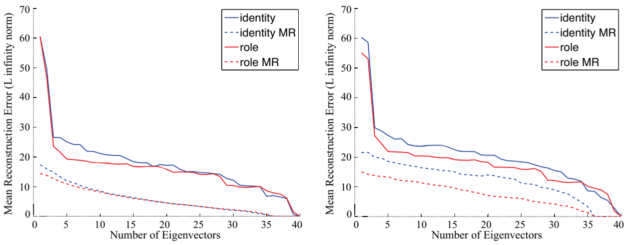

Figure 3. Plot showing the reconstruction error as a function of the number of eigenvectors used to reconstruct the signal using the L_∞ norm for non-mean and mean-normalized features for both identity and role representations on train (seen) and test (unseen) data.

Because the spatial relationships of a formation are defined in terms of roles (and not individualistic attributes like name) and players swap roles during the game, we expect the spatiotemporal patterns in $\{\mathbf{r}_1^{\mathcal{T}}, \mathbf{r}_2^{\mathcal{T}}, \ldots, \mathbf{r}_T^{\mathcal{T}}\}$ to be more compact compared to $\{\mathbf{p}_1^{\mathcal{T}}, \mathbf{p}_2^{\mathcal{T}}, \ldots, \mathbf{p}_T^{\mathcal{T}}\}$. Additionally, we expect a team to maintain its formation while moving up and down the field. As a result, position data $\tilde{\mathbf{r}}_t^{\mathcal{T}}$ expressed relative to the mean (x, y) location of the team should be even more compressible. To test these conjectures, we manually tracked all players over 25000 time-steps (which equates to $8 \times 25000 = 200,000$ frames across 8 cameras), and asked a field hockey expert to assign roles to the player locations in each frame. A breakdown of the manually labelled data is given in Table 1.

For brevity, we explain the analysis in terms of roles $\mathbf{r}_t^{\mathcal{T}}$ since the original player ordering $\mathbf{p}_t^{\mathcal{T}}$ is just a special non-permuted case $\mathtt{x}_t^{\mathcal{T}} = \mathtt{I}$. We ran PCA on the temporal data series produced by both teams $\{\mathbf{r}_1^{\mathcal{T}}, \mathbf{r}_2^{\mathcal{T}}, \ldots, \mathbf{r}_{25000}^{\mathcal{T}}, \overleftrightarrow{\mathbf{r}}_1^{\mathcal{T}}, \overleftrightarrow{\mathbf{r}}_2^{\mathcal{T}}, \ldots, \overleftrightarrow{\mathbf{r}}_{25000}^{\mathcal{T}}\}$. This was to measure how well the low-dimensional representation $\hat{\mathbf{r}}_t^{\mathcal{T}}$ matches the original data $\mathbf{r}_t^{\mathcal{T}}$ using the L_∞ norm of the residual $\Delta\mathbf{r} = \hat{\mathbf{r}}_t^{\mathcal{T}} - \mathbf{r}_t^{\mathcal{T}}$

$$\|\Delta\mathbf{r}\|_\infty = \max(\|\Delta\mathbf{r}(1)\|_2, \ldots, \|\Delta\mathbf{r}(P)\|_2) \tag{2}$$

where $\|\Delta\mathbf{r}(p)\|_2$ is the L_2 norm of the p^{th} x and y components of $\Delta\mathbf{r}$. We chose the L_∞ norm instead of the L_2 norm because large deviations may signify very different formations, *e.g.* a single player could be breaking away to score. Figure 3 illustrates how both $\mathbf{p}_t^{\mathcal{T}}$ and $\mathbf{r}_t^{\mathcal{T}}$ are quite compressible on the training data. However, when we test on unseen data (with role labels), the dynamic role-based ordering $\mathbf{r}_t^{\mathcal{T}}$ is much more compressible than the static ordering $\mathbf{p}_t^{\mathcal{T}}$. Relative positions are more compressible than absolute positions in both orderings.

3.2. Incorporating Adversarial Behavior

A player's movements are correlated not only to teammates but to opposition players as well. Therefore, we anticipate that player location data can be further compressed

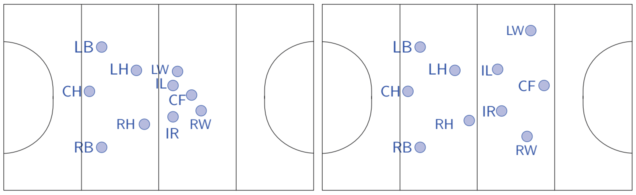

Figure 4. Examples showing the difference between the mean formations using the: (left) identity and (right) role representations on one of the matches.

if the locations of players on teams A and B are concatenated into a single vector $\mathbf{r}_t^{AB} = [\mathbf{r}_t^A, \mathbf{r}_t^B]^\mathsf{T}$.

In Figure 4, we show the mean formations for the identity and role representation. We can see that the role representation has a more uniform spread between the players, while the identity representation has a more crowded shape, which highlights the constant swapping of roles during a match. In terms of compressibility, Table 2 shows that using an adversarial representation gains better compressibility for both cases, and that using both a role and adversarial representation yields the most compressibility.

3.3. Bilinear Spatiotemporal Analysis

The representation of time-varying spatial data is a well-studied problem in computer vision (see [9] for overview). Recently, Akhter *et al.* [4], presented a bilinear spatiotemporal basis model which captures and exploits the dependencies across both the spatial and temporal dimensions in an efficient and elegant manner, which can be applied to our problem domain. Given we have P players per team, we can form our role-based adversarial representation, $\mathbf{x}$, as a spatiotemporal structure $\mathbf{S}$, given $2P$ total players sampled at F time instances as

$$\mathbf{S}_{F\times 2P} = \begin{bmatrix} x_1^1 & \dots & x_{2P}^1 \\ \vdots & & \vdots \\ x_1^F & \dots & x_{2P}^F \end{bmatrix} \tag{3}$$

where x_j^i denotes the jth index within the role representation at the ith time instant. Thus, the time-varying structure matrix $\mathbf{S}$ contains $2FP$ parameters. This representation of the structure is an over parameterization because it does not take into account the high degree of regularity generally exhibited by motion data. One way to exploit the regularity in spatiotemporal data is to represent the 2D formation or shape at each time instance as a linear combination of a small number of shape basis vectors $\mathbf{b}_j$ weighted by coefficients ω_j^i as $\mathbf{s}^i = \sum_j \omega_j^i \mathbf{b}_j^T$ [11, 8]. An alternative representation of the time-varying structure is to model it in the trajectory subspace, as a linear combination of trajectory basis vectors, θ_i as $\mathbf{s}_j = \sum_i a_i^j \theta_i$, where a_i^j is the coefficient weighting each trajectory basis vector [31, 2]. As a result, the structure matrix can be represented as either

$$\mathbf{S} = \mathbf{\Omega}\mathbf{B}^T \quad \text{or} \quad \mathbf{S} = \mathbf{\Theta}\mathbf{A}^T \tag{4}$$

	Compressibility	
Representation	Identity	Role
Single Team	30%	25%
Adversarial Teams	20%	**15%**

Table 2. Showing the compressibility of different representations. Compressibility in this context refers to the percentage of features required to represent 95% of the original signal.

where $\mathbf{B}$ is a $P \times K_s$ matrix containing K_s shape basis vectors, each representing a 2D structure of length $2P$, and $\mathbf{\Omega}$, is an $F \times K_s$ matrix containing the corresponding shape coefficients ω_j^i; and $\mathbf{\Theta}$ is an $F \times K_t$ matrix containing K_t trajectory basis as its columns, and $\mathbf{A}$ is a $2P \times K_t$ matrix of trajectory coefficients. The number of shape basis vectors used to represent a particular instance of motion data is $K_s \leq \min\{F, 2P\}$, and $K_t \leq \{F, 2P\}$ is the number of trajectory basis vectors spanning the trajectory subspace.

Both representations of $\mathbf{S}$ are over parameterizations because they do not capitalize on either the spatial or temporal regularity. As $\mathbf{S}$ can be expressed exactly as $\mathbf{S} = \mathbf{\Omega}\mathbf{B}^T$ and also $\mathbf{S} = \Theta\mathbf{A}^T$, then there exists a factorization

$$\mathbf{S} = \mathbf{\Theta}\mathbf{C}\mathbf{B}^T \tag{5}$$

where $\mathbf{C} = \mathbf{\Theta}^T\mathbf{\Omega} = \mathbf{A}^T\mathbf{B}$ is a $K_t \times K_s$ matrix of spatiotemporal coefficients. This equation describes the bilinear spatiotemporal basis, which contains both shape and trajectory bases linked together by a common set of coefficients.

Due to the high degree of temporal smoothness in the motion of humans, a predefined analytical trajectory basis can be used without significant loss in representation. A particularly suitable choice of a conditioning trajectory basis is the Discrete Cosine Transform (DCT) basis, which has been found to be close to the optimal Principal Component Analysis (PCA) basis if the data is generated from a stationary first-order Markov process [25]. Given the high temporal regularity present in almost all human motion, it has been found that the DCT is an excellent basis for trajectories of faces [3, 4] and bodies [6]. Figure 5 shows that due to the highly structured nature of the game, and the fact that humans motion is over short periods of time is very simple, we can gain enormous dimensionality reduction especially in the temporal domain. From this, we can effectively represent 5 second plays plays with no more than $K_t = 3$ and $K_s = 33$ with an maximum error of less than 2 meters. In terms of dimensionality reduction, this means we can represent temporal signals using $3 \times 33 = 99$ coefficients. For 5 second plays, this means a reduction of over 60 times. We found greater compressibility could be achieved on longer plays.

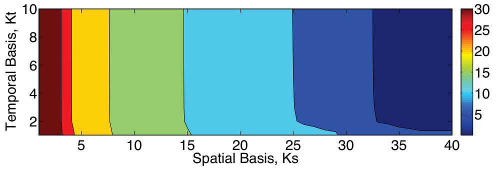

Figure 5. Plot showing the mean reconstruction error of the test data as the number of temporal basis (K_t) and spatial basis (K_s) vary for 5 second plays (*i.e.* $K_{t\max} = 150$). We magnified the plot to only show the first 10 temporal basis to highlight that only only $K_t = 3$ is required to represent coarse player motion.

4. The Assignment Problem

In the previous section, roles were specified by a human expert. We now address the problem of automatically assigning roles to an arbitrary ordering of player locations $\mathbf{p}_t^\tau$. Assuming a suitably similar vector $\hat{\mathbf{r}}^\tau$ of player locations in role order exists, we define the optimal assignment of roles as the permutation matrix $\mathtt{x}_t^{\tau\star}$ which minimizes the square L_2 reconstruction error

$$\mathtt{x}_t^{\tau\star} = \arg\min_{\mathtt{x}_t^\tau} \|\hat{\mathbf{r}}^\tau - \mathtt{x}_t^\tau \mathbf{p}_t^\tau\|_2^2. \qquad (6)$$

This is the linear assignment problem where an entry $\mathtt{C}(i,j)$ in the cost matrix is the Euclidean distance between role locations

$$\mathtt{C}(i,j) = \|\hat{\mathbf{r}}^\tau(i) - \mathbf{p}_t^\tau(j)\|_2. \qquad (7)$$

The optimal permutation matrix can be found in polynomial time using the Hungarian (or Kuhn-Munkres) algorithm [18].

4.1. Assignment Initialization

To solve the assignment problem, we need a reference formation to compare to. Using the mean formation (see Figure 4) is a reasonable initialization as the team should maintain that basic formation in most circumstances. However, in different areas of the field there are subtle changes in formation due to the what the opposition are doing as well as the game-state. To incorporate these semantics, we used a codebook of formations which consists of every formation within our training set. However, this mapping is difficult to do as the input features have no assignment. Given we have the assignment labels of the training data, we can learn a mapping matrix $\mathbf{W}$ from the mean and covariances of the training data to its assignment labels via the linear transform $\mathbf{X} = \mathbf{W}^T\mathbf{Z}$. Given we have N training examples, we can learn $\mathbf{W}$ by concatenating the mean and covariance into an input vector $\mathbf{z}_n$, which corresponds to the labeled formation $\mathbf{x}_n$. We compile all these features into the matrices $\mathbf{X}$ and $\mathbf{Z}$, and given these, we can use linear regression to learn $\mathbf{W}$ by solving

$$\mathbf{W} = \mathbf{X}\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T + \lambda\mathbf{I})^{-1} \qquad (8)$$

where λ is the regularization term. Using this approach, we can estimate a labelled formation from the training set which best describes the current unlabeled one. In terms of assignment performance on the test set, this approach works very well compared to using the mean formation for both the identity and role labels as can be seen in Table 3. Figure 6 shows the confusion matrices for both Team A and Team B for both representations. It worth noting that the role representation gave far better results than the identity representation, which is not surprising seeing that only spatial location is used. In terms of the role representation (bottom two plots), it can be seen that there is little confusion between the 3 defenders (LB, CH, RB) and the 3 forwards (LW, CF, RW). However, the midfield 4 (LH, RH, IL, IR) tend to interchange position a lot causing high confusion. Noticeably, there is a discrepancy between Team A and Team B which is understandable in this case as Team B interchanges positions more than twice the amount than Team A upon analysis of the ground-truth.

	Protoype	Hit Rate	
		Team A	Team B
Identity	Mean Formation	38.36	29.74
	Codebook	49.10	37.15
Role	Mean Formation	49.47	50.30
	Codebook	**74.18**	**69.70**

Table 3. Accuracy of the assignment using a mean formation as well as a codebook of possible formations.

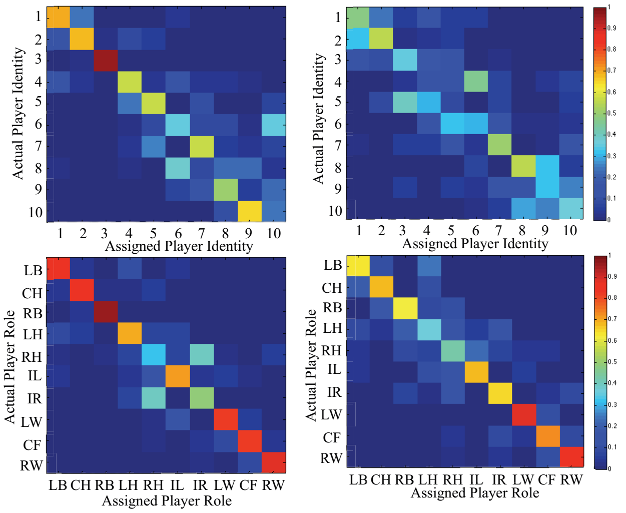

Figure 6. Confusion matrix showing the hit-rates for correctly assigning identity (top row) and role (bottom) for Team1 (left) and Team 2 (right) on the test set.

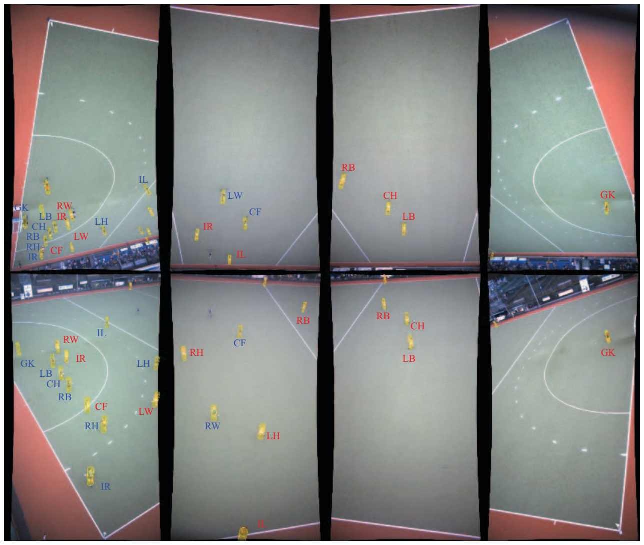

Figure 7. Players are detected by interpreting background subtraction results in terms of 3D geometry, where players are coarsely modeled as cylinders. Based on these detections, we assign player role for each team.

	Raw Detections		With Assignment	
	Precision	Recall	Precision	Recall
Detections	77.49	89.86	**91.90**	80.46
Team A	72.54	86.14	**86.69**	74.17
Team B	79.84	89.66	**92.91**	82.85

Table 4. Precision-Recall rates for the raw detections (left) and with the initialized assignments (right).

5. Interpreting Noisy Data

In practice, we will not obtain perfect data from a vision-system so our method has to be robust against both missed and false detections. To evaluate our approach, we employed a real-time state-of-the-art player detector [10] that detects player positions at 30fps by interpreting background subtraction results based on the coarse 3D geometry of a person (Figure 7). Once the locations of all players were determined, we classified the players into their respective teams using a color model for each team. Each player image was represented as a histogram in LAB color space and K-means clustering using the Bhattacharyya distance was performed to learn a generalized model for each team and camera. The precision and recall rates for the detector and the team affiliation are given in the left side of Table 4. In this work, we consider a detection to be made if a player was was within two meters of a ground-truth label.

5.1. Assigning Noisy Detections

To determine whether or not we should make the assignment or discard the detection, some type of game context feature is required (*i.e.* the part of the field most of the players are located). To do this, we employed a similar strategy to the one we proposed in Section 4.1. However, instead of learning the mapping from the clean features $\mathbf{Z}$, we learn from the noisy features $\mathbf{Z}_{\text{noisy}}$. As the player detector has systematic errors (there are some "black-spots" on the field due to reduced camera coverage, or game situations where players bunch together), we include the number of players detected from the system as well as the mean and covariance in our noisy game context feature $\mathbf{z}_{\text{noisy}}$, which we can then use to learn $\mathbf{W}_{\text{noisy}}$. We are able to do this as we make the assumption that the clean centroid is good approximation to the noisy centroid which was found was a valid one as can be seen in Figure 8. Using this assumption, we can obtain a reasonable prototypical formation to make our player assignments.

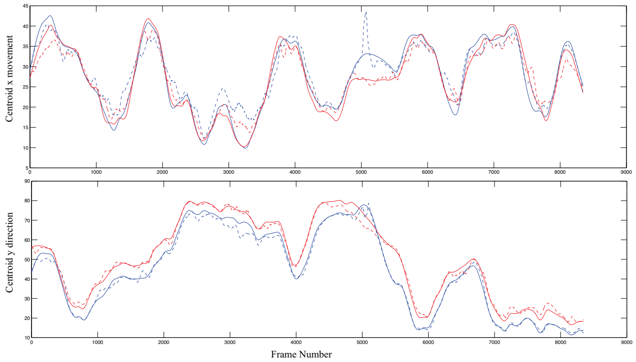

Figure 8. As the centroids of both the clean (solid) and noisy (dashed) of both teams (blue=team1, red=team2) are roughly equivalent, we learn a mapping matrix using linear regression to find a formation from the training set which can best describe the noisy test formation.

Using the estimated prototype, we then make an assignment by using the Hungarian algorithm. This is challenging however, as we may have missed or false detections which alters the one-to-one mapping between the prototype and input detections. To counter this, we employed an "exhaustive" approach, where if we have fewer detections than the number of players in the prototype, we find all the possible combinations that the labels could be assigned then use the combination which yielded the lowest cost from the assignments made. Conversely, if we had more detections than the number of players, we find all the possible combinations that the detections could be and then use the combination of detections which had the lowest cost.

For example, given we have only 9 detections for a team, we first find the 10 possible combinations that prototype could be (i.e. $[1, \ldots, 9]$, $[2, \ldots, 10]$, $[1, 2, 4, \ldots, 10]$,

	Correct	Incorrect	Missed	Hit Rate
Team A	41.89	32.89	25.22	56.02
Team B	45.92	35.56	18.53	56.36

Table 5. Detection rates assigning roles to the noisy data. The column on the far right gives the effective hit-rate (i.e. missed detections omitted) of the correct assignments.

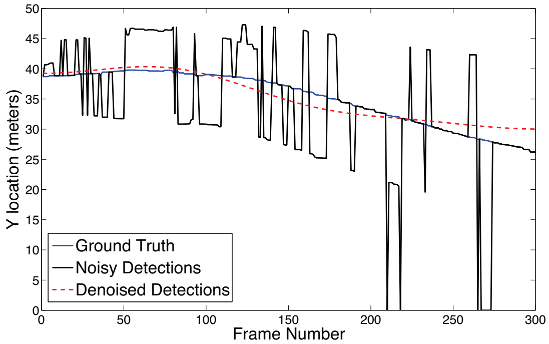

Figure 9. Given our noisy detections (black), using our bilinear model we can estimate the trajectory of each player over time. We can see our estimate (red) is close to the ground-truth (blue).

$[1, 2, 3, 5, \ldots, 10]$ etc.). For each one of these combinations, we then perform the Hungarian algorithm and calculate the cost of the made assignments. After we have exhaustively gone through all possible combinations, we make the assignment based on the combination with the lowest cost. Or given we have 11 detections for a team, we first find the 11 possible combinations that the detections could be, find the cost for each set and choose the one with the lowest cost. However, sometimes we get false positives which means that even though we may get 10 detections for a team we may only have 7 or 8 valid candidates. Employing this approach greatly improves the precision rate, while the recall rate decreases which is to be expected (see right side of Table 4). Even despite the drop in recall, we still assign role reasonably well (over 55% compared to 66% on the clean data) as can be seen in Table 5.

5.2. Denoising the Detections

While our precision and recall rates from the detector are relatively high, to do useful analysis we need a continuous estimate of the player label at each time step to do formation and play analysis. This means that we need a method which denoise the signal - that is a method which can impute missing data and filter out false detections. Given the spatial bases, the bilinear coefficients and an initial estimate of the player labels, we can use an Expectation Maximization (EM) algorithm to denoise the detections. The approach we use is similar to [4]. Using this approach, the expectation step is simplified to making an initial hard assignment of the labels which can be gained by finding the initial assignments using the method described in the previous section. From this initialization, we have an initial guess of $\hat{\mathbf{S}}$. In the maximization step, we can calculate $\mathbf{C} = \Theta^T \hat{\mathbf{S}} \mathbf{B}$, and then estimate $\mathbf{S}$ from our new $\mathbf{C}$ as well as our spatial and temporal basis $\mathbf{B}$ and Θ. Examples of the cleaned up detections using this approach are shown in Figure 9.

As the recall rate of the denoised data is 100%, we are interested to see how precise our method is in inferring player position based on their label. To test this, we calculated the

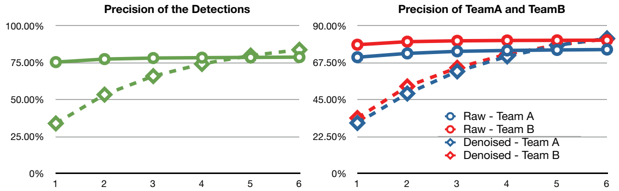

Figure 10. Precision accuracy vs the distance threshold from ground-truth for: (left) the overall detections, (right) the detections based on team affiliation. The solid lines refer to the raw detections and the dashed lines refer to the denoised signal.

precision rate for the detections and the denoised detections against a distance threshold - that is, the minimum distance a player had to be to ground-truth to be recognized as a correct detection). The results are shown in Figure 10. As can be seen from these figures, the detections from the player detector are very accurate and do not vary with respect to the error threshold (i.e. it either detects a player very precisely or not at all). Conversely, the denoised data is heavily smoothed due to the bilinear model, so we lose some of the finer detail to gain a continuous signal.

5.3. Formation and Play Analysis

To check the usefulness of our cleaned-up signal, we conducted cluster analysis on both static formations and dynamic plays to see whether we could replicate what could achieve with manually labelled data. The first analysis we conducted was to find the top five formations that could best describe the test data (i.e. the 3 most likely formations that occurred). The results are shown in Figure 11. From the figure it can be seen that despite small differences, we go close to replicating what we get from manually labelled data – formations 1 correspond and 3 and 2 are reversed. A similar trend is observed for the play analysis where we clustered 10 second plays (see Figure 12) . As can be seen from the denoised data, the bilinear model has smoothed out the trajectory, although it is unrealistic in some cases. Additionally, this analysis can be done with a fraction of the amount of features due to the high compressibility of the signal ($D = 200$ vs $D = 12000$).

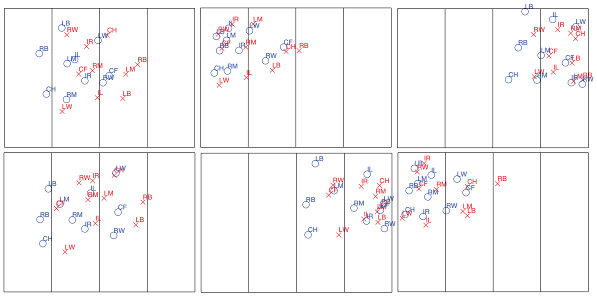

Figure 11. Cluster analysis of the top three formations (1-3, ordered left-to-right) which best represent the test data using (top) manually labelled data and (bottom) our denoised data. The blue team is attacking from left-to-right.

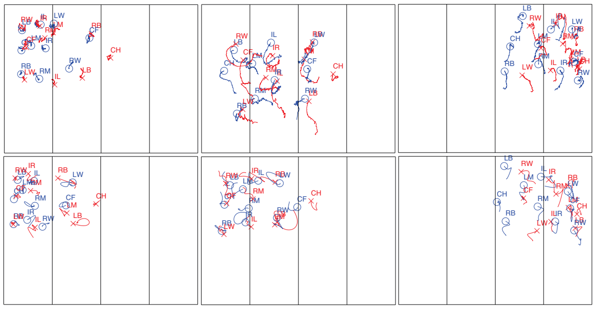

Figure 12. Cluster analysis of the top three 10 second plays on the test data using (top) manually labelled data and (bottom) our denoised data. The x's and the o's refer the position of the player at the end of the 10 second play.

6. Summary and Future Work

In this paper, we presented a representation which utilized player role labels to exploit the heavy spatiotemporal correlations that exist within adversarial domains. As this representation is highly correlated in both space and time, we showed that a spatiotemporal bilinear basis model can leverage this trait to compress the incoming signal by up to two orders of magnitude without much loss of information. Our final contribution of this paper was the use of the bilinear model to effectively clean up noisy player detections from a state-of-the-art detector, which enables analysis of static formations as well as temporal plays. To enable this, we used the Hungarian algorithm in an exhaustive way based on a prototype formation which was found using a codebook of possible formations. The implications of this work are important, as having the ability to identify formations and plays from a large repository can enhance realtime commentary in sports by helping highlight recurrent team strategies and long-term trends. The process of post-game play annotations , which coaches and their teams spend hours performing manually could be automated. Our future work will be focussing on large quantities of data to enable this to occur.

References

[1] J. Aggarwal and M. Ryoo. Human Activity Analysis: A Review. *ACM Computing Surveys*, 2011.

[2] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Nonrigid Structure from Motion in Trajectory Space. In *NIPS*, 2008.

[3] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Trajectory Space: A Dual Representation for Nonrigid Structure from Motion. *T. PAMI*, 2010.

[4] I. Akhter, T. Simon, S. Khan, I. Matthews, and Y. Sheikh. Bilinear Spatiotemporal Basis Models. *ACM Transactions on Graphics*, 2012.

[5] S. Ali and M. Shah. Floor Fields for Tracking in High Density Crowd Scenes. In *ECCV*, 2008.

[6] O. Arikan. Compression of Motion Capture Databases. *ACM Transactions on Graphics*, 25(3), 2006.

[7] M. Beetz, N. von Hoyningen-Huene, B. Kirchlechner, S. Gedikli, F. Siles, M. Durus, and M. Lames. ASPOGAMO: Automated Sports Game Analysis Models. *International Journal of Computer Science in Sport*, 8(1), 2009.

[8] C. Bregler, A. Hertzmann, and H. Biermann. Recovering Non-Rigid 3D Shape from Image Streams. In *CVPR*, 2000.

[9] A. Bronstein, M. Bronstein, and R. Kimmel. *Numerical Geometry of Non-Rigid Shapes*. Springer, 2008.

[10] P. Carr, Y. Sheikh, and I. Matthews. Monocular Object Detection using 3D Geometric Primitives. In *ECCV*, 2012.

[11] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models - their training and applications. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[12] A. Gupta, P. Srinivasan, J. Shi, and L. Davis. Understanding Videos, Constructing Plots: Learning a Visually Grounded Storyline Model from Annotated Videos. In *CVPR*, 2009.

[13] A. Hervieu and P. Bouthemy. Understanding sports video using players trajectories. In J. Zhang, L. Shao, L. Zhang, and G. Jones, editors, *Intelligent Video Event Analysis and Understanding*. Springer Berlin / Heidelberg, 2010.

[14] S. Intille and A. Bobick. A Framework for Recognizing Multi-Agent Action from Visual Evidence. In *AAAI*, 1999.

[15] S. Khan and M. Shah. Detecting Group Activities Using Rigidity of Formation. In *ACM Multimedia*, 2005.

[16] K. Kim, M. Grundmann, A. Shamir, I. Matthews, J. Hodgins, and I. Essa. Motion Fields to Predict Play Evolution in Dynamic Sports Scenes. In *CVPR*, 2010.

[17] K. Kitani, B. Ziebart, A. Bagnell, and M. Herbert. Activity Forecasting. In *ECCV*, 2012.

[18] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[19] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfield. Learning Realistic Human Actions from Movies. In *CVPR*, 2008.

[20] R. Li and R. Chellappa. Group Motion Segmentation Using a Spatio-Temporal Driving Force Model. In *CVPR*, 2010.

[21] R. Li, R. Chellappa, and S. Zhou. Learning Multi-Modal Densities on Discriminative Temporal Interaction Manifold for Group Activity Recognition. In *CVPR*, 2009.

[22] V. Morariu and L. Davis. Multi-Agent Event Recognition in Structured Scenarios. In *CVPR*, 2011.

[23] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. You'll Never Walk Alone: Modeling Social Behavior for Multi-Target Tracking. In *CVPR*, 2009.

[24] M. Perse, M. Kristan, S. Kovacic, and J. Pers. A Trajectory-Based Analysis of Coordinated Team Activity in Basketball Game. *Computer Vision and Image Understanding*, 2008.

[25] K. Rao and P. Yip. *Discrete Cosine Transform: Algorithms, Advantages, Applications*. Academic, New York, NY, 1990.

[26] M. Rodriguez, J. Sivic, I. Laptev, and J. Audibert. Data-Driven Crowd Analysis in Video. In *ICCV*, 2011.

[27] A. Sadilek and H. Kautz. Recognizing Multi-Agent Activities from GPS Data. In *AAAI*, 2008.

[28] B. Siddiquie, Y. Yacoob, and L. Davis. Recognizing Plays in American Football Videos. Technical report, University of Maryland, 2009.

[29] D. Stracuzzi, A. Fern, K. Ali, R. Hess, J. Pinto, N. Li, T. Konik, and D. Shapiro. An Application of Transfer to American Football: From Observation of Raw Video to Control in a Simulated Environment. *AI Magazine*, 32(2), 2011.

[30] G. Sukthankar and K. Sycara. Activity Recognition for Dynamic Multi-Agent Teams. *ACM Trans. Intell. Syst. Technol*, 2012.

[31] L. Torresani and C. Bregler. Space-Time Tracking. In *CVPR*, 2002.

[32] D. Tran and J. Yuan. Optimal Spatio-Temporal Path Discovery for Video Event Detection. In *CVPR*, 2011.

[33] C. Xu, Y. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang. Using Webcast Text for Semantic Event Detection in Broadcast. *T. Multimedia*, 10(7), 2008.

[34] Y. Zhang, W. Ge, M. Chang, and X. Liu. Group Context Learning for Event Recognition. In *WACV*, 2012.