Identifying Team Style in Soccer using Formations from Spatiotemporal Tracking Data

Alina Bialkowski[1,2], Patrick Lucey[1], Peter Carr[1], Yisong Yue[1,3], Sridha Sridharan[2] and Iain Matthews[1]
[1]Disney Research, Pittsburgh, USA, [2]Queensland University of Technology, Australia, [3]California Institute of Technology, USA
Email: a.bialkowski@connect.qut.edu.au, {patrick.lucey, peter.carr, iainm}@disneyresearch.com
yyue@caltech.edu, s.sridharan@qut.edu.au

***Abstract*—To the trained-eye, experts can often identify a team based on their unique style of play due to their movement, passing and interactions. In this paper, we present a method which can accurately determine the identity of a team from spatiotemporal player tracking data. We do this by utilizing a formation descriptor which is found by minimizing the entropy of role-specific occupancy maps. We show how our approach is significantly better at identifying different teams compared to standard measures (i.e., shots, passes etc.). We show the utility of our approach using a entire season of Prozone player tracking data from a top-tier professional soccer league.**

I. Introduction

The question we ask in this paper is: *given all the player and ball tracking data of a team in a season, what team-based features can adequately discriminate a team's behavior?* In practice, an expert human is able to do this but this is very labor intensive and is inherently subjective. Having a method which can quantify these behaviors should be possible with the prevalence of spatiotemporal tracking data of player and ball movement being captured in most professional sports (e.g., [1], [2]). However, this task is challenging due to the complexities dealing with adversarial multi-agent trajectory data. A major issue centers on the *alignment* of individual player trajectories within a team setting which is a source of noise. In this paper, we align the data based on a role-based method which is learnt directly from data [3]. We show that using this approach, semantically meaningful team-based strategic features can be obtained which are highly predictive of their identity. We compare this descriptor to other features such as match statistics (e.g., shots, passes, fouls) and ball movement and show that the formation descriptor is far more superior in discriminating unique team characteristics (Fig. 1).

A. Related Work

With the recent deployment of player tracking systems in professional sports, a recent influx of research has been conducted on how to use such data sources. Most of the work has centered on individual player analysis. In basketball, Goldsberry [4] used player tracking data to rank the best shooters in the NBA according to their shot location. Maheswaran et al. [5], [6] used the tracking data to analyze the best method to obtain a rebound. Similarly, Wiens et al. [7] looked at how teams should crash the backboard to get rebounds. Recently, Lucey et al. [8] used tracking data to discover how teams achieved open three-point shots. Bocskocksy et al. [9] re-investigated the hot-hand theory. Miller et al., [10] analyzed the shot selection process of players using non-negative matrix factorization. Cervone et al. [11] used basketball tracking data to predict points and decisions made during a play. Carr et al. [12] used real-time player detection data to predict the future location of play and point a robotic camera in that location for automatic sport broadcasting purposes. In tennis, Wei et al. [13], [14] used Hawk-Eye data to predict the type and location of the next shot. Ganeshapillai and Guttag [15] used SVMs to predict pitching in baseball while Sinha et al. [16] used Twitter feeds to predict NFL outcomes.

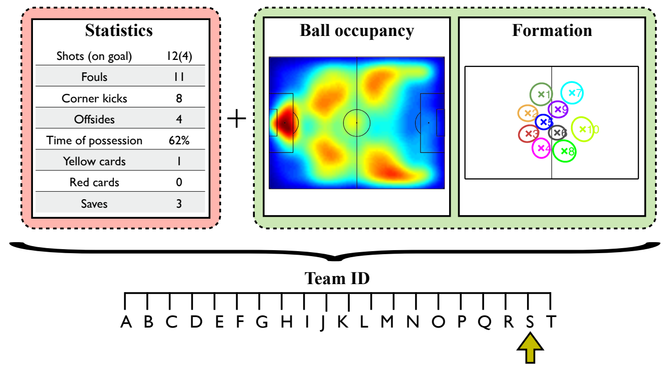
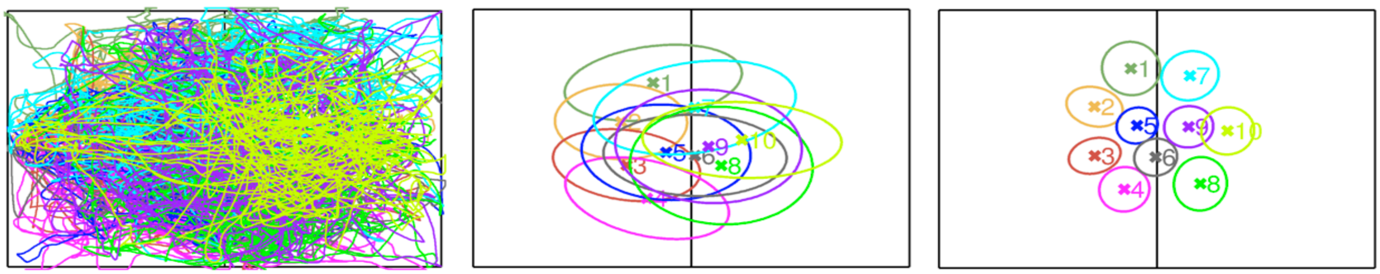


Fig. 1. In this paper, based solely on (left) match statistics, (middle) ball movement patterns, and (right) formation descriptor - we can predict with high accuracy the identity of a team soccer. We show the formation descriptor is the best discriminator of team style.

In terms of analyzing a team's style of play, most work has centered on soccer. Lucey et al. [17] used entropy maps to characterize a team's ball movement patterns using data from Opta [18]. This was followed by [19], which showed that a team's home and away style varied, highlighting the home teams had more possession in the forward third as well as shots and goals. Bialkowski et al. [20] examined the rigidity of a team's formation across a season and showed that home teams tended to player higher up the pitch both in offense and defense. Outside of the sporting realm, there has been plenty of work focusing on identifying style. In the seminal work on separating style from content, Tenenbaum and Freeman [21] used a bilinear model to decouple the raw content for improved recognition on a host of different tasks. More recently, Doersch et al. [22] used discriminative clustering to discover the attributes that distinguished images of one city from another. They followed this work by exploring the visual style of objects (e.g., cars and houses) and how they vary over time [23]. The contribution of this paper is using a *formation* descriptor to identity the unique style of a team.



(a) (b) (c)

Fig. 2. (a) Given the player trajectory of each player during an entire half, we see that players continually swap positions. (b) Shown are the mean-normalized covariances of player positions which again highlights the overlap. (c) Using our iterative approach (which is very similar to k-means with the constraint that at every frame each detection requires a unique role), a role label is assigned to each player at the frame-level, we see the underlying structure of the team.

Statistic	Frequency
Teams	20
Games	375
Data Points	3.89M
Ball Events	721K

TABLE I. INVENTORY OF DATASET USED FOR THIS WORK.

II. DATA: PLAYER TRACKING IN SOCCER

For this work, we utilized an entire season of player tracking data from Prozone. The data consisted of 20 teams who played home and away, totaling 38 games for each team or 380 games overall. Five of these games were omitted for various reasons. We refer to the 20 teams using arbitrary labels $\{A, B, \ldots, T\}$. Each game consists of two halves, with each half containing the (x, y) position of every player at 10 frames-per-second. This results in over 1 million data-points per game, in addition to the 43 possible annotated ball events (e.g., passes, shots, crosses, tackles etc.). Each of these ball events contained the time-stamp as well as location and players involved. An inventory of the data is given in Table I.

III. DISCOVERING FORMATIONS FROM DATA

In sports, there exists a well established vocabulary for describing the responsibility each player has within a team. Even though it varies from sport to sport, within each sport these descriptions generalize. The language used is in terms of *formations*, which is effectively a strategic concept (i.e., different teams can use the same formation simultaneously). As a result, we refer to a formation's generic players using a set of identity agnostic labels which we denote *roles*. A formation is generally shift-invariant and allows for non-rigid deformations. Therefore, we define each *role* by its position relative to the other roles (i.e., insoccer a left-midfielder plays in-front of the left-back and to the left of the center-midfielder). Each role within a formation is unique (i.e., no two players within the same formation can have the same role at the same time), and players can swap roles throughout the match. Additionally, multiple formations may exist which can be interpreted as different sets of roles. A role represents any arbitrary 2D probability density function. Therefore, we can represent it non-parametrically by quantizing the field into a discrete number of cells, or parametrically using a mixture of 2D Gaussians. We can then represent the formation by concatenating the features of each role into a single vector.

Pass	Foul - Direct FK	Cross	Catch Drop Save
Pass Assist	Foul - Indirect FK	Cross Assist	Catch Save
Corners	Foul - Penalty	Reception	Punch
Shot on Target	Foul - Throw-in	Reception Assist	Punch Save
Shot off Target	Offside	Reception Save	Diving
Goal	Yellow Card	Catch	Diving Save
Own Goal	Red Card	Catch Drop	Drop of Ball
Neutral Clear Save	Running with Ball	Chance	Substitution
Block	Drop Kick	Pass Save	Hold of Ball
Clearance Uncontrolled	Neutral Clearance	Player Out	Clearance

TABLE II. LIST OF MATCH STATISTICS USED TO DESCRIBE TEAM BEHAVIOR.

Role is a dynamic label, meaning that a player can be fulfill many roles during the game (e.g., left-winger switches to the right-wing and is not characterize as the left-winger because he/she started there). However, each role needs to be assigned to a player in every frame (i.e., two players can not be in the same role at the same time).

As a formation basically assigns an area or space to each player at every frame, this problem can be framed as a *minimum entropy data partitioning problem* [24], [25]. Bialkowski et al., [3] show the full derivation, but in practice it is similar to k-means clustering with the caveat of instead of assigning each data point to its closest cluster, we solve a linear assignment problem between identities and roles using the Hungarian algorithm [26]. The process is shown in Fig 2. The formation of every team in every half we analyzed in shown in Fig 3. We compare the formation descriptor to other match factors in the next section.

IV. PREDICTING TEAM IDENTITY

To determine if teams had a distinct playing style, we conducted a series of team identity experiments. The challenge was, *given only player tracking data and ball events, can we predict the identity of each team?* To do this, we need descriptors of team behaviors during a match. For this paper, we generated three types of match descriptors: 1) match statistics, 2) ball occupancy, and 3) team formation.

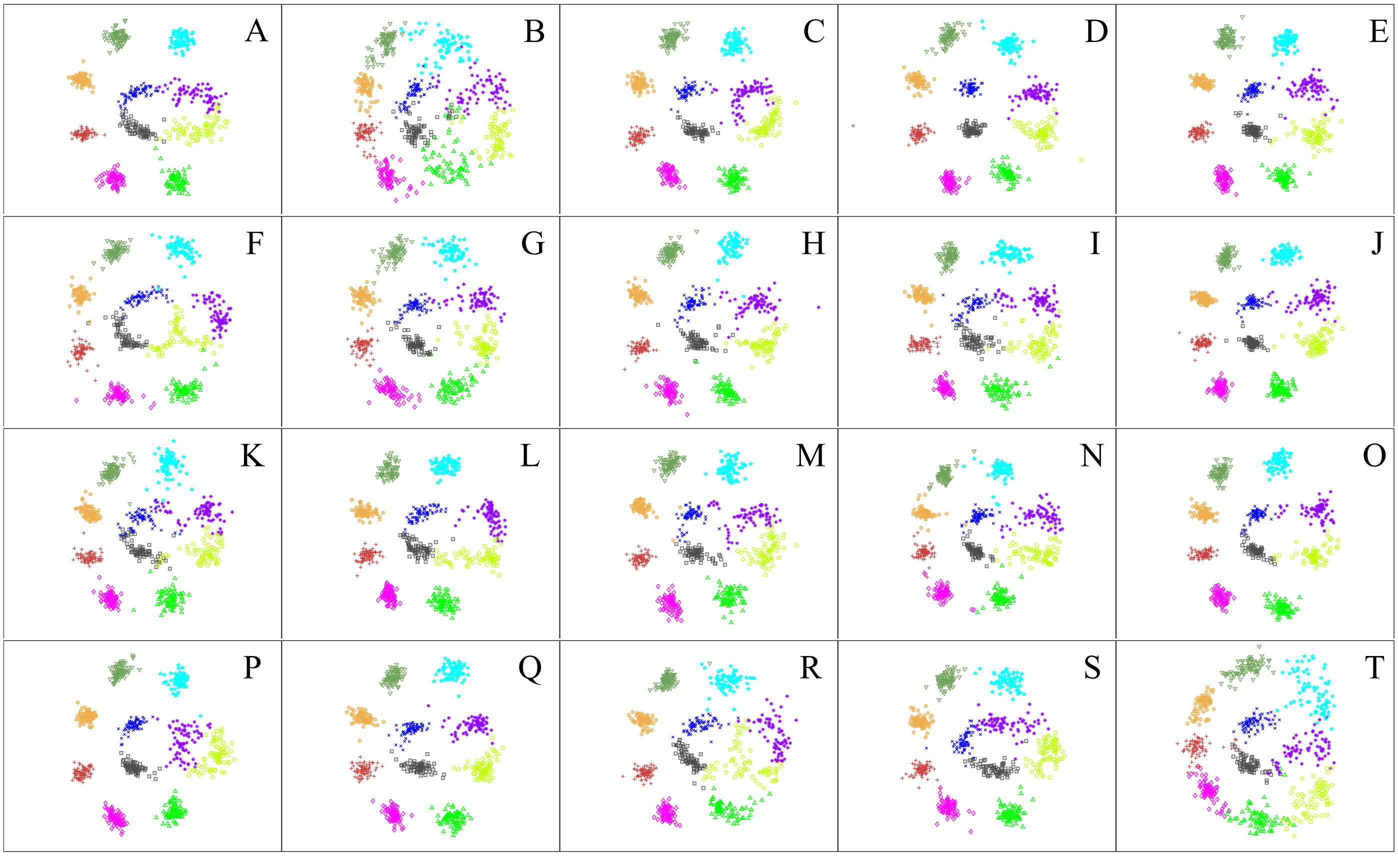


Fig. 3. Example of our formation descriptors for each team. The colors represent different roles. For visualization purposes we have just plotted the centroid for each role for each match.

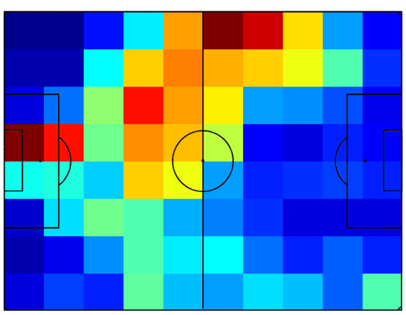


Fig. 4. Example of ball occupancy map of a team from a match (attacking left to right).

A. Match Descriptors

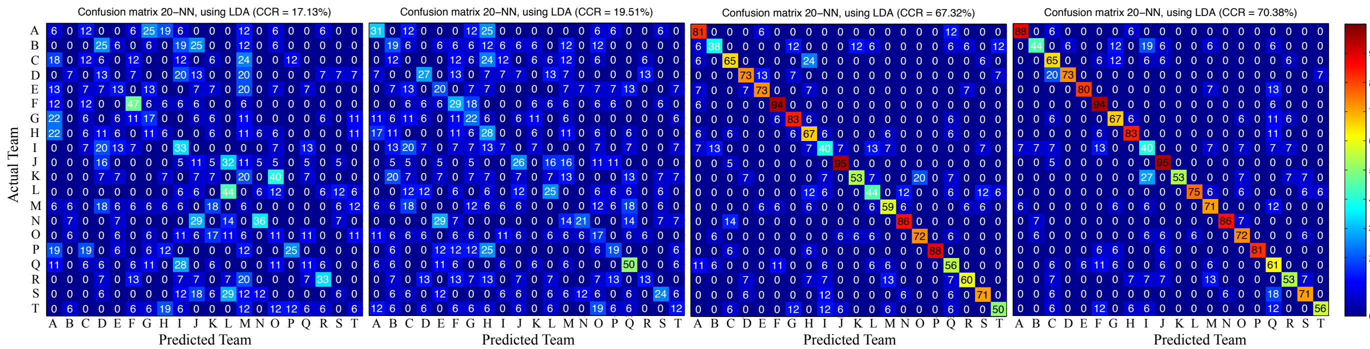
Match Statistics: During a match, various statistics that capture team and individual behavior are annotated. Table II shows the list of statistics which we used in this paper. While the number of these match statistic are quite large, the majority of them are quite sparse with only a couple of these events labelled per match. In reporting in the match, only a half-dozen of the most important match statistics are normally documented (i.e., goals, shots on target, shots off target, passes, corners, yellow and red-cards).

Ball Occupancy: Associated with the match statistics are the time and location that the event occurred. To form a representation of this information, we adopted the approached used in [17], [19] which consists of estimating the continuous ball trajectory at each time-stamp as well as which team had possession (we ignore stoppages). We then broke the field into a 10×8 spatial grid and calculated the ball occupancy of each of these grids for each team (i.e. how often the team was in possession of the ball in this location over the match). A visualization of the resulting occupancy is shown in Figure 4.

Formation Descriptor: For each match half, we found the formation descriptor $\mathcal{F}^*$ by using the method described in Section III. This gave a $M \times N$ matrix where M refers to the number of cells in the field and N is the number of roles which was 10 (we omitted the goal-keeper as well as games which had a player sent off). A depiction of the formation descriptors for each team for all matches are shown in Figure 3. For clarity of presentation, we have only plotted the centroid of each role for each match with each team attacking from left-to-right. Each different color marker corresponds to a different role for that team. It can be seen from the plot that teams are rather rigid in the way they play across a season which suggest that this is a useful feature in discriminating between different teams. Another interesting point, as teams vary little in terms of playing style throughout the season, this could be used as a powerful prior in opposition teams preparing for upcoming matches.

B. Experiments

The team identity experiments were performed using a "leave-one-match-out" cross-validation strategy where one match was left out to test against, and the remaining matches



(a) (b) (c) (d)

Fig. 5. Team identity results for the various descriptors: (a) match statistics, (b) ball occupancy, (c) formation descriptor and (d) fused all descriptors.

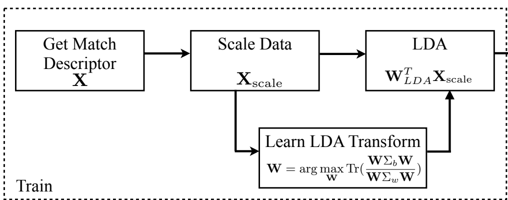


Fig. 6. Given a match descriptor, we first scale the data and then multiply it by $\mathbf{W}^T$ which is found using LDA to yield a discriminative feature vector. The LDA matrix is learnt using the team identity labels and their match descriptors in the training set. Team identity is predicted using k-NN.

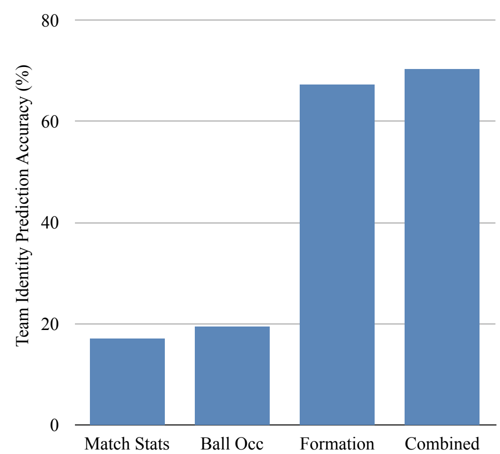


Fig. 7. Comparison of the accuracy of predicting team identity based on the different descriptors.

were used as our train set. The results are shown in Figure 5 and the block-diagram shown in Figure 6 describes the process. Firstly, we generate the descriptors described above and then scale the features. To obtain a compact but discriminative representation, we perform linear discriminant analysis (LDA) by learning the transformation matrix $\mathbf{W}$ from the training set where we used the team identity as the class labels (i.e., $C = 20$). We learn a $\mathbf{W}$ for each descriptor and then multiply the features by $\mathbf{W}$ to yield a $C-1$ feature vector. To predict the identity label of the teams in the test match, we use a k-nearest-neighbor classifier ($k = 20$) using the Euclidean norm as our distance metric.

The results for the various descriptors are shown in Figure 5. In the first experiment, (Figure 5(a)) we can see that using the match statistics are quite low with an overall accuracy of 17% (chance is 5%). This result makes sense as the match statistics only contain coarse event information without any spatial or temporal information about the ball or the players. Using the ball occupancy only gave marginally improved performance over the match statistics with an accuracy of 19% (Figure 5(b)). This is well below the 33% which was obtained in the previous works [17], [19]. A possible explanation of the performance difference could be due to the coarse estimation of the possession strings and the ball occupancy maps from the event data.

The most impressive performance by far is the formation descriptor which obtains over 67% accuracy, which clearly shows that teams have a true underlying signal which can be encapsulated in the way the team moves in formation over time (Figure 5(c)). We also fused together these descriptors by late-integration method where the number of k-nearest neighbors for each stream was dependent on how reliable the stream was (i.e., the formation descriptor received the most neighbors/votes). This approached improved the overall performance to over 70% which shows there is complimentary information within the other descriptors. A bar-graph comparing the overall performance for each descriptor is given in Figure 7.

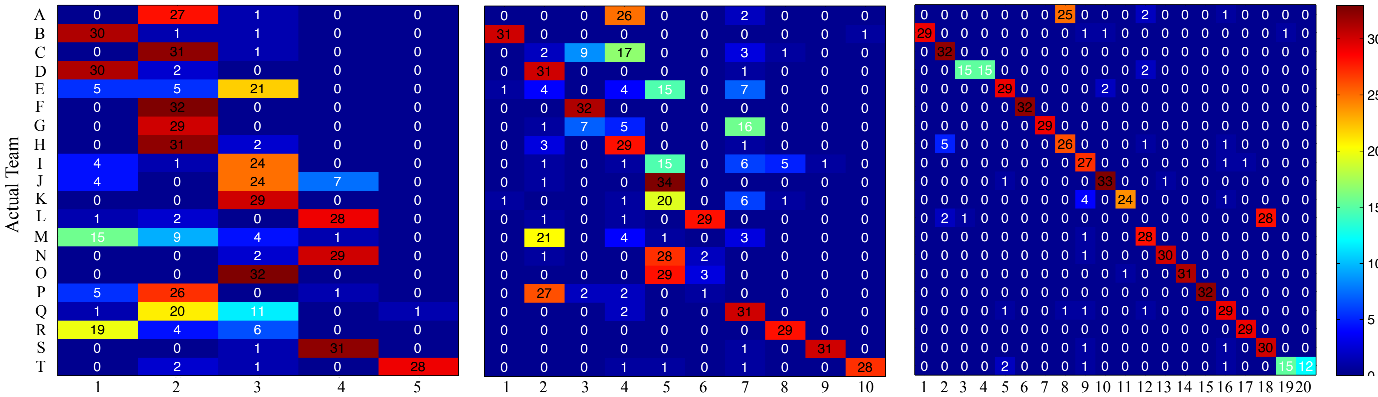
V. Analyzing Team Behaviors

In this section we explore how we can learn and represent the characteristic style of teams, and use this for analysing team behaviours in prediction and anomaly detection tasks.

A. Team Style

Team style is a very subjective and high-level attribute to label, especially in continuous sports like soccer. This is in part due to the dynamic and low-scoring nature of such sports, as it is hard to segment the game into discrete parts and assign a label when style encompasses all aspects of play. Due to the global nature of style, one way to quantify a team's style is via a linear combination of prior behaviour styles.

Given a training set of team behaviorr descriptors, we can discover a discrete set of styles using k-means clustering. For evaluation, we exclude the last two rounds of the season for testing, and use the remaining games to train the style models. We first project the match features into a lower dimensional, discriminative space using LDA, as in the team



(a) (b) (c)

Fig. 8. Shows the clustering results based on style when we set the number of styles to: (a) 5, (b) 10, and (c) 20. These can be used as a style prior for predicting the results of future matches.

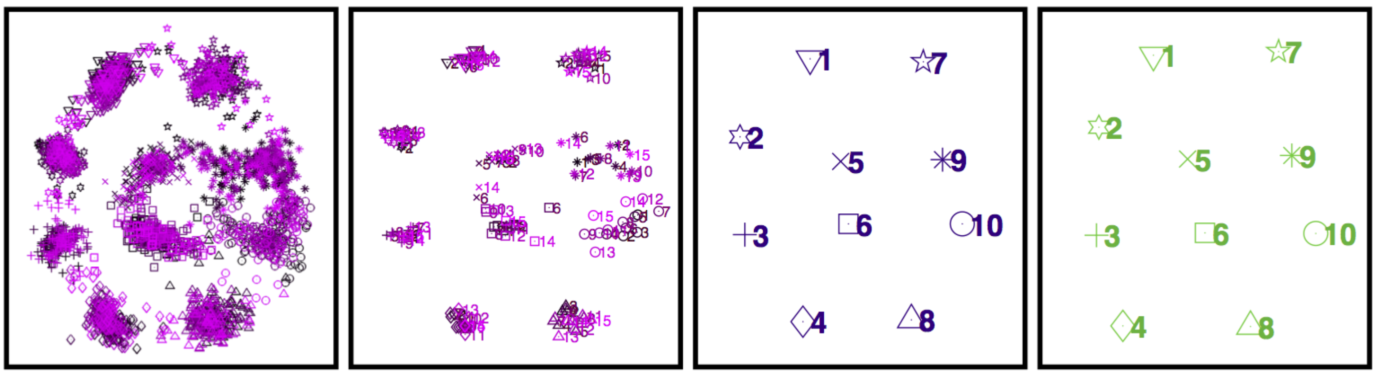


Fig. 10. Prediction of formation using k-NN regression. (a) all training examples, (b) retrieved examples according to style prior, (c) the predicted formation (= mean(retrieved examples)), (d) the actual formation.

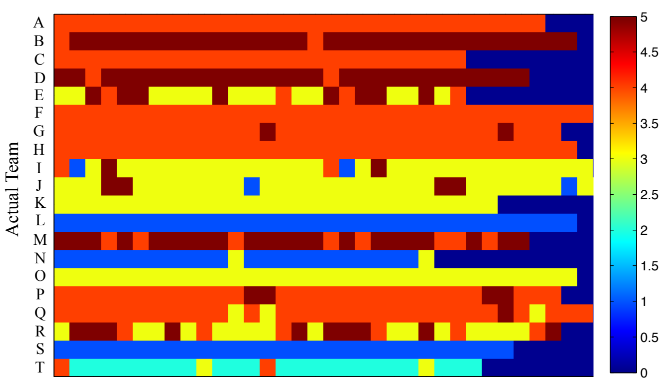


Fig. 9. Shows the variation in style each team has when we cluster to only 5 styles.

identity experiments (Fig. 6), and then cluster similar examples in this space. The style clustering results for k = 5, 10 and 20, are shown in Figure 8.

Observing Figure 8, there is some overlap in styles between certain teams, and some teams exhibit multiple styles. The variation in style for each team using k = 5 styles, is shown in Figure 9. Team T stands out, being in a style cluster of its own, which could be explained by the distinctly different formation from all other teams, with 3 defenders at the back (see Fig. 3). Most teams play a single style, while teams E and R vary their playing styles more frequently than other teams.

To encapsulate the behaviour styles that teams adopt, we define the playing style of a team as the normalised weights from the style clustering matrices (e.g. for the 5 style clusters used in Figure 8(a), the style vector for Team A=$[0, \frac{27}{28}, \frac{1}{28}, 0, 0]$, Team B=$[\frac{30}{32}, \frac{1}{32}, \frac{1}{32}, 0, 0]$, etc.). Modeling teams as a combination of the styles they play makes intuitive sense, as sometimes a team could play a pressing game and on other occasions the team may play defensively, so they would be weighted according to these performances. Another team may be very rigid and play the same style every game - so the weight for that game may be very high. These style vectors can then be used to assist prediction.

B. Prediction and Anomaly Detection

Previously, given the ball and player tracking data, we predicted the team identity. In this section, we want to do the reverse - *given we just have the identity of the two teams playing, can we predict how the game will be played by estimating what the match features will be?*

To predict the most likely features, we use K-NN regression using the learnt team style priors as the input, which allows us to select which of the training matches to regress from for our prediction. That is, for each match in the training set, we

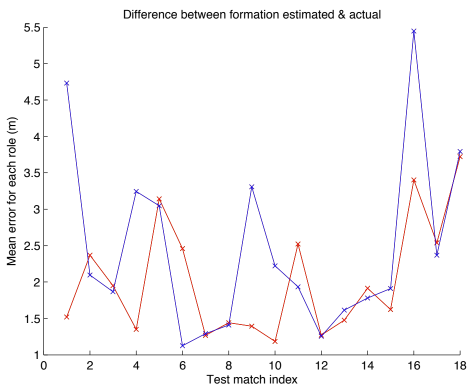


Fig. 11. Evaluation of formation prediction results (prediction vs actual, Red = home team, Blue = away team)

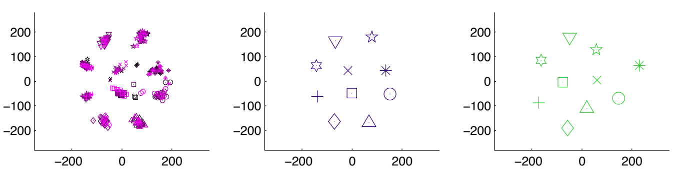


Fig. 12. Example of a poor formation estimate, which appears to be due to an anomaly in the team's behaviour. (a) retrieved examples, (b) predicted formation, (c) actual formation

compare the two team styles to the test match's team style priors. We then extract the matches which are most similar in terms of team styles, and calculate the mean features to predict the outcome of the test match. We can then compare this prediction with the actual result. The procedure, demonstrating formation prediction is shown in Figure 10.

We performed prediction of team formation on the last two rounds of the season (containing 18 matches). We evaluated the results by comparing the predicted formation to the actual formation played, presented in Figure 11. It can be seen that most matches are estimated within 2 m average error per role, while Match 1 and 16 are most poorly estimated. This suggests that the teams were not playing their normal formation style in these matches (i.e. anomalous behaviour). The predictions allow us to visualise the most likely formation given prior examples and when anomalies occur, such as in Figure 12.

VI. Summary and Future Work

In this paper, we first presented a *formation descriptor* which was found by minimizing the entropy of a set of player roles. Using an entire season of player tracking data, we generated the formation descriptor by projecting the set of occupancy maps of each role into a low-dimensional discriminative feature space using linear discriminating analysis (LDA). We showed that this approach characterizes individual team behavior significantly better (3 times more) than other match descriptors which are normally used to describe team behavior. We then conducted a series of analysis and predictions which showed the utility of our approach. In future work, we plan to use this descriptor for short-term prediction (i.e., who will the next pass go to etc.), as well as long-term prediction (i.e., match result).

References

[1] Prozone, www.prozonesports.com.

[2] STATS SportsVU, www.sportvu.com.

[3] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews, "Large-Scale Analysis of Soccer Matches using Spatiotemporal Tracking Data," in *ICDM*, 2014.

[4] K. Goldsberry, "CourtVision: New Visual and Spatial Analytics for the NBA," in *MITSSAC*, 2012.

[5] R. Masheswaran, Y. Chang, A. Henehan, and S. Danesis, "Destructing the Rebound with Optical Tracking Data," in *MIT Sloan Sports Analytics Conference*, 2012.

[6] R. Masheswaran, Y. Chang, J. Su, S. Kwok, T. Levy, A. Wexler, and N. Hollingsworth, "The Three Dimensions of Rebounding," in *MIT Sloan Sports Analytics Conference*, 2014.

[7] J. Wiens, G. Balakrishnan, J. Brooks, and J. Guttag, "To Crash or Not to Crash: A quantitative look at the relationship between the offensive rebounding and transition defense in the NBA," in *MITSSAC*, 2013.

[8] 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," in *MITSSAC*, 2014.

[9] A. Bocskocsky, J. Ezekowitz, and C. Stein, "The Hot Hand: A New Approach to an Old "Fallacy"," in *MIT Sloan Sports Analytics Conference*, 2014.

[10] A. Miller, L. Bornn, R. Adams, and K. Goldsberry, "Factorized Point Process Intensities: A Spatial Analysis of Professional Basketball," in *ICML*, 2014.

[11] D. Cervone, A. D'Amour, L. Bornn, and K. Goldsberry, "POINTWISE: Predicting Points and Valuing Decisions in Real Time with NBA Optical Tracking Data," in *MIT Sloan Sports Analytics Conference*, 2014.

[12] P. Carr, M. Mistry, and I. Matthews, "Hybrid Robotic/Virtual Pan-Tilt-Zoom Cameras for Autonomous Event Recording," in *ACM Multimedia*, 2013.

[13] X. Wei, P. Lucey, S. Morgan, and S. Sridharan, "Sweet-Spot: Using Spatiotemporal Data to Discover and Predict Shots in Tennis," in *MIT Sloan Sports Analytics Conference*, 2013.

[14] ——, "Predicting Shot Locations in Tennis using Spatiotemporal Data," in *DICTA*, 2013.

[15] G. Ganeshapillai and J. Guttag, "A Data-Driven Method for In-Game Decision Making in MLB," in *MIT Sloan Sports Analytics Conference*, 2014.

[16] S. Sinha, C. Dyer, K. Gimpel, and N. Smith, "Predicting the NFL Using Twitter," in *ECML Workshop on Machine Learning and Data Mining for Sports Analytics*, 2013.

[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.

[18] Opta Sports, www.optasports.com.

[19] P. Lucey, D. Oliver, P. Carr, J. Roth, and I. Matthews, "Assessing team strategy using spatiotemporal data," in *ACM SIGKDD*, 2013.

[20] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, and I. Matthews, "Win at home and draw away: Automatic formation analysis highlighting the differences in home and away team behaviors," in *MIT Sloan Sports Analytics Conference*, 2014.

[21] J. Tenenbaum and W. Freeman, "Separating Style and Content with Bilinear Models," *Neural Computation*, vol. 12, no. 6, pp. 1247–1283, 2000.

[22] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros, "What Makes Paris Look Like Paris?" *ACM Transactions on Graphics (SIGGRAPH)*, vol. 31, no. 4, 2012.

[23] Y. Lee, A. Efros, and M. Hebert, "Style-Aware Mid-Level Representation for Discovering Visual Connections in Space and Time," in *ICCV*, 2013.

[24] S. Roberts, R. Everson, and I. Rezek, "Minimum Entropy Data Partitioning," *IET*, pp. 844–849, 1999.

[25] Y. Lee and S. Choi, "Minimum Entropy, K-Means, Spectral Clustering," in *International Joint Conference on Neural Networks*, 2004.

[26] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.