"Win at Home and Draw Away": Automatic Formation Analysis Highlighting the Differences in Home and Away Team Behaviors

Alina Bialkowski, Patrick Lucey, Peter Carr, Yisong Yue and Iain Matthews
Disney Research, Pittsburgh, PA, USA, 15213
Email:{alina.bialkowski, patrick.lucey, peter.carr, yisong.yue, iainm}@disneyresearch.com

Abstract

In terms of analyzing soccer matches, two of the most important factors to consider are: 1) the formation the team played (e.g., 4-4-2, 4-2-3-1 etc.), and 2) the manner in which they executed it (e.g., conservative - sitting deep, or aggressive - pressing high). Despite the existence of ball and player tracking data, no current methods exist which can automatically detect and visualize formations. In this paper, we present an automatic method which can detect formations using an Expectation-Maximization (EM) approach. Using an entire season of Prozone data which consists of ball and player tracking information from a recent top-tier professional league, we showcase our method by investigating the "home advantage". In a paper we published recently, using an entire season of ball tracking data we showed that home teams had significantly more possession in the forward third which correlated with more shots and goals while the shooting and passing proficiencies were the same. Using our automatic formation analysis, we extend this analysis, and show that teams tend to play the same formation at home as they do away, but the manner in which they execute it is significantly different. Specifically, we show that the formation of teams at home is significantly higher up the field compared to when they play away. This conservative approach at away games suggests that coaches aim to *win their home games and draw their away games*. Additionally, we also show that our method can visually summarize a game which gives an indication of dominance and tactics. While enabling new discoveries of team behavior which can enhance analysis, it is also worth mentioning that our automatic formation detection method is the first to be developed.

1 Introduction

As chronicled in Jonathan Wilson's *Inverting the Pyramid* [1], the many tactical and strategic revolutions that have occurred in soccer over the last century can be observed in the variations in formations utilized by coaches and managers over this time. For example, in the very first international match in 1872, Wilson notes that England played what looked like a 1-2-7 (one full-back, two mid-fielders and seven attackers), while Scotland played a 2-2-6. As the game evolved through various rule-changes (i.e., off-side rule) and professionalism (i.e., players could train full-time), so too did formations where coaches/managers set up their team's formation to best maximize the chances of their team winning while trying to minimize the chances of the opposition. Prime examples of such formations are the dour and stifling "Catenaccio", often employed by Italian teams where the emphasis is getting behind the ball waiting for the opposition to make a mistake and hit them on the counter-attack, or the dynamic and fluid "Total-Football" introduced by Rinus Michaels through the Ajax and Dutch teams in the 60's and 70's which has evolved to the "Tiki-Taka" style of football Barcelona play today using a 4-3-3 formation (N.B. Lobanovskyi used a similar style with the Dynamo Kiev team at the same time with similar success). While the differences in these formations are readily evident to the trained eye and despite the fact that most professional soccer teams are extremely sophisticated in terms of their knowledge of team tactics and strategy, analytical measures which can quantify such team behaviors are lacking. This is understandable though as compared to other sports, soccer is a dynamic, continuous game with events (e.g., shots, goals etc.) occurring sparsely. Coupled with the fact that players continuously swap position or role within a formation, making meaningful comparisons is very difficult.

While challenging, some objective measures are starting to emanate from new data sources that contain ball-events (Opta [2]), or ball and player tracking information (Prozone [3]). In *The Numbers Game* [4], the authors highlight some recent measures which debunk some commonly-held beliefs. Notable examples include, "a team is most likely to concede a goal just after it has scored", and "the greater the number of corners a team has, the more likely they are to score". In both these examples they show that this is in fact *not* the case, with teams *least* likely to concede after they score a goal in the first example, and that the relationship between goals and corners is *essentially zero* in the second example. Contrastingly though, another commonly-held belief in soccer is that of the "home advantage" - where teams are more likely to win at home compared to away. In *Scorecasting* [5], Moskowitz and Wertheim uphold this belief by highlighting that the home advantage exists in all professional sports and suggest that referees play a significant role by giving home teams favorable calls at critical moments. Specifically for soccer, they show that event statistics such as the amount of injury time, number of yellow cards and number of penalties awarded to home-teams reinforce their hypothesis.
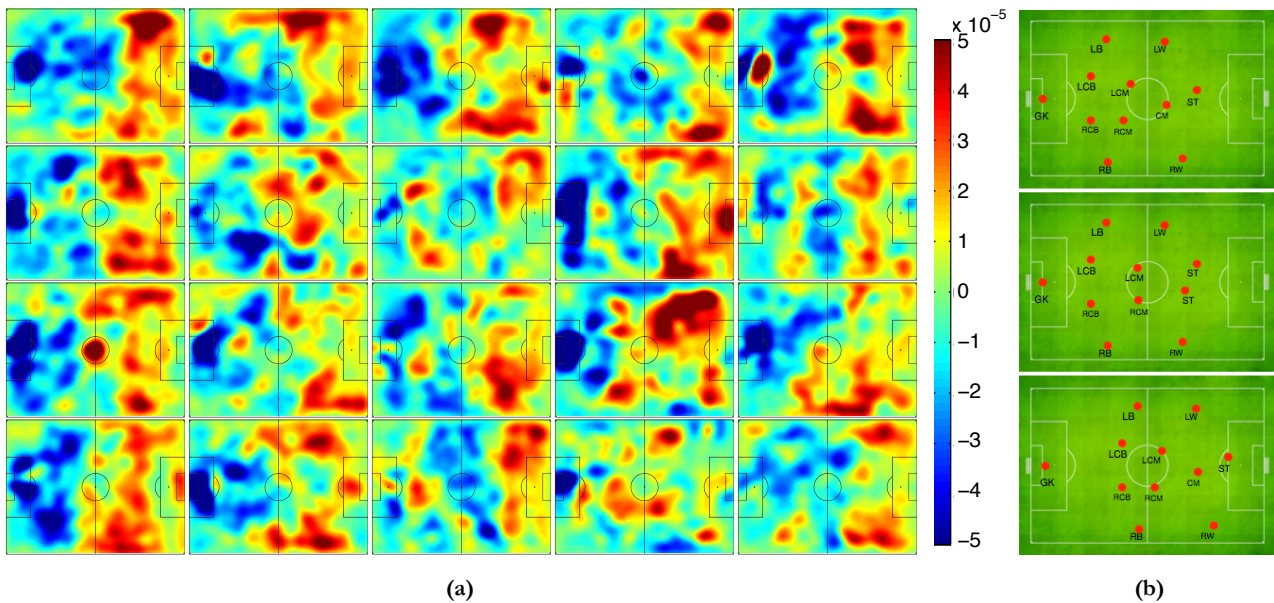
Figure 1: (a) Using only ball tracking data from Opta, in our recent paper [6] we showed that home teams have far more possession in the forward third which correlates with a significantly higher number of shots and goals. We observe the same finding in this new dataset, and above show the normalized difference maps between home and away ball occupancy for all 20 teams. (All teams attack left-to-right, and positive values refer to teams having more possession in this location in home games.). (b) Using player tracking data from Prozone, in this paper we dig a bit deeper by looking at the formations teams use and seeing if they play different formations at home games - e.g., 2 strikers instead of 1 (middle), or if they just play it differently - e.g., higher up the field (bottom).

To explore if there was any strategic explanation, we recently investigated this problem by analyzing an entire season's worth of ball tracking data [6]. In that paper, we found that even though there was no difference in shooting or passing proficiency, home teams had significantly more shots and goals. Most notably, we found that home teams had significantly more possession in the forward third, compared to away teams (see Figure 1(a)), which suggests that away teams play more conservatively than home teams. This backs up another commonly held belief that teams should aim to "win their home games and draw their away ones", which suggests that managers employ a conservative strategy at away games. In this paper, we dig deeper into this phenomenon by investigating whether:

1) teams play different formations at away games (i.e., do managers elect to play only one striker away rather than two), or
2) the team plays the formation differently (i.e., the formation is the same, but they just played more defensively).

With the aid of a whole season of player and ball tracking data from Prozone [3] from a top-tier professional soccer league, we reinvestigate this phenomenon by analyzing each team's formation at home and compare it to their away formations. As each game has more than a million data points (i.e., all 22 players at 10 frames-per-second), having an expert human label formations across an entire season is prohibitive and so requires an automatic solution. Even though formation analysis is done today in a very qualitative way by an expert human analyst (e.g., zonalmarking [7]), no method exists which can automatically represent and detect formations of a team in soccer. The biggest problem is dealing with the constant swapping of player roles that occurs within a formation as this introduces noise into the signal, making comparisons difficult. This problem though, also provides a clue to a solution as the dynamic swapping of player positions suggests that our method also needs to dynamically update over time. Leveraging an Expectation-Maximization (EM) method, we show how this can be done to detect and visualize formations from player tracking data (Section 3). Using this analysis, we discover many unique characteristics that teams employ during a season (Section 4). Additionally, we show that our method can be used to dynamically visualize and summarize the game in a very meaningful and quick way (Section 5). Before we showcase this work, we first re-explore the "home advantage".

2   Re-Exploring the Home Advantage

In our previous work [6], we analyzed an entire season of ball-tracking data from the English Premier League data from Opta and we found teams had approximately the same number of passes, passing accuracy and shooting accuracy when playing at home and away. However, there was significantly more shots and goals for home teams, and we found that this coincided with home teams having more possession in the forward third (see Figure 1(a)). Before proceeding, we wanted to see if this same phenomenon occurred across a season of data from a different data source. To explore formations and the way players move during games, we used an entire season of ball and player tracking data from Prozone [3]. Due to the proprietary nature of the data, we unfortunately

Event Statistic	Mean for Home Team	Mean for Away Team	P-Value
Points (win = 3pts, draw = 1pt, loss = 0pts)	1.61 per match	1.10 per match	<0.001
Goals	1.57 per match	1.21 per match	<0.001
Shots on target	6.48 per match	5.21 per match	<0.001
Shots not on target	8.94 per match	6.97 per match	<0.001
Shooting accuracy	41.7%	42.4%	0.5046
Passes	451 per match	436 per match	0.1483
Occupancy in final third (of time in play)	14.1% per match	11.8% per match	<0.001
Team x-centroid	-1.79 m	-4.90 m	<0.001
Team y-centroid	0.064 m	-0.047 m	0.4397

(a)   **(b)**

Figure 2. (a) Table showing that the home advantage exists on the dataset that we worked on for this paper, obtaining similar results to our paper in [6]. (b) Using the player tracking data, we show the mean team centroid (top) and team spread (bottom) and can see that the home teams had a mean position which was significantly more forward than when they played away (highlighted by the second last entry in the table).

can not disclose the identity of the league or the teams. The dataset consisted of 20 teams who played every team home and away and the data consisted of the location of every player at 10 frames-per-second as well as the ball events. To anonymize the identity of each team, we randomly assigned each team a unique identifier "A-T". The results of this analysis is shown in Figure 2.

As can be seen in Figure 2(a), the same pattern exists in the dataset that we used for this work. Most notably, home teams had more shots and goals, as well as more points and possession in the forward third, while the shooting proficiency and number of passes was roughly the same. Additionally in Figure 2(b), we show that home teams are more advanced than away teams in terms of their mean position, while the position across the field is the same. This is a rather interesting finding as it suggest something is different in the way the teams play: does it mean that teams play with two strikers at home while they play with an extra midfielder away, or is this a global trend of the formation where all the players pushed forward? To do this we first have to detect and visualize the formations.

## 3 Automatically Detecting Formations from Tracking Data

Coordination between players within a team is essential for a team's success, in terms of providing coverage of the large field area, and in scoring and defending against their opposition from scoring. To achieve this, the coach/manager designates a formation or overall structure of play to coordinate the team's movements and each player is assigned a particular responsibility or *role* within the formation. For example, in a 4-3-3 formation, the roles may be defined as R = {goal-keeper, left-back, right-back, left center back, right center back, left center midfield, right center midfield, left wing, right wing, attacking center midfield, striker}, which represent

the players' positions relative to one another (i.e. the left wing, by definition, is positioned to the left relative to other players) and their general location on the field. As a formation reflects the relative spatial arrangement of players and not so much with the absolute position of the player, we normalized for translation (i.e. removed the centroid of position of the formation so that they are centered on the half-way line - goal-keeper excluded) in each frame. We then characterized the formation over a match, by calculating the mean and covariances of the role positions about the centroid as shown in Figure 3.

Due to the dynamic nature of soccer, players frequently swap roles throughout a match, making the representation and automatic

(a)   **(b)**
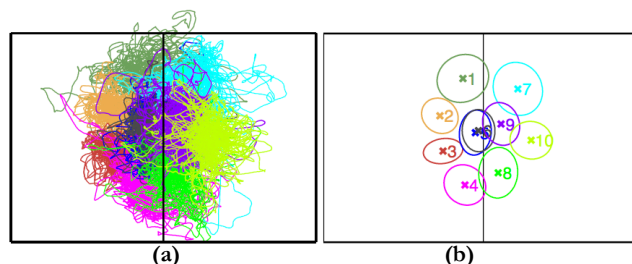
Figure 3. (a) The positions of the players at every frame across a match relative to the centroid and ordered by player identity. (b) The formation can be represented using the mean and covariances of the player positions about the team centroid (drawn relative to the field-centre).
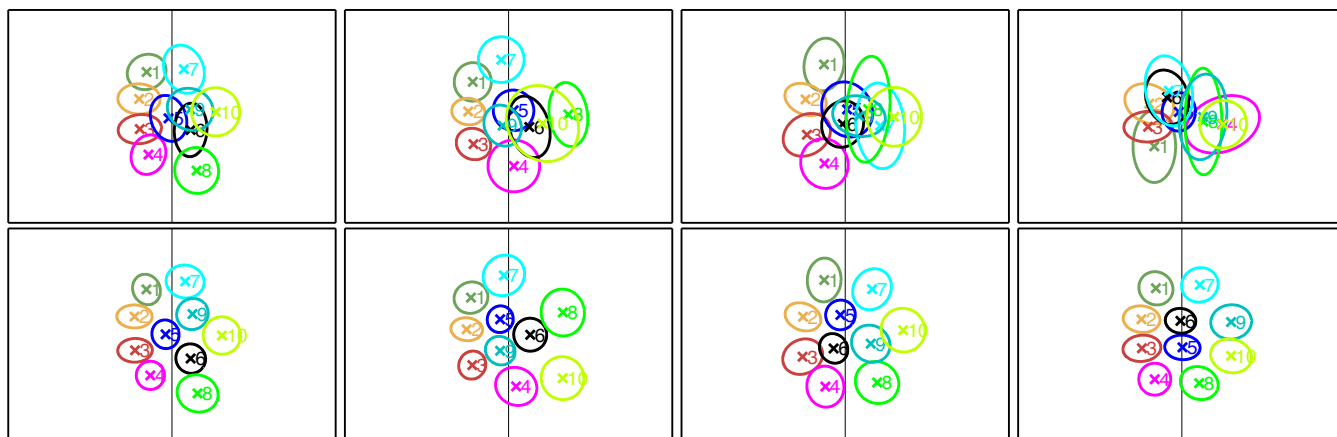
Figure 4. (Top) Shows the mean positions of the players over a half of a game - notice the heavy overlapping that occurs between the midfielders and the forwards. (Bottom) Using our approach we can accurately assign player roles at each frame which disambiguates role and allows formation analysis to occur.
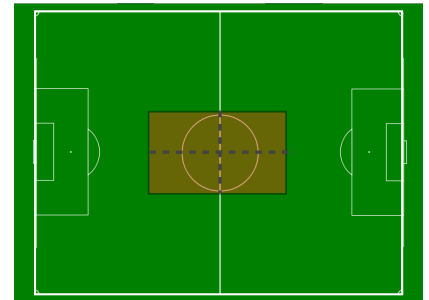
detection of a formation challenging. For example, using a static ordering of players throughout a match (e.g. ordering players by their starting roles or identity), the frequent role swaps throughout a match cause great overlap in the player distributions, as shown in the top row of Figure 4, and hence, a static ordering of players does not accurately represent the structure or formation of the team. To overcome this, we use a dynamic ordering of players by the role that they occupy at a given instant in time. However, a big challenge in achieving this is that we don't know what formation a team is playing in advance, so we can not assign roles. We solve the two tasks of formation discovery and role assignment simultaneously using an Expectation-Maximization approach on the player tracking data, where we first assign initial player role based on the mean positions of the players. We then use the Hungarian algorithm [9] to update the role of each player in every frame and calculate the error between the previous and updated assignments which is similar to [8]. We then continue to iterate this process until convergence. This process normally takes 5-10 iterations to converge for an entire half of soccer and some results of this process are shown at the bottom of Figure 4.
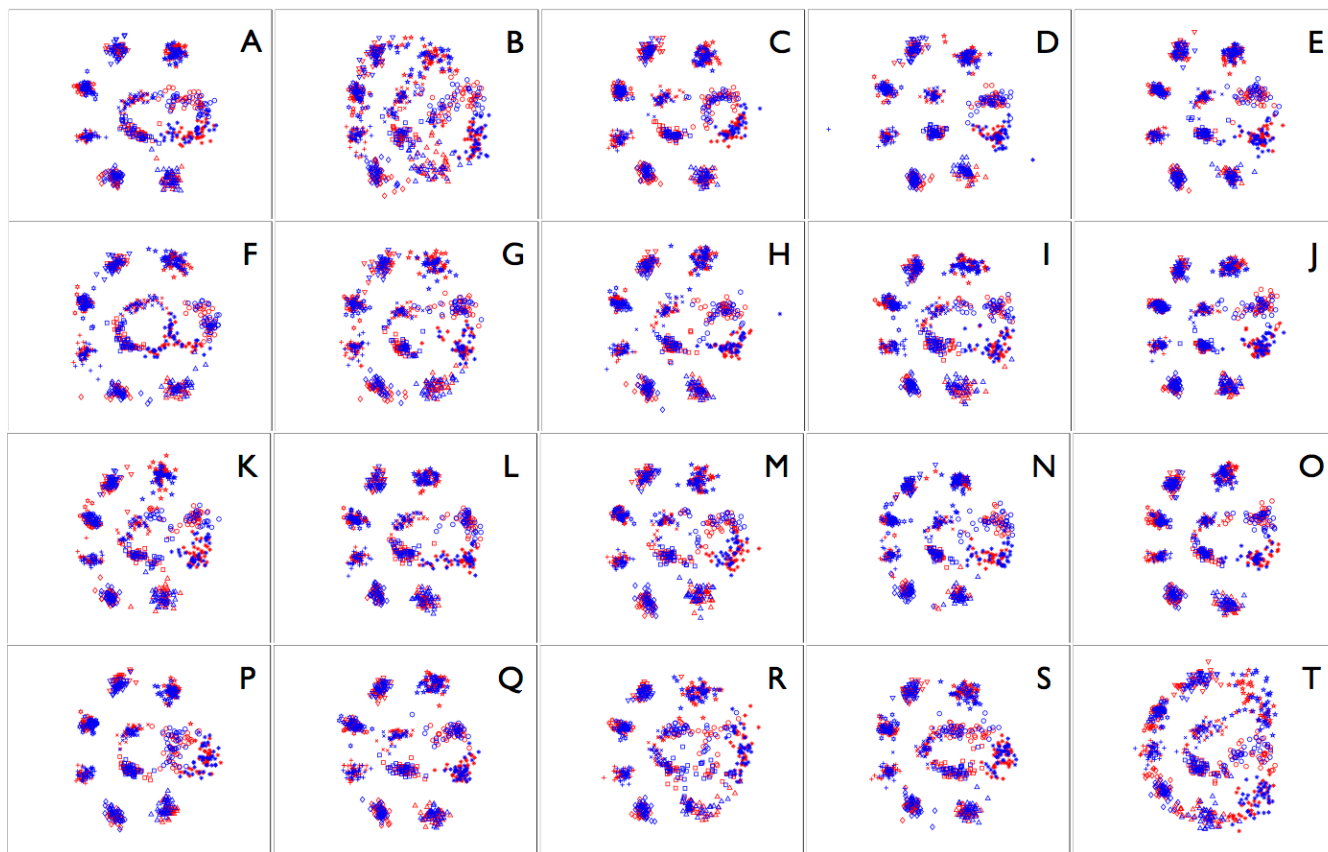
Figure 5. Mean formation for each match half relative to the centroid (centre of each rectangle) overlaid over one another for each team (A-T). Red represents home games and blue represents away games, with all teams normalized to attack from left to right.

Given that we have disentangled the player positions and generated a dynamic role for each player at each frame, we can use these labels to represent the mean formations for each match for every team. The formations for each of the 20 teams (A-T) for every match are visualized in Figure 5, where the red corresponds to their home performances and blue corresponds to away. As can be seen in this figure, most of the teams tend to play the same formation with only a slight variation occurring in some of the positions. For example, only teams B and T seem to have some variation across the course of a season, while others like teams A, F, P and R only have a minor change in the midfield (i.e. player 1 holding midfielder vs 2, or player with one striker vs two). Other than that, most teams tend to be rather staunch in what they play regardless of whether they are playing at home or away. The most dominant formation appears to be a 4-4-2, with some teams varying the midfield as described above. Only one team appeared to dabble with a back three (team T) and this variation is quite obvious but it was consistent for both home and away performances.

4   Formation Style Analysis

In the previous section we showed that each team played pretty much the same formation regardless whether they were at home or away. This begs the question then, *how did they play the formation? Were they more attacking, or were they more defensive?*

Figure 6. To get a closer look at the formation differences, we conducted analysis on a zoomed in area of the field.
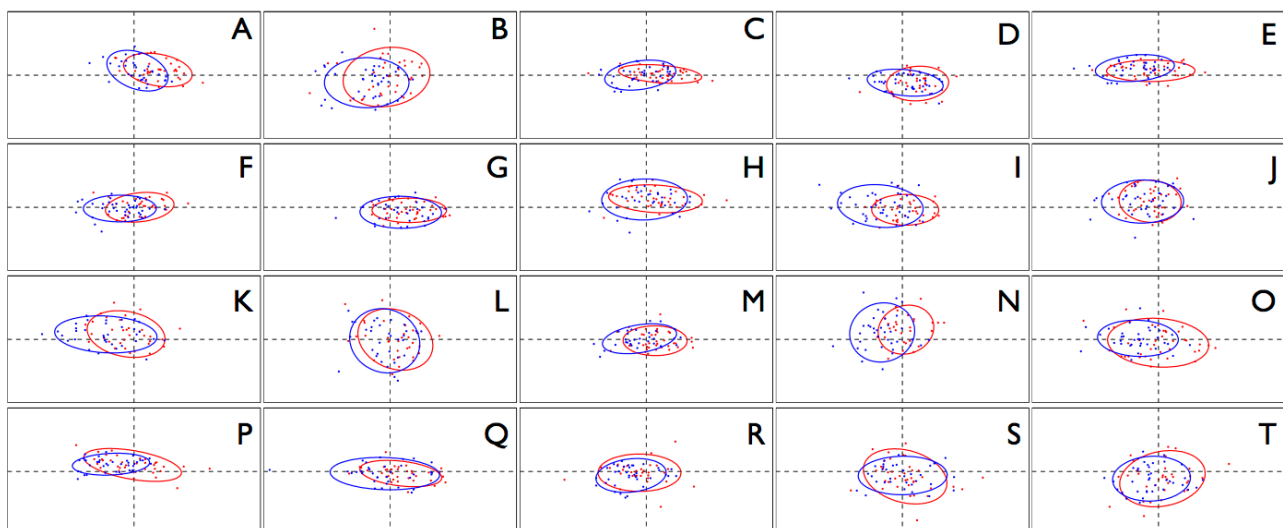
(a) Mean position of the formation when they were in possession
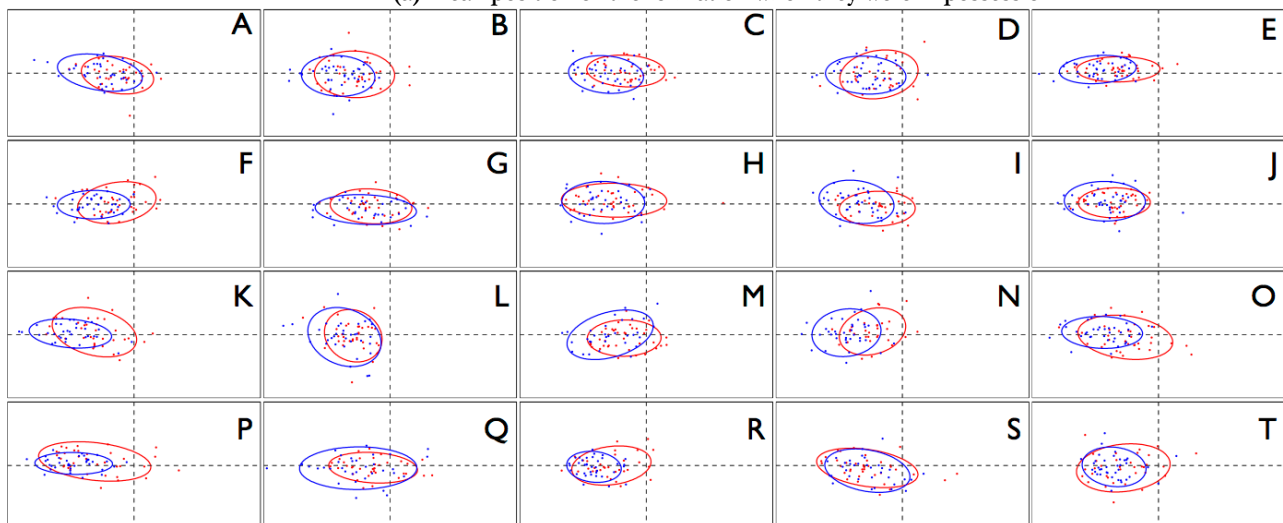
(b) Mean position of the formation when the opposition were in possession

Figure 7. Plots of the mean position of the team when they were in (a) possession, and (b) when the opposition were in possession. The red ellipses refer to their behavior when they were at home and the blue ellipses when they played away.
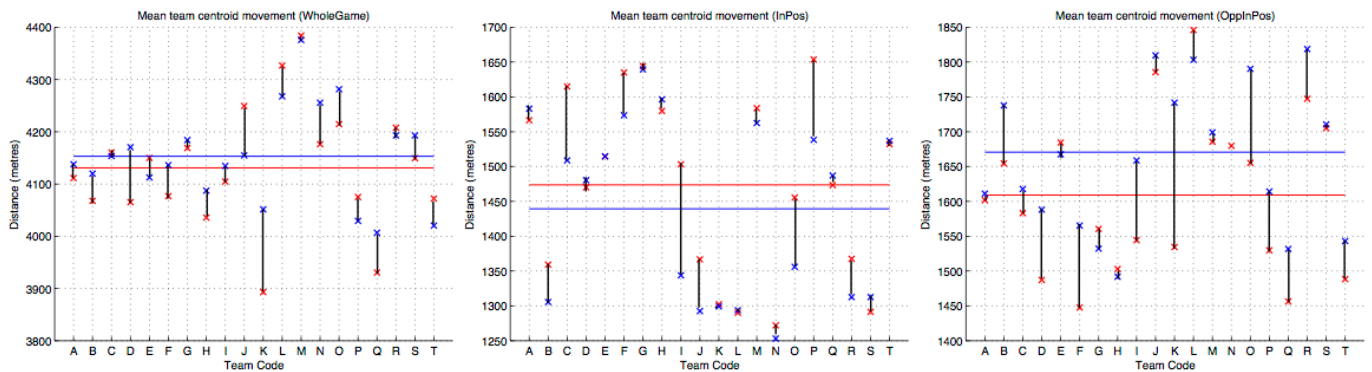
Figure 8. Plots showing the distance ran for each team when they played at home (red) compared to when they played away (blue). (Left) Overall, across the league we can see that the away teams (blue-line) covered more distance than the home team (red-line). (Middle) While in attack, home teams tended to run more than away teams, mostly because they are exploiting more area, but the discrepancy between the two is small compared to the amount of distance away teams had to cover in defense (right). As the amount of possession for both with similar, these findings are significant.

To answer these questions, we looked at the centroid of each formation and looked at each team's behavior when they were: a) in possession, and b) when the opponent was in possession (we disregarded the times when the ball was out for a stoppage, which is amazingly around 50% of the time). These plots for the zoomed in area of the field (see Figure 6) are shown in Figure 7.

As it is very obvious from these plots, it can be seen that nearly all teams are more forward at home than they are away, both when they attack and when they defend. This can help explain how teams have more possession in the forward third at home, as simply having the players in more advanced positions suggest that they would have more possession in these advanced areas. Similarly, when they are defending higher up the pitch, it means that they are more likely to regain possession higher up the pitch. The downside to this tactic, is that they leave more space exposed behind the defensive line which allows teams to be hit on the counter attack. However, in addition to winning the ball in more advanced positions (which is closer to the opponent's goal, which will possibly lead to more shots and more goals seeing the accuracy is roughly the same), the home team will expend less energy. The implications of this are large as the less energy a team expends, the longer and more sustained attacks a team can lead. In Figure 8, we show the difference in distance that teams run at home compared to away games (again red refers to home teams, and blue refers to away teams).

5   Match Summarization via Formation Visualization

Statistics given at half-time are quite poor as they only contain sparse statistics which are not symbolic of how the team played or what the opposition are doing both in terms of formation and how they executing the formation (i.e. attacking or defending deep). A powerful statistic which broadcasters use during a live-broadcast is to give the possession of both teams over the last 5 minutes which gives a indication on which team is dominating or not. While this is very insightful, it does not give any information about where this is happening. Using our formation detection approach, we can dynamically see how the match is played quickly over the half by using a sliding window of 5 minutes to see what formation they are playing and where it is relative to the other team. A film-strip of our approach is shown in Figure 9.
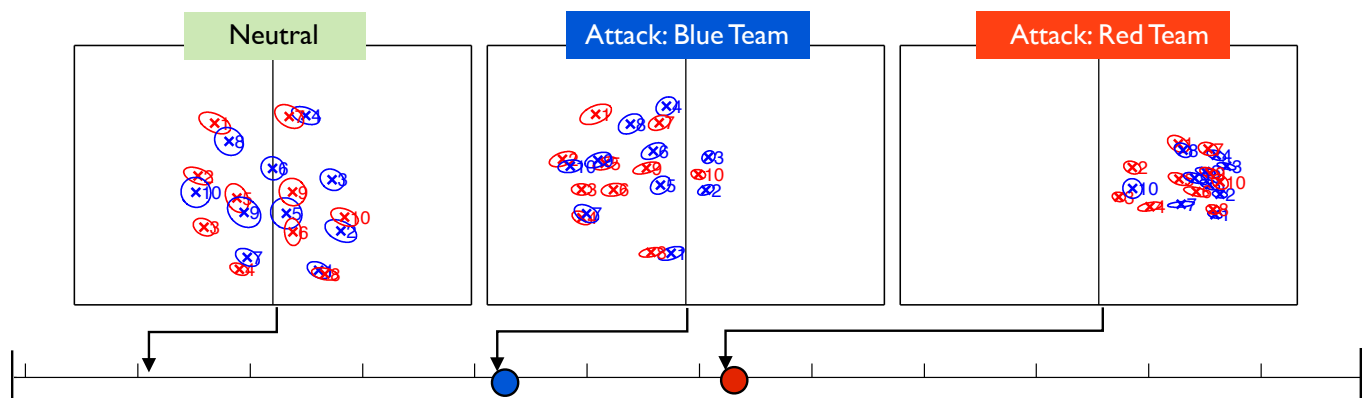
Figure 9. By representing formations within a match using a sliding window, we can see how the game progresses in terms of formations and location on the field. A 45 min half timeline is shown (circles = goals, yellow squares = yellow cards), and the utility of our formation and role representation is demonstrated above, with the home team (red), attacking from left to right. We can see that during a neutral portion of the game, the teams are playing what appears to be a 4-1-4-1 (red), and 4-2-3-1 (blue). Before the blue team scores, we can see a midfielder (role 9) moves forward to aid in the attack. In the final example, the red team scores, with the whole team positioned close to the goal.

6 Summary and Future Work

In this paper, we presented a novel method which could automatically detect and visualize formations from soccer player tracking data. Utilizing an entire season's data from Prozone from a top-tier professional league, we used our automatic approach to see whether the formation they played had anything to do with explaining why teams are more successful at home rather than away. Our analysis showed that nearly all teams tend to play the same formation at home as they do away, however, the way they executed the formation was significantly different. Specifically, we were able to show that at home, teams played significantly higher up the field compared to when they played away (or conversely, teams sat much deeper at away games). This conservative approach at away games suggests that coaches aim to win their home games and draw their away games. Additionally, we also show that our method can visually summarize a game which gives an indication of dominance and tactics. While enabling new discoveries of team behavior which can enhance analysis, it is also worth mentioning that our automatic formation detection method is the first to be developed. In future work, we aim to use the team's formation information as additional contextual features which will enhance short-term prediction (i.e., see if we can predict the next pass), as well as using it to discovery discriminative or unique patterns teams use in offensive and defensive situations.

References

[1] J. Wilson, "Inverting the Pyramid: The History of Football Tactics", Orion Books, 2008.

[2] Opta Sports. http://www.optasports.com

[3] Prozone. http://www.prozonesports.com

[4] C. Anderson and D. Sally, "The Numbers Game: Why Everything You Know About Soccer is Wrong", Penguin Books, 2013.

[5] T. Moskowitz and L. Wertheim. Scorecasting: The Hidden Influences Behind How Sports Are Played and Games Are Won. Crown Publishing Group, 2011.

[6] P. Lucey, D. Oliver, P. Carr, J. Roth and I. Matthews, "Assessing Strategy using Spatiotemporal Data", in KDD, 2013.

[7] Zonalmarking. http://www.zonalmarking.net

[8] 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.

[9] H. W. Kuhn, "The Hungarian Method for the Assignment Problem", in Naval Research Logistics Quarterly, vol. 2, no 1-2, pp. 83-97, 1955.