Assessing Team Strategy using Spatiotemporal Data

Patrick Lucey
Disney Research
Pittsburgh, PA, USA
patrick.lucey@disneyresearch.com

Dean Oliver
ESPN
Bristol, CT, USA
dean.oliver@espn.com

Peter Carr
Disney Research
Pittsburgh, PA, USA
peter.carr@disneyresearch.com

Joe Roth
Disney Research
Pittsburgh, PA, USA
joe.roth@disneyresearch.com

Iain Matthews
Disney Research
Pittsburgh, PA, USA
iaim@disneyresearch.com

ABSTRACT

The Moneyball [14] revolution coincided with a shift in the way professional sporting organizations handle and utilize data in terms of decision making processes. Due to the demand for better sports analytics and the improvement in sensor technology, there has been a plethora of ball and player tracking information generated within professional sports for analytical purposes. However, due to the continuous nature of the data and the lack of associated high-level labels to describe it - this rich set of information has had very limited use especially in the analysis of a team’s tactics and strategy. In this paper, we give an overview of the types of analysis currently performed mostly with hand-labeled event data and highlight the problems associated with the influx of spatiotemporal data. By way of example, we present an approach which uses an entire season of ball tracking data from the English Premier League (2010-2011 season) to reinforce the common held belief that teams should aim to “win home games and draw away ones”. We do this by: i) forming a representation of team behavior by chunking the incoming spatiotemporal signal into a series of quantized bins, and ii) generate an expectation model of team behavior based on a code-book of past performances. We show that home advantage in soccer is partly due to the conservative strategy of the away team. We also show that our approach can flag anomalous team behavior which has many potential applications.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous; I.2.6 [Learning]: General

Keywords

Sports Analytics; Spatiotemporal Data; Representation

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD ’13, August 11–14, 2013, Chicago, Illinois, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
Copyright 2013 ACM 978-1-4503-2174-7/13/08 ...$15.00.

1. INTRODUCTION

In his 2003 book Moneyball [14], Michael Lewis documented how Oakland A’s General Manager Billy Beane was able to effectively use metrics derived from hand-crafted statistics to exploit the inefficiencies in the value of individual baseball players. Around the same time, Basketball on Paper [24] was published which outlined methods for valuing player performance in basketball which is a far more challenging problem because it is a continuous team sport. Due to the popularity and effectiveness of the tools that emanated from these works, there has been enormous interest in the field of sports analytics over the last 10 years with many organizations (e.g. professional teams, media groups) housing their own analytics department. However, nearly all of the analytical works have dealt solely with hand-labeled event data which describes what happened (e.g. basketball - rebounds, points scored, assists; football - yards per carry, tackles, sacks; soccer - passes, shots, tackles; – see Figure 1(a)). Once the data is in this form, most approaches query for relevant data from a database, then applying sport-based rules and standard statistical methods, including regression and optimization.

As most sporting environments tend to be dynamic with multiple players continuously moving and competing against each other, simple event statistics do not capture the complex aspects of the game. To gain an advantage over the rest of the field, sporting organizations have recently looked to employ commercial tracking technologies which can locate the position of the ball and players at each time instant.
in professional leagues [9, 26, 29, 32] - to determine where and how events happen. Even though there is potentially an enormous amount of hidden team behavioral information to mine from such sources, due to the sheer volume as well as the noisy and variable length of the data, methods which can adequately represent team behaviors are yet to be developed. The value of this data is limited as little analysis can be conducted. What compounds the difficulty of this task is the low-scoring and/or continuous nature of many team sports (e.g. soccer, hockey, basketball) which makes it very hard to associate segments of play with high-level behaviors (e.g. tactics, strategy, style, system, formations). Without these labels, which essentially give the game context, inferring team strategy or behavior is impossible as there are no factors to condition against (see Figure 1(b)).

Due to these complexities, there has been no effective method of utilizing spatiotemporal data in continuous sports. Having suitable methods which can first develop suitable representations from imperfect (e.g. noisy or impartial) data, and then learn team behaviors in an unsupervised or semi-supervised manner, as well as recognize and predict future behaviors would greatly enhance decision making in all areas of the sporting landscape (e.g. coaching, broadcasting, fantasy-games, video games, betting etc.). We call this the emerging field of Sports Spatiotemporal Analytics - and we show an example of this new area of analytics by comparing the strategies of home and away teams in the English Premier League using ball tracking data.

2. RELATED WORK

The use of automatic sports analysis systems have recently graduated from the virtual to the real-world. This is due in part to the popularity of live-sport, the amount of live-sport being broadcasted, the proliferation of mobile devices, the rise of second-screen viewing, the amount of data/statistics being generated for sports, and demand for in-depth reporting and analysis of sport. Systems which use match statistics to automatically generate narratives have already been deployed [2, 33]. Although impressive, these solutions give a low-level description of match statistics and notable individual performances without any tactical analysis about factors which contributed to the result. In tennis, IBM Slamtracker [10] provides player analysis by finding patterns that characterize the best chance a player has to beat another player from an enormous amount of event labeled data - although no spatiotemporal data (i.e. player or ball tracking information) has been used in their analysis yet.

Spatiotemporal data has been used extensively in the visualization of sports action. Examples include vision-based systems which track baseball pitches for Major League Baseball [31], and ball and players in basketball and soccer [29, 32]. Hawk-Eye deploy vision-based systems which track the ball in tennis and cricket, and is often used to aid in the officiating of these matches in addition to providing visualizations for the television broadcasters [9]. Partial data sources normally generated by human annotators such as shot-charts in basketball and ice-hockey are often used for analysis [23], as well as passing and shot charts in soccer [26]. Recently, ESPN developed a new quarterback rating in American Football called “TotalQBR” [25] which assigns credit or blame to the quarterback depending on a host of factors such as pass or catch quality, importance in the match, pass thrown under pressure or not. As these factors are quite subjective, annotators who are reliable in labeling such variables are used. In terms of strategic analysis, zonalmarking.net [37] describes a soccer match from a tactical and formation point of view. This approach is very qualitative and is based solely on the opinion of the analyst.

As the problem of fully automatic multi-agent tracking from vision-based systems is still an open one, most academic research has centered on the tracking problem [1, 5, 17, 27]. The lack of fully automated tracking approaches has limited team behavioral research to works on limited size datasets. The first work which looked at using spatiotemporal data for team behavior analysis was conducted over 10 years ago by Intille and Bobick [12]. In this seminal work, the authors used a probabilistic model to recognize a single football play from hand annotated player trajectories. Since then, multiple approaches have centered on recognizing football plays [15, 16, 30, 34], but only on a very small number of plays (i.e. 50-100). For soccer, Kim et al. [13] 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. [4] 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. [28] used trajectories of player movement to recognize three types of team offensive patterns. Morariu and Davis [21] integrated interval-based temporal reasoning with probabilistic logical inference to recognize events in one-on-one basketball. Hervieu et al. [11] 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 [8, 36]. The lack of extensive tracking data to adequately train models has limited the success of the above research.

It is clear from the overview given above, that there exists a major disparity in resources between industry and academia to deal with this problem domain. Sporting organizations receive large volumes of spatiotemporal data from third-party vendors but often lack the computational skills or resources to make use of it. Research groups on the other-hand, have the necessary analysis tools but cannot access these large data repositories due to the proprietary nature of commercial tracking systems. Recently however, due to the potential payoff, some industry groups are investing in analytical people with these skill sets, or have teamed up with research groups to help facilitate a solution. The release of STATS Sports VU data [32] to some research groups has enabled interesting analysis of shots and rebounding in the NBA [7, 20]. In tennis, Wei et al. [35] used ball and player tracking information to predict shots using data from the 2012 Australian Open. For soccer, researchers have charac-

\(^1\) As nearly all professional leagues currently forbid the use of wearable sensors on players, unobtrusive data capture methods such as vision-based systems or armies of human annotators are used to provide player and ball tracking information. However, this restriction may change soon as monitoring the health and well-being of players has attracted significant interest lately, especially for concussions in American Football [19], as well as heart issues in soccer [3].

\(^2\) These works only capture a portion of the field, making group analysis very difficult as all active players are rarely present in the all frames.
### 3. CASE STUDY: HOME ADVANTAGE IN SOCCER

#### 3.1 “Win at Home and Draw Away”

In a recent book [22], Moskowitz and Wertheim highlight that the home advantage in soccer is due to the conservative strategy that away teams use (or more aggressive approach of the home team) which reinforces the commonly held belief that teams aim to win their home games and draw their away ones.

### Table 1: Table showing the statistics for the home and away performances for each team in the 2010 EPL season: (left columns) home matches (right columns) away columns (Key: W = wins, D = draws, L = losses, P = points (3 for a win, 1 for a draw, 0 for a loss), GF = goals for, SF = shots for, GA = goals against, SA = shots against).

<table>
<thead>
<tr>
<th>Team</th>
<th>Home</th>
<th>Away</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>D</td>
<td>L</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Man Utd</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>Chelsea</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>Man City</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>Arsenal</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>Tottenham</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Liverpool</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>Everton</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>Fulham</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Aston Villa</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Sunderland</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>West Brom</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>Newcastle</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Stoke City</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>Bolton</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Blackburn</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Wigan</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Wolves</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Birmingham</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Blackpool</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>West Ham</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

**SUM:** 179 111 90 648 617 616 446 4926 90 111 179 381 446 4926 617 6162

**AVG (per game):** 0.47 0.29 0.24 1.71 1.62 1.62 1.17 13.0 0.24 0.29 0.47 1.00 1.17 13.0 1.62 16.2

Table 1: Table showing the statistics for the home and away performances for each team in the 2010 EPL season: (left columns) home matches (right columns) away columns (Key: W = wins, D = draws, L = losses, P = points (3 for a win, 1 for a draw, 0 for a loss), GF = goals for, SF = shots for, GA = goals against, SA = shots against).

This table shows various statistics for the home and away performances of different teams in the 2010-2011 English Premier League season. The columns include the number of wins (W), draws (D), losses (L), points (P), goals for (GF), goals against (GA), shots for (SF), and shots against (SA). The data is broken down by home and away matches for each team. The average points scored per game and other relevant statistics are also provided.

The data highlights several key observations:

- **Goal Scoring:** Home teams scored more goals on average (1.62 vs 1.17) compared to away teams.
- **Shots:** Home teams had more shots on target (16.2 vs 13.0) and scored more goals from those shots (1.62 vs 1.17) compared to away teams.
- **Points:** Home teams earned more points compared to away teams (1.62 vs 1.30).

These statistics reinforce the commonly held belief that teams aim to win their home games and draw their away ones.

#### 3.2 Ball Tracking and Event Data

Due to the difficulty associated with accurately tracking players and the ball, data containing this information is still scarce. Most of the data collected is via an array of human annotators who label all actions that occur around the ball - which they call *ball actions*. The F24 soccer data feed collected for the English Premier League (EPL) by Opta [26] is a good example of this. The F24 data is a time coded feed that lists all player action events within the game with a player, team event type, minute and second for each action. Each event has a series of qualifiers describing it. This method, we first describe the data.
type of data is currently used for the real-time online visualiza-
tions of events, as well as post-analysis for prominent tele-
vision and newspaper entities (e.g. ESPN, The Guardian).
Even though this data has been widely used, there are no
systems which use this data or data like this for automatic
tactical analysis.
For this work we used the 2010-2011 EPL season F24 Opta
feed, which consists of 380 games and more than 760,000
events. Each team plays 38 games each, which corresponds
with each team playing each other team twice (once home
and once away). The team names and rankings for the 2010-
2011 EPL data is given in Table 1. In our approach, to
analyze the tactics of a team we are required to know the
position of the ball and which team has possession of it at
every time step (i.e. every second). To do this, we infer the
ball location from the data feed. We describe our method
via Figure 2. At $t_0$ the ball is passed to the location at $t_1$.
The next action labeled is at $t_4$ where a player takes on
an opponent player. As nothing occurred between the time
$t_1$ and $t_4$, we infer the ball was dribbled in a straight line
and at a uniform velocity between these two locations. We
do the same thing between $t_5$ and $t_6$, before the ball is
passed from the location at $t_6$ to $t_7$. Using the same procedure,
we can estimate the ball position and team possession for
the remaining times - as the stoppages are tagged in the data
feed. It is worth noting here that all data is normalized onto
a field of size 100 x 100, with all positions given for teams
attacking left to right.

4. OCCUPANCY MAPS
Given we have ball tracking and team possession informa-
tion, we can partition the ball tracking data into pos-
session strings (i.e. continuous movement of the ball for
a single team without turnover or stoppage), where $\mathcal{O}^A =
\{o^A_0, \ldots, o^A_{J-1}\}$ and $\mathcal{O}^B = \{o^B_0, \ldots, o^B_{J-1}\}$ refer to the posses-
sion strings associated with each team and $I-1$ and $J-1$ are
the number of possessions. We then quantize the field
into $D$ bins where $D = I \times W$ equal size areas and vectorize
the field via the columns. As the possession strings vary in
length, we apply a sliding window of length $T$ to quantize or
chunk each possession string, $o$, into a set of play-segments
$S = \{s_0, \ldots, s_{N-1}\}$ of equal length, where $N$ is the total
number of play-segments for a possession, and $M$ is the to-
tal number of play-segments for a team over a match. Given
that a team’s possession string is of length $T$, the resulting
number of play-segments for each possession is therefore:
$N = (T - T) + 1$. If the possession string is smaller in dur-
tion than $T$, we discard it. To represent each play-segment
$s = \{s_0, \ldots, s_{T-1}\}$, the quantized ball position is tabulated
at each time step. An example of the description is given in
Figure 3.

From our set of play-segments for a given team, $S_A$ or $S_B$.
we can build a distribution to characterize the expected be-
havior in each location. We do this as follows: Given the
observations of a team, we determine the subset of play-
segments $S_d$, which originated in quantized area $d$. From
the play-segment vectors within $S_d$, we keep a count of the
locations of where the ball travels or occupies during these
play-segments. As this representation will yield a very high-
dimensional feature vector ($D^3$), we would prefer the dimen-
sionality to be $D$ for visualization purposes. To achieve this,
we can use either the mean, median, mode, total count or
even an entropy measure to describe each $p(S_d)$, which will
yield an occupancy or team behavioral map. Vectorizing the
occupancy map, gives us our $D$-dimensional spatiotemporal
feature vector $x$. The mean occupancy maps using entropy
to describe each area is shown in Figure 4 and can give a
indication of redundant patterns. The top ranked teams have
higher entropy over most of the field compared to the lower
ranked teams - this gives an indication that these teams util-
ize more options (i.e. less predictable) which is intuitive
as these teams normally have more skilled players. As the
frequency counts incorporate temporal information (more
counts, means the ball is moving quicker), we used the to-
tal count of entities as the entropy measure normalizes this
information.

Figure 2: (From the XML feed, we can infer the ball position and possession at every time step (solid lines and dots are annotated, dotted lines are inferred).

Figure 3: Given a possession string, we break it up into $N$ equal chunks and use the quantized ball position values as our play-segment representation, $s$. 

Table 2: Table showing the identification rate by breaking the plans into different time series and the field into different quantization areas. The results are the trace of the team actions using the MAPR framework we described previously is when they deviate from this style and quantify it? Moreover, can we detect their overall goal (i.e., did they win the match or not?), we can find the plans or a sequence of plans which correlate with a player winning and losing. These sequence of plans generated by humans. 

Figure 2: An example of IBM SlamTracker (IBM SlamTracker). 2012.
Figure 4: The mean entropy maps for each of the twenty English Premier League teams characterizing their ball movement patterns. The maps have been normalized for teams attacking from left to right. The bright red refers to high entropy scores (i.e., high variability) and the blue areas refer to low entropy scores (i.e., very predictable behavior).

5. STRATEGY ANALYSIS

5.1 Discriminating Team Behavior

Evaluating team strategy is a very difficult task. The major hurdle to overcome is the absence of strategy labels. But given we know the team identity, and assuming that teams exhibit similar behaviors over time, we can treat the task as an identification problem. We can do this by answering the question: using only ball movement information, can we accurately identify the most likely team?

We model team behavior using a codebook of past performances. If a team’s behavior is consistent, then previous matches will be a good predictor of future performances. For the experiments, we used 380 games of the season and used a leave-one-out cross validation strategy to maximize training and testing data. Before we investigate the difference between home and away performances, we first need to obtain the best possible representation. To evaluate this, we wanted to see how effective event-labeled data was in discriminating between different teams, and if having knowledge of where teams operate could boost the discriminating power.

To conduct the experiments, we compared our occupancy representation, to twenty-three match statistics currently used in analysis (e.g., passes, shots, tackles, fouls, aerials, possession, time-in-play etc.). We also combined the two inputs by concatenating the feature vectors. For classification, we used a k-Nearest Neighbor approach (with \( k = 30 \)) and all experiments were conducted using \( D = 10 \times 8 \) spatial areas (heuristically, we found those values of \( k \) and \( D \) gave the best performance). To reduce the dimensionality but maintain class separability for the spatiotemporal representation and the combined representation, we used linear discriminant analysis (LDA). We used the number of teams

Figure 5: The identification rate for correctly identifying home performances using event-labeled data (i.e., no location information), occupancy maps (i.e., no event information, just location), and a combination of the two.
As classes ($C = 20$), which resulted in a $C - 1 = 19$ dimensional input feature, $y = W^T x$, where $W$ can be found via solving

$$\arg \max_W \text{Tr}(W \Sigma_y W^T)$$

where $\Sigma_y$ and $\Sigma_w$ are the between-scatter and within-scatter matrices respectively. To test the various representations, only identity experiments on home performances were tested (home and away comparisons will be presented next). Results for these experiments are shown in Figure 5. This figure illustrates how spatiotemporal data greatly improves the discriminability between different teams (19.26% vs 38.79%), and fusing traditional and spatiotemporal statistics boosts performances again (46.70%). The confusion matrix of the combined representation shows that teams who play a similar style often get confused with each other (e.g. the top 5 teams and teams 13-15) and from viewing the entropy maps in Figure 4, we can see that these teams have similar characteristics.

### 5.2 Comparing Home vs Away Behavior

If a team plays in the same manner at home as they do away, the home model should be able to yield similar performance in identifying away matches as they do to the home matches. To test this theory, we used our home models to identify away performances and our away models to identify home performances. From the results (see Table 2), we can see that there is a drop in the hit-rate of the occupancy map representation – 8.69% for the home model tested on the away matches and 6.07% on the reverse case. Even though not excessive, this drop in performance suggests there is a change in the spatial behavior between home and away performances.

To explore this aspect further, we visualized the difference in occupancy between the home and away performances. To do this, we simply subtracted the home occupancy maps from the away maps and divided by the away occupancy to obtain a relative measure. The difference maps for all twenty teams is given in Figure 7. To make it easier to quantify the difference in occupancy, we calculated the difference with respect to certain areas of the field. Specifically we calculated the difference: for the whole field (W), the attacking half (H), and the attacking-third (T) - these values are listed below each difference map. As can be seen from the difference maps, nearly all the teams (18 out of 20) had more possession in the attacking half. More telling is that 19 out of the 20 teams had more possession in the attacking third. Seeing that the shooting proficiency is essentially the same (10% vs 9%), we can point to the observation that the more possession in the attacking third leads to more chances, which in-turn leads to more goals. A potential statistic to back this observation up is that Chelsea - who were the only team to have less occupancy in the attacking third - had the smallest discrepancy between home and away shots (only 12, the next was Arsenal with 45).

With the absence of labels to compare against it is impossible to say whether this was actually this case as other factors may have contributed to the home advantage (i.e. referee’s [22], shooting chance quality, game context (winning, losing, red-cards, key injury, derby matches etc.)) – essentially our analysis provides correlation and not causation. However, this analysis does provide a effective method of automatically flagging behavioral differences which can quickly alert an expert (such as a coach or commentator) into further or deeper analysis which can aid in the decision making process. In the next section we show some methods which can do this.

### 6. PRE/POST GAME ANALYSIS

Given a coach or analyst is preparing for an upcoming match, having a measure of how variable a team’s performance is would be quite beneficial. For example, the coach or analyst may have viewed a previous match and formed a qualitative model based on their expert observation. However, this model is only formed by a single observation and may be subject to over-fitting. Having a measure which could indicate how variable a team’s performance is would be quite useful. Given they have a feature representation of each of the previous performances of a team, our approach could be a method of determining the performance variance. To do this, the distance in feature space between each of the past performances, $y$, and the mean, $\hat{y}$, can be calculated where the mean is $\hat{y} = \frac{1}{m} \sum_{i=1}^{M} Y_i$, and where $M$ is the number of previous performances. A distance measure such as the $L_2$ norm could be used given that the input space has been scaled appropriately, which generates the distance measure via

$$\text{dist}_{m} = \| y_m - \hat{y} \|_2$$

where $m$ refers to the game of interest. Returning to our home and away performance example, one way to quantify
Figure 7: The normalized difference maps between home and away performances for all the 20 EPL teams. In all maps, teams are attacking from left-to-right and a positive value refers to a team having more occupancy in home games, while a negative value refers to more occupancy in away games. Percentages underneath each team give a value on this difference (Key: W is whole field, H refers to forward half and T refers to the attacking third.)

Figure 8: Variance in distortion for (a) home and (b) away performances.

the variation in performance for each team in the EPL is through calculating the “feature variance” or “variance in distortion” using Equation 2 (see Figure 9).

As can be seen in Figure 9(a), the variation in playing style of most home team’s is small - with the exception of teams 10 and 11 (Sunderland and West Brom). Conversely, the majority of team’s away performances were highly variable, with only teams 10 and 14 (Sunderland and Bolton) having low variance. These results suggest that predicting the playing style of home teams is quite feasible while predicting performances of away teams would be unreliable.

In terms of post-match analysis, a similar approach could be used to see if a team’s performance was within the expectation range (i.e. ±σ). A good example was Fulham’s away performance against Manchester United. In this match, they lost 2-0 and conceded both goals in the first half (12th and 32nd minute). The comparison between this game and their typical performance can be visualized through their occupancy maps which are given in Figure 10. In this match, Fulham occupied a lot more possession in the middle of the field than they normally do. This example highlights the importance of the match context, as after scoring two early goals, based on the data it appeared that Manchester United sat back and allowed Fulham to have the majority of possession (52% of overall possession) [6], but this possession was in non-threatening positions (i.e. possession outside of the attacking third). To account for this behavior, analysis would have to normalize for variations in behavior caused by changes in match context (i.e. score, strength of opposition etc.). However, this is a major problem as we would limit the amount of data we would have to train our model. Future work will be focussed on clustering styles unsupervised to maximize the amount of context dependent data.
7. SUMMARY AND FUTURE WORK

Most sports analytics approaches only use event-based statistics to drive analysis and decision-making despite the influx of ball and player tracking data becoming available. The reason why this new and rich source of information is being neglected stems from the fact that it is continuous, high-dimensional and difficult to segment into semantic groups. The emerging field of sports spatiotemporal analytics uses spatiotemporal data such as ball and player tracking data to drive automatic team behavior/strategy analysis which would be extremely useful in all facets of the sports industry (e.g. coaching, broadcasting, fantasy-games, video games, betting etc.). In this paper, we gave an overview of the types of sports analytics work being done both in industry as well as academia. Additionally, we gave a case study which investigated possible reasons for why the home advantage exists in continuous sports like soccer. Using spatiotemporal data, we were able to show that teams at home play have more possession in the attacking third. Coupled with the fact that the shooting and passing efficiencies are not significantly different, this observation can partially explain why home teams have more shots and score more, and in-turn win more at home compared to away matches. Using our feature representation, we also showed examples where pre and post game analysis can be performed. Specifically, we were able to show the variation in home and away performances for each team, as well as the ability to flag anomalous performances.

Our work also highlighted the importance of match context and the limiting factor it could have on training examples. In future work, we will look at unsupervised methods to cluster playing similar playing styles, which can enrich our training data set, without effecting its discriminating power. Additionally, we will explore how player tracking information can be used to discover team formations and plays. Predicting team interactions and subsequent performances and outcomes, especially when they have not played each other, is another application of our research. As reliable high-level labels are almost impossible to obtain, predicting match outcomes may be the best indicator of of successful analytical modeling.

8. REFERENCES


