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
iainm@disneyresearch.com

ABSTRACT

The *Moneyball* 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

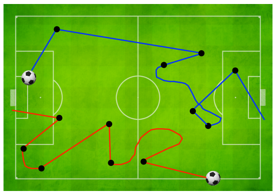
Keywords

Sports Analytics, Spatiotemporal Data, Representation

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.
Copyright 2013 ACM 978-1-4503-2174-7/13/08 ...$15.00.

Canada		USA
9(3)	Shots (on Goal)	12(4)
13	Fouls	11
3	Corner Kicks	8
1	Offsides	4
38%	Time of Possession	62%
3	Yellow Cards	1
0	Red Cards	0
4	Saves	3

(a)



(b)

Figure 1: (a) An example of standard soccer statistics based on hand-labeled event data which describe *what* happened. (b) Spatiotemporal data has the potential to describe the *where* and *how*, but as it is a continuous signal which is not associated with a fixed event, using this data for analysis is difficult.

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 just relate to parsing in the 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 [29, 32, 9, 26][1] - 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 by 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 [33, 2]. Although impressive, these solutions just give a low-level description of match statistics and notable individual performances without giving any tactical analysis about factors which contributed to the result. In tennis, IBM has created Slamtracker [10] which can provide 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 [32, 29]. 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 attempts to assign 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] attempts to describe a soccer match from a tactical and formation point of view. Whilst interesting, this approach is still 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, 27, 17, 5]. 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 [16, 30, 15, 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 type 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 [36, 8][2]. Even though notable, the lack of tracking data to adequately train models has limited the usefulness 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 that receive large volumes of spatiotemporal data from third-party vendors but often the people within these organizations lack the computational skills or resources to make use of it. Contrastingly, due to the proprietary nature of commercial tracking systems, and the cost and method of generating the tracking data, research groups who have the necessary skills can not access these large data repositories. 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]

[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.

N	Team	Home								Away							
		W	**D**	**L**	**P**	**GF**	**SF**	**GA**	**SA**	**W**	**D**	**L**	**P**	**GF**	**SF**	**GA**	**SA**
1	Man Utd	18	1	0	55	49	347	12	191	5	10	4	25	29	272	25	271
2	Chelsea	14	3	2	45	39	379	13	233	7	5	7	26	30	367	20	208
3	Man City	13	4	2	43	34	306	12	216	8	4	7	28	26	240	21	298
4	Arsenal	11	4	4	37	33	350	15	154	8	7	4	31	39	305	28	251
5	Tottenham	9	9	1	36	30	383	19	228	7	5	7	26	25	274	27	339
6	Liverpool	12	4	3	40	37	319	14	220	5	3	11	18	22	266	30	270
7	Everton	9	7	3	34	31	321	23	227	4	8	7	20	20	259	22	279
8	Fulham	8	7	4	31	30	307	23	262	3	9	7	18	19	245	20	271
9	Aston Villa	8	7	4	31	26	273	19	263	4	5	10	17	22	233	40	340
10	Sunderland	7	5	7	26	25	287	27	243	5	6	8	21	20	246	29	311
11	West Brom	8	6	5	30	30	329	30	237	4	5	10	17	26	273	41	297
12	Newcastle	6	8	5	26	41	300	27	250	5	5	9	20	15	209	30	256
13	Stoke City	10	4	5	34	31	298	18	256	3	3	13	12	15	186	30	294
14	Bolton	10	5	4	35	34	311	24	256	2	5	12	11	18	261	32	346
15	Blackburn	7	7	5	28	22	254	16	259	4	3	12	15	24	200	43	360
16	Wigan	5	8	6	23	22	290	34	227	4	7	8	19	18	221	27	284
17	Wolves	8	4	7	28	30	256	30	266	3	3	13	12	16	205	36	306
18	Birmingham	6	8	5	26	19	231	22	324	2	7	10	13	18	174	36	362
19	Blackpool	5	5	9	20	30	296	37	297	5	4	10	19	25	240	41	441
20	West Ham	5	5	9	20	24	325	31	317	2	7	10	13	19	250	39	378
SUM		179	111	90	648	617	6162	446	4926	90	111	179	381	446	4926	617	6162
AVG(per game)		**0.47**	**0.29**	**0.24**	**1.71**	**1.62**	**16.2**	**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).

used ball and player tracking information to predict shots using data from the 2012 Australian Open. For soccer, researchers have characterized team behaviors in the English Premier League using ball-motion information across an entire season using OPTA data [18]. In this paper, we extend this method to explain 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.*

3. CASE STUDY: HOME ADVANTAGE IN SOCCER

3.1 "Win at Home and Draw Away"

In a recent book by Moskowitz and Wertheim [22], they highlight that the home advantage exists in all professional sports (i.e. teams win more at home than away). The authors hypothesized that referees play a significant role by giving home teams favorable calls at critical moments. They then quantitatively showed this in baseball through the use of pitch tracking data. For soccer, hand-labeled event statistics such as the amount of injury time, number of yellow cards and number of penalties awarded to reinforce their hypothesis. As soccer is a very tactical game, we hypothesize that the strategy of the home and away teams also plays a role in explaining the home advantage.

A great case study of home advantage is the 2010-2011 English Premier League soccer season. In that season, the home team earned an average 1.71 points out of a total 3 points per match. This is in stark contrast with the away team, which earned only 1.00 points per game: a rather large discrepancy, especially considering that teams play every other team both at home and away so that any talent disparities apply to both home and away averages. In terms of shooting and passing proficiency, there was no significant difference between teams at home and away (10.01% vs 9.05% for shooting and 73.46% vs 72.99% for passing - see the bottom row in Table 1).

However, there is a large difference between the amount of shots (16.2 vs 13.0) and goals scored (1.62 vs 1.17) at home and away. An illuminating example is the league champions for that season, Manchester United (see top row in Table 1). At home, they were unbeaten (winning 18 and drawing 1), but away from home they only won 5 games, drew 10 and lost 4. The telling statistic is that at home they scored 49 goals from 347 shots, compared to only 29 goals from 272 shots away from home. Comparatively, the opposition at home games only scored a paltry 12 goals from 191 shots while at away games they scored 25 goals from 271 shots. In soccer, there is the commonly held belief that team should aim to *win their home games and draw their away ones.* If you skim down Table 1, you will find that: i) all teams won more home games (except for Blackpool who won the same amount), ii) all teams score more goals at home (except Arsenal and Blackburn), iii) all teams had more shots at home compared to away, and iv) all teams gained more points at home. These event statistics tell us *what* has occurred, in the rest of the paper we use spatiotemporal data to help explain *where* and *why* this occurred. Before we detail the method, we first describe the data.

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 army 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 ac-

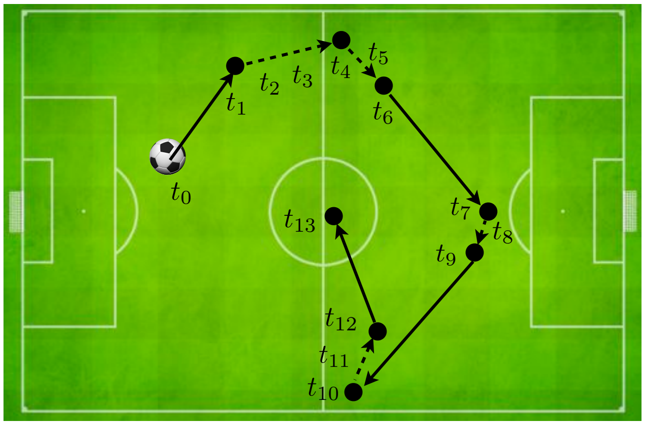


Figure 2: (From the XML feed, we can infer the ball position and possession at every time step (solid lines and dots are annotated, dotted lines are inferred).

tion. Each event has a series of qualifiers describing it. This type of data is currently used for the real-time online visualizations of events, as well as post-analysis for prominent television 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 ranking 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 opposition 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 × 100, with all positions given for teams attacking left to right.

4. OCCUPANCY MAPS

Given we have ball tracking and team possession information, we can partition the ball tracking data into possession strings (i.e. continuous movement of the ball for a single team without turnover or stoppage), where $\mathcal{O}^A = \{\mathbf{o}_0^A, \ldots, \mathbf{o}_{I-1}^A\}$ and $\mathcal{O}^B = \{\mathbf{o}_0^B, \ldots, \mathbf{o}_{J-1}^B\}$ refer to the possession 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 $D = l \times w$ 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, $\mathbf{o}$, into a set of play-segments $\mathcal{S} = \{\mathbf{s}_0, \ldots, \mathbf{s}_{N-1}\}$ of equal length, where N is the total number of play-segments for a possession, and M is the total number of play-segments for a team over a match. Given that a team's possession string is of length T_1, the resulting number of play-segments for each possession is therefore: $N = (T_1 - T) + 1$. If the possession string is smaller in duration than T, we discard it. To represent each play-segment $\mathbf{s} = \{s_0, \ldots, s_{T-1}\}$, the quantized ball position is tabulated at each time step. An example of the description is given in Figure 3.

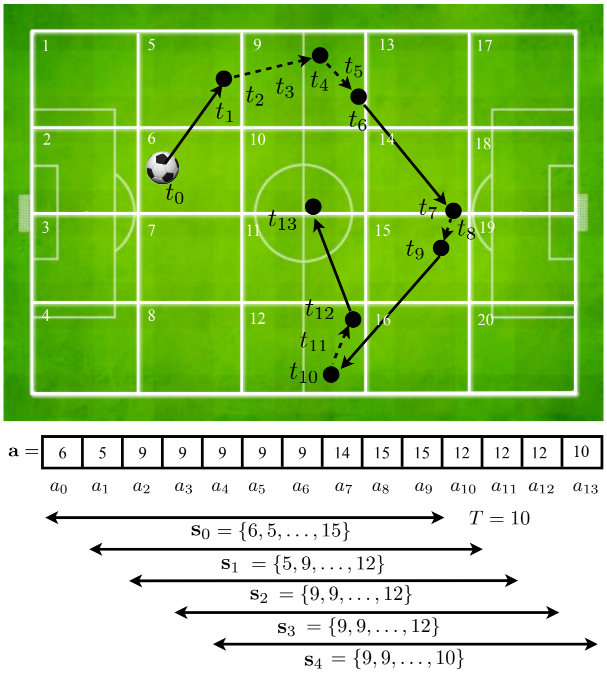

1
5
9
13
17
2
6
10
14
18
3
7
11
15
19
4
8
12
16
20
t_0
t_1
t_2
t_3
t_4
t_5
t_6
t_7
t_8
t_9
t_{10}
t_{11}
t_{12}
t_{13}
$\mathbf{a} =$ 6 5 9 9 9 9 9 14 15 15 12 12 12 10
a_0 a_1 a_2 a_3 a_4 a_5 a_6 a_7 a_8 a_9 a_{10} a_{11} a_{12} a_{13}
$T = 10$
$\mathbf{s}_0 = \{6, 5, \ldots, 15\}$
$\mathbf{s}_1 = \{5, 9, \ldots, 12\}$
$\mathbf{s}_2 = \{9, 9, \ldots, 12\}$
$\mathbf{s}_3 = \{9, 9, \ldots, 12\}$
$\mathbf{s}_4 = \{9, 9, \ldots, 10\}$

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.

From our set of play-segments for a given team, $\mathcal{S}_A$ or $\mathcal{S}_B$, we can build a distribution to characterize the expected behavior in each location. We do this as follows: Given the observations of a team, we determine the subset of play-segments $\mathcal{S}_d$, which originated in quantized area d. From the play-segment vectors within $\mathcal{S}_d$, 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^2), we would prefer the dimensionality 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 $\mathbf{p}(\mathcal{S}_d)$, which will yield an occupancy or team behavioral map. Vectorizing the occupancy map, gives us our D-dimensional spatiotemporal feature vector $\mathbf{x}$. The mean occupancy maps using entropy to describe each area is shown in Figure 4 and can give a indication of redundant patterns. As can be seen the top teams have higher entropy over most of the field compared to the lower teams - this gives an indication that these teams utilize 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

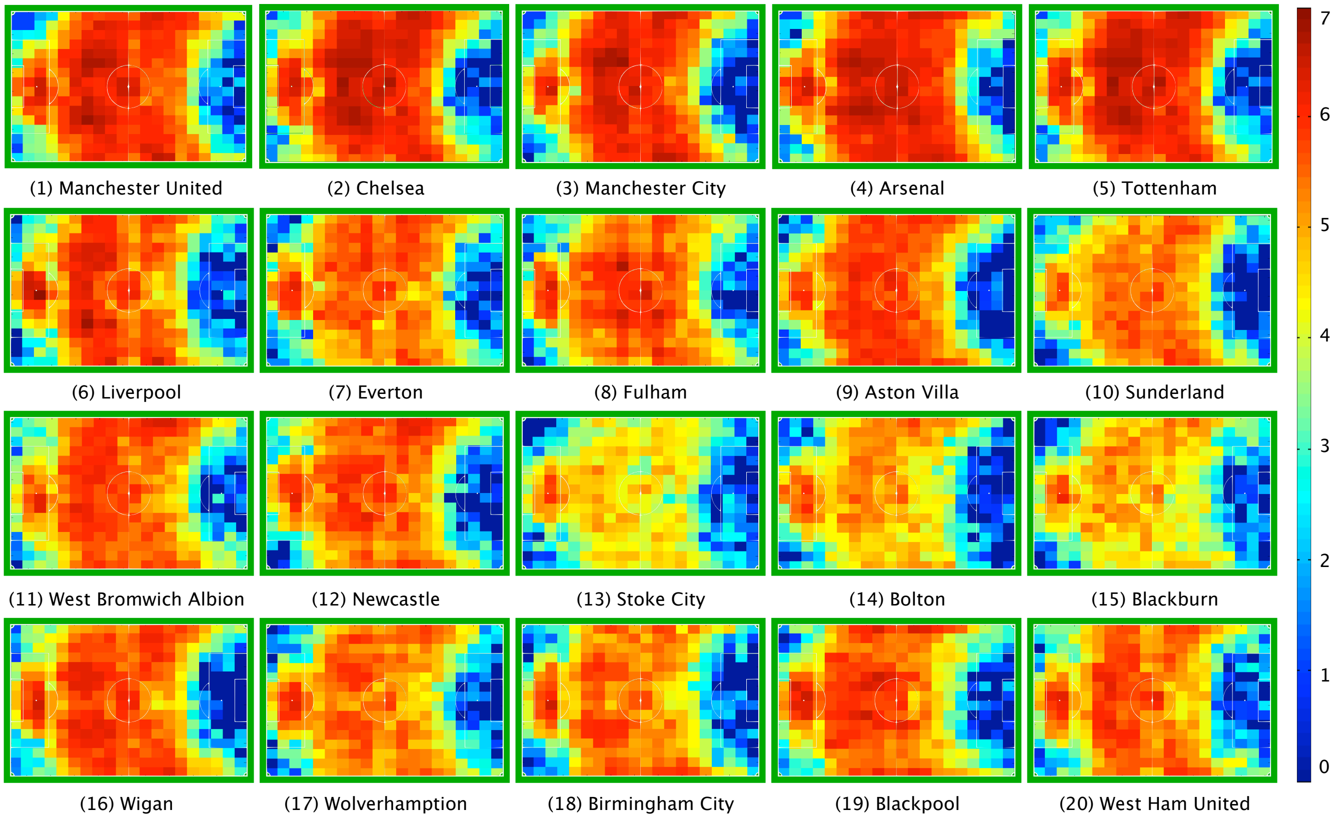


Figure 4: The mean entropy maps for each of the twenty English Premier League teams characterizing their ball movement patterns. The maps have been normalized for teams attacking from left to right. The bright red refers to high entropy scores (i.e. high variability) and the blue areas refer to low entropy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

counts, means the ball is moving quicker), we used the total count of entires as the entropy measure normalizes this information.

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 *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-match-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.

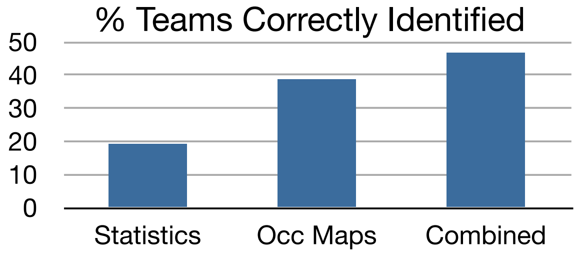
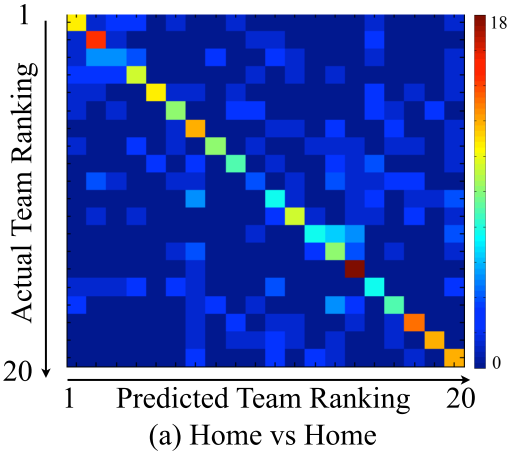


Figure 5: The identification rate for correctly identifying home performances using event-labeled data (i.e. no location information), occupancy maps (i.e. no event information, just location), and a combination of the two.

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 $D = 10 \times 8$



(a) Home vs Home

Figure 6: Confusion matrices of the team identification experiments using the combined representation.

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 as classes ($C = 20$), which resulted in a $C - 1 = 19$ dimensional input feature, $\mathbf{y} = \mathbf{W}^T\mathbf{x}$, where $\mathbf{W}$ can be found via solving

$$\arg\max_{\mathbf{W}} \mathrm{Tr}(\frac{\mathbf{W}\Sigma_b\mathbf{W}}{\mathbf{W}\Sigma_w\mathbf{W}}) \tag{1}$$

where Σ_b and Σ_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 are be done next). Results for these experiments are shown in Figure 5. From this figure, it can be seen using spatiotemporal data greatly improves the discriminability between different teams (19.26% vs 38.79%), and fusing the two together boosts performances again (46.70%). This result makes sense as the location of where teams play in addition to what they do should characterize their behavior. 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 look similar.

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. The difference maps for all twenty teams is given in Figure 7 and it makes for compelling viewing. 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, spatially, nearly all the teams (18 out of 20) had more possession in the attacking half and probably 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).

Exp	Event-Labeled	Occupancy Maps
H v H	19.26	38.79
H v A	16.09	30.08
A v A	13.98	36.41
A v H	16.36	30.34

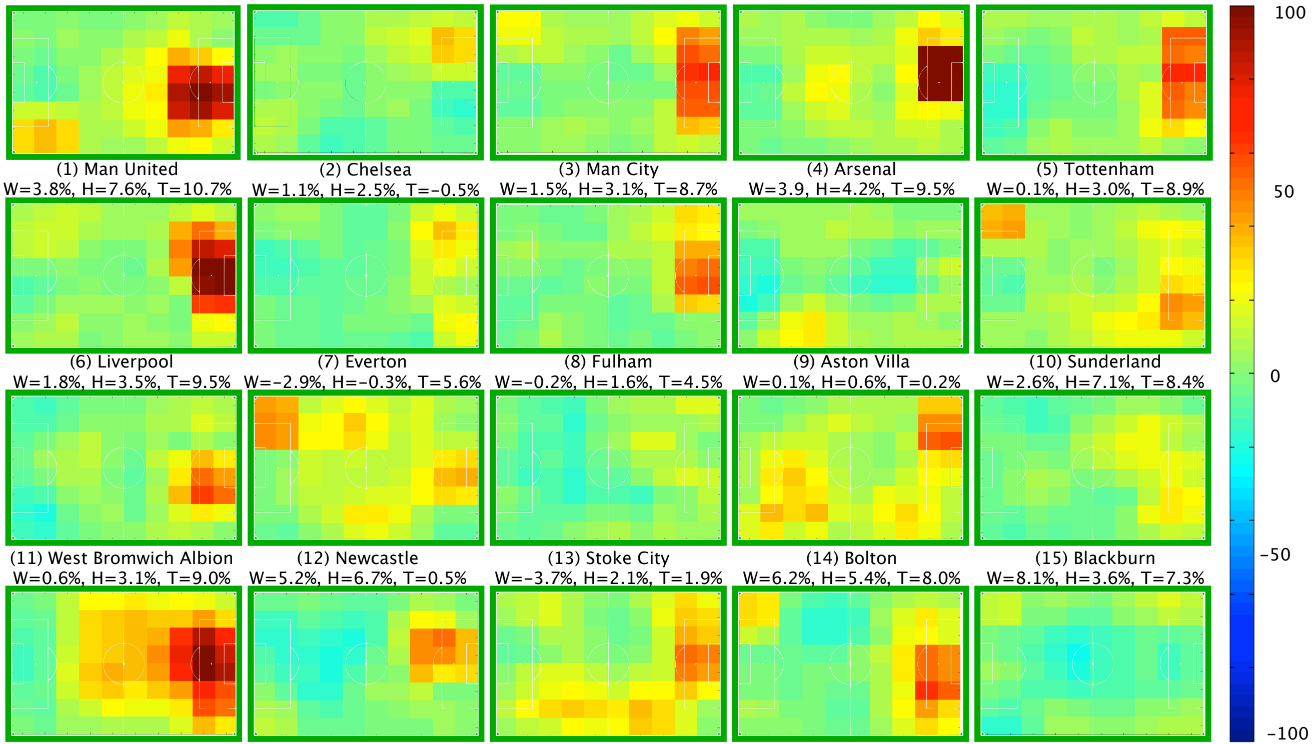
Table 2: The hit rate accuracy of experiments which tested home (H) and away (A) models against home and away matches (e.g. HvH refers to home model tested on home matches and HvA refers to the home model being tested on away matches).

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 (i.e. winning, losing, red-cards, key injury, derby matches etc.). However, through the use of spatiotemporal data, we can provide evidence of behavioral differences which can aid in the analysis of performance and decision making. This approach can also be used to flag and predict individual team performances, and in the next section we show methods in which these can be applied.

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 an 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, $\mathbf{y}$, and the mean, $\hat{\mathbf{y}}$, can be calculated where the mean is

$$\hat{\mathbf{y}} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{y}_i \tag{2}$$



(1) Man United W=3.8%, H=7.6%, T=10.7%

(2) Chelsea W=1.1%, H=2.5%, T=−0.5%

(3) Man City W=1.5%, H=3.1%, T=8.7%

(4) Arsenal W=3.9, H=4.2%, T=9.5%

(5) Tottenham W=0.1%, H=3.0%, T=8.9%

(6) Liverpool W=1.8%, H=3.5%, T=9.5%

(7) Everton W=−2.9%, H=−0.3%, T=5.6%

(8) Fulham W=−0.2%, H=1.6%, T=4.5%

(9) Aston Villa W=0.1%, H=0.6%, T=0.2%

(10) Sunderland W=2.6%, H=7.1%, T=8.4%

(11) West Bromwich Albion W=0.6%, H=3.1%, T=9.0%

(12) Newcastle W=5.2%, H=6.7%, T=0.5%

(13) Stoke City W=−3.7%, H=2.1%, T=1.9%

(14) Bolton W=6.2%, H=5.4%, T=8.0%

(15) Blackburn W=8.1%, H=3.6%, T=7.3%

(16) Wigan W=8.1%, H=12.4%, T=8.1%

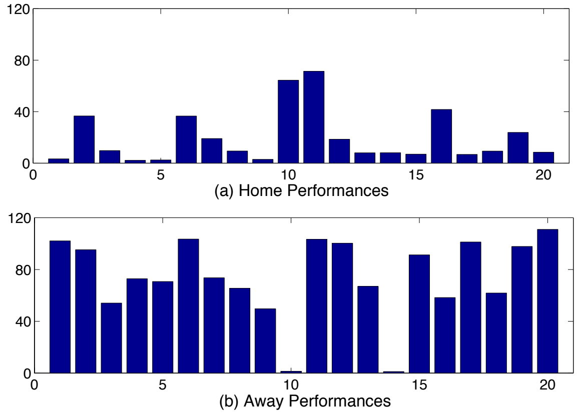
(17) Wolverhampton W=−4.3%, H=0.4%, T=4.1%

(18) Birmingham City W=5.0%, H=8.3%, T=13.8%

(19) Blackpool W=−1.0, H=3.4%, T=11.9%

(20) West Ham United W=−2.2%, H=−0.7%, T=3.6%

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.)



(a) Home Performances

(b) Away Performances

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

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 (as is the case in our work), which generates the distance measure via

$$\text{dist}_m = \|\mathbf{y}_m - \hat{\mathbf{y}}\|_2 \tag{3}$$

where m refers to the game of interest. Returning to our home and away performance example, we can show the variation in performance for each team in the EPL by finding the variance in distortion (Figure 9). As can be seen in this figure, each team's home performance has quite a low variance which gives an indication that when team's play at home they do not vary their approach too much. Conversely, it seems that the away behavior is quite random so forecasting away performance may 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. $\pm\sigma$). 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). As can be seen by comparing both occupancy maps, in their match against Manchester United they occupied a lot more possession in the middle of the field then normal. This highlights the importance of context, as after scoring two early goals Manchester United sat back and allowed Fulham to have the majority of possession in non-threatening regions (52% of overall possession) [6]. To counter this, we would have to normalize for 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.

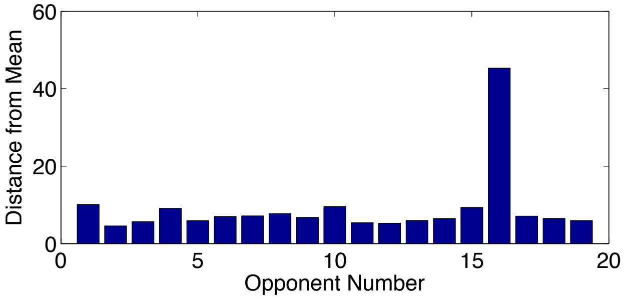


Figure 9: Example of the distortion for each away performance of Fulham in the 2010-2011 season. In match 16 they played Manchester United, and the performance on the day was far different from their other away performances (they lost 2-0 on this occasion).

7. SUMMARY AND FUTURE WORK

Most sports analytics approaches still only use event-labeled 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 for the fact that it is continuous and is extremely difficult to segment into categories which would enable high-level analysis (e.g. team strategy labels). The emerging field of *sports spatiotemporal analytics* attempts to use 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 proficiencies 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 our future work, we will look at unsupervised methods which cluster playing similar playing styles which can enrich our training data set, without effecting its discriminating power. Additionally, we are looking to extend this approach to focus on using player tracking information to discover team formations and plays. Predicting team interactions and subsequent performances and outcomes, especially when they have not played each other is another area focus of our research. As reliable high-level labels are almost impossible to obtain, predicting match outcomes as our evaluation tool seems to be the best indicator of improved team modeling.

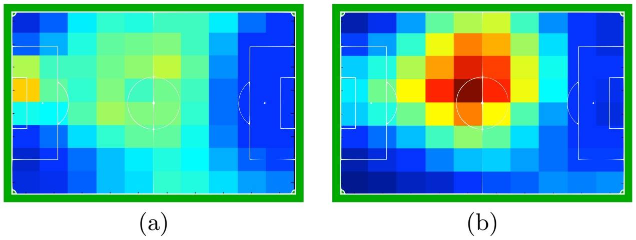


Figure 10: Occupancy maps of the: (top) mean away performance for Fulham, and (bottom) their performance against Manchester United - in this match they lost 2-0 and conceded early in the match.

8. REFERENCES

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

[2] N. Allen, J. Templon, P. McNally, L. Birnbaum, and K. Hammond. StatsMonkey: A Data-Driven Sports Narrative Writer. In *AAAI Fall Symposium Series*, 2010.

[3] BBC-Sports. Footballers may trial wearing microchips to monitor health. `www.bbc.co.uk/sport/0/football/21460038`, 14 Feb 2013.

[4] 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.

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

[6] ESPNFC. `http://espnfc.com/us/en/report/292849/report.html?soccernet=true&cc=5901`.

[7] K. Goldsberry. CourtVision: New Visual and Spatial Analytics for the NBA. In *MIT Sloan Sports Analytics Conference*, 2012.

[8] 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.

[9] Hawk-Eye. `www.hawkeyeinnovations.co.uk`.

[10] D. Henschen. IBM Serves New Tennis Analytics At Wimbledon. `www.informationweek.com/software/business-intelligence/ibm-serves-new-tennis-analytics-at-wimbl/240002528`, 23 June 2012.

[11] 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.

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

[13] 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.

[14] M. Lewis. *Moneyball: The Art of Winning an Unfair Game*. Norton, 2003.

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

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

[17] W. Lu, J. Ting, K. Murphy, and J. Little. Identifying Players in Broadcast Sports Videos using Conditional Random Fields. In *CVPR*, 2011.

[18] 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.

[19] L. Madden. NFL to Follow Army's Lead on Helmet Sensors in Attempt to Prevent Head Injury. `www.forbes.com/sites/lancemadden/2012/07/16/nfl-to-follow-armys-lead-on-helmet-sensors-in/-attempt-to-prevent-head-injury/`, 16 July 2012.

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

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

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

[23] NBA Shot Charts. `www.nba.com/hotspots`.

[24] D. Oliver. *Basketball on Paper: Rules and Tools for Performance Analysis*. Brassey's, Incorporated, 2004.

[25] D. Oliver. Guide to the Total Quarterback Rating. `espn.go.com/nfl/story/_/id/6833215/explaining-statistics-total-quarterback-rating`, 4 August 2011.

[26] Opta Sports. `www.optasports.com`.

[27] 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.

[28] 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.

[29] Prozone. `www.prozonesports.com`.

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

[31] SportsVision. `www.sportsvision.com`.

[32] STATS SportsVU. `www.sportvu.com`.

[33] Statsheet. `www.statsheet.com`.

[34] 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.

[35] 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.

[36] 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.

[37] Zonalmarking. `www.zonalmarking.net`.