Using Machine Learning to Draw Inferences from Pass Location Data in Soccer

Joel Brooks*, Matthew Kerr and John Guttag

Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

received 1 July 2015; revised 15 January 2016; accepted 11 May 2016
DOI:10.1002/sam.11318
Published online in Wiley Online Library (wileyonlinelibrary.com).

Abstract: In this paper, we present two approaches to analyzing pass event data to uncover sometimes-nonobvious insights into the game of soccer. We illustrate the utility of our methods by applying them to data from the 2012–2013 La Liga season. We first show that teams are characterized by where on the pitch they attempt passes, and can be identified by their passing styles. Using heatmaps of pass locations as features, we achieved a mean accuracy of 87% in a 20-team classification task. We also investigated using pass locations over the course of a possession to predict shots. For this task, we achieved an area under the receiver operating characteristic (AUROC) of 0.785. Finally, we used the weights of the predictive model to rank players by the value of their passes. Shockingly, Cristiano Ronaldo and Lionel Messi topped the rankings. © 2016 Wiley Periodicals, Inc. Statistical Analysis and Data Mining: The ASA Data Science Journal, 2016

Keywords: machine learning; sports analytics; soccer analytics

1. INTRODUCTION

Although soccer is by far the world's most popular sport [1], published work in soccer analytics has yet to achieve the same level of sophistication as analytics being performed in other professional sports. There has been work on developing new summary statistics to measure and confirm traditional beliefs about the game [2]. But such analyses do not fully leverage the newly available rich datasets in soccer, such as the dataset of ball-events we use in this paper. In this paper, we describe and evaluate two applications of machine learning to build predictive models from these kinds of data.

One of the key problems in building predictive models from soccer data is the sparseness of positive outcomes. In high-scoring sports such as basketball, there are unambiguous and frequent positive outcomes (e.g. a basket) to relate to sequences of events through statistical models. In soccer, goals are so rare that such direct associations are hard or impossible to establish. A team may completely dominate statistical measures such as possessions in the offensive area or even shots, and fail to score even one goal. This weak correlation implies that more data is needed to build models than is typically needed for other sports.

* *Correspondence to:* Joel Brooks (brooksjd@mit.edu)

Despite this problem, we demonstrate in this paper that by using machine learning techniques on passing data from the 2012–2013 La Liga season, we could uncover relevant data-driven insights into soccer.

1. We show that heatmaps built using only the origins of passes provide fingerprints that can be used to identify teams with 87% accuracy.

2. We further show that even when we only consider passes originating from the midfield, the resulting heatmaps can still be used to identify teams.

3. We construct a model relating pass origins and destinations during a possession with the probability of a shot. The resulting weights offer insights into the offensive utility of passes.

4. We utilize this model to rank players by the frequency with which their passes are highly valued by the model.

The rest of the paper is organized as follows. In Section 2, we outline some previous related work on using machine learning for knowledge discovery in soccer and other sports. In Section 3, we describe the event-based dataset we used

in the reported work. We describe our experiments for classifying teams by their passes in Section 4. In Section 5, we present our work on predicting possessions ending in shots. Finally, in Section 6 we summarize the overall contributions of this paper and discuss possible future work.

2. RELATED WORK

Much of the published research in sports analytics, especially research that utilizes spatiotemporal data, has focused on sports that are easily discretized, such as baseball, American football, and tennis [3]. These sports are easily broken up into individual events (at bats, plays, or points) that have obvious immediate outcomes, such as hits, yards gained, or a point won. For example, Intille and Bobick used player tracking data to recognize different plays in American football [4]. It is more difficult to perform similar work for sports that are not as easily discretized, such as basketball and soccer, because the continuous nature of play makes the connections to outcomes less obvious. In both [5] and [6], the authors were able to utilize player tracking data for basketball to classify offensive plays and the movement patterns of offensive players. Similar work in soccer has proven to be more difficult because it is not obvious how to break up gameplay to understand strategy.

Soccer analytics has focused on building probabilistic models to simulate game actions and predict the outcomes of matches. Reep and Benjamin developed models for the success rates of different length passing sequences [7]. This work was limited by the lack of access to data about the location of passes. More recently, both [8] and [9] predicted the outcome of matches by using possession rates of different teams and other historical statistics to develop probabilistic models. In contrast, our work focuses on developing insights into styles of play rather than on predicting outcomes of games, extending previous work done by Kerr [10]. In [11], the authors investigated the frequency of passing that occurs immediately before and after a goal has been scored. They found that in the 5 minutes preceding a goal, the team that scores makes a significantly greater frequency of successful passes than their average for the half, whereas the conceding team played significantly fewer successful passes.

As the amount of spatiotemporal soccer data has increased, there has been more work that leverages the information the datasets provide. Spatiotemporal data allows analysts to study the underlying mechanics of the game, such as style or player movements. Bloomfield *et al.* used player tracking information to investigate the physical demands of soccer and the work rates of different players [12]. In [13], the authors leveraged ball-event data and passing sequences to cluster the playing styles of different teams but not, as we do, to classify teams. They described a passing sequence by the number of unique players involved in the sequence, and observed that different teams use different sequences at different rates. Lucey *et al.* used ball-event data to infer the location of the ball throughout a game. Using this information, they constructed 'entropy-maps' to characterize how different teams move the ball during a match [14]. Using entropy-maps and in-game statistics, they were able to classify teams with 47% accuracy. In more recent work, the authors combine match statistics, event data, and player tracking data to identify the teams playing in a given game with 70% accuracy in [15]. Since their data and task formulation differ from ours, no direct comparison with our results is possible.

3. THE DATA

The dataset we use throughout all of the experiments presented in this paper was collected by Opta Sports [16]. The data are hand-labeled annotations of each ball-event that took place during the course of a match, for example, each pass, tackle, shot, etc. A ball-event is recorded any time a player makes a play on the ball, apart from dribbling. The dataset also includes additional information for each ball-event such as the location, the player involved, and the outcome. We refer to these as 'descriptors'. Some descriptors, such as location, time of event, and player involved, are common for each ball-event type. Other descriptors, such as pass length, exist only for specific kinds of ball-events. We assume that the data are clean and accurate, since it was collected by a professional enterprise dedicated to collecting sports data. In this paper, we use data collected from the 2012−2013 La Liga season. La Liga is the premier league in Spain and is comprised of 20 teams.

3.1. Passes

We utilize the following information recorded about each pass in the dataset:

1. In which game the pass was attempted
2. The time of the pass
3. Which team attempted the pass
4. Which player attempted the pass
5. The intended recipient of the pass
6. The origin of the pass

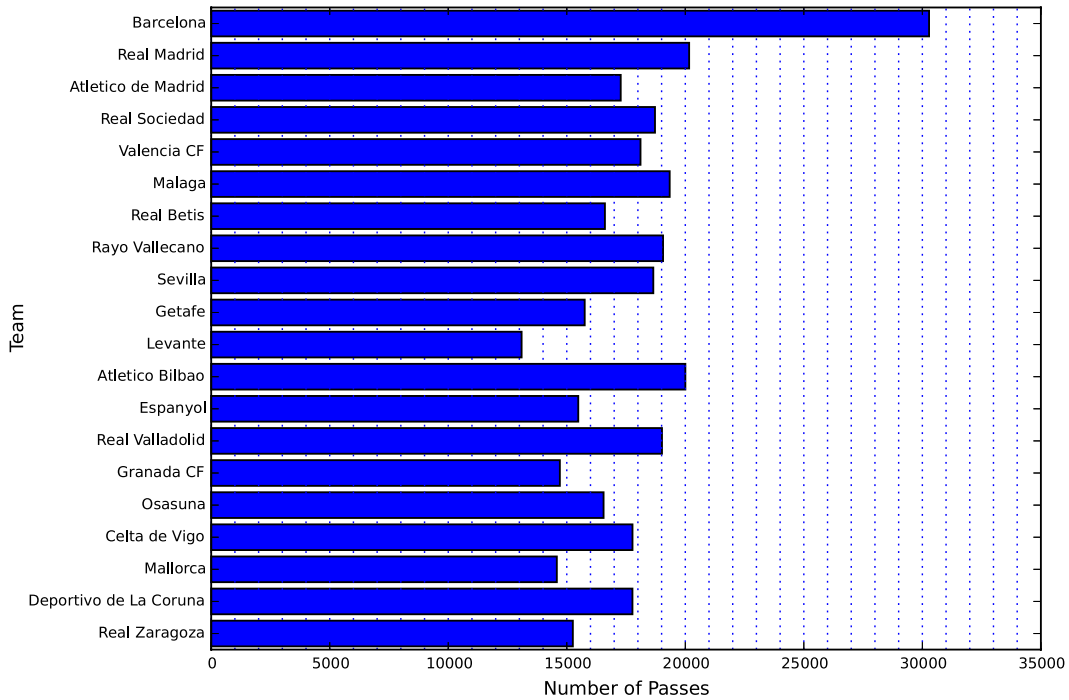

Fig. 1 Total number of passes attempted by each team during the 2012–2013 La Liga season. The teams are ordered from top to bottom by their final league standing. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

7. The destination of the pass

8. The outcome of the pass (was it successful or not).

On average, each team in La Liga attempted ∼18,000 passes during the entire season, with Barcelona making the most passes by a wide margin at 30,283, and Levante attempting the smallest number of passes at 13,094. Figure 1 shows the number of passes each team attempted during the season. The teams are also ordered from top to bottom by their final position in the league (Barcelona finished first).

3.2. Pitch Discretization

For all the work we describe in this paper, we focused on the locations of passes because we hypothesized that pass location is a strong indicator of team strategy and personnel. To generalize between passes that are near each other but not in the same location, we discretized the pitch into 18 zones, as shown in Fig. 2. This representation has previously been shown to identify critical zones on the pitch associated with offensive outcomes such as shots and goals [17].

4. CHARACTERISTIC PASSING STYLES

In this section, we present our investigation of whether teams have a characteristic passing style. We first visualize the similarity of the passing styles of different teams. We then solve the following classification task: given a random sample of passes made by a single team, identify which team attempted those passes.

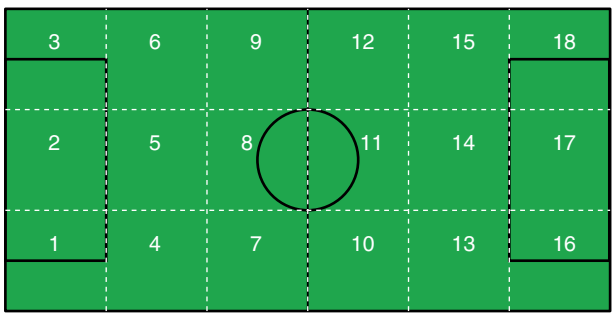

Fig. 2 Playing area split up into 18 zones. The left side of the pitch (zones 1–3) is the defensive side of the pitch, and the right side (zones 16–18) is the offensive area. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

4.1. Pass Heatmaps

For a given set of attempted passes, we counted how many passes originated from each zone in our discretized pitch and normalized by the total number of passes to produce a heatmap of the origins of passes. The heatmap can be represented by a 3×6 matrix of frequency values.

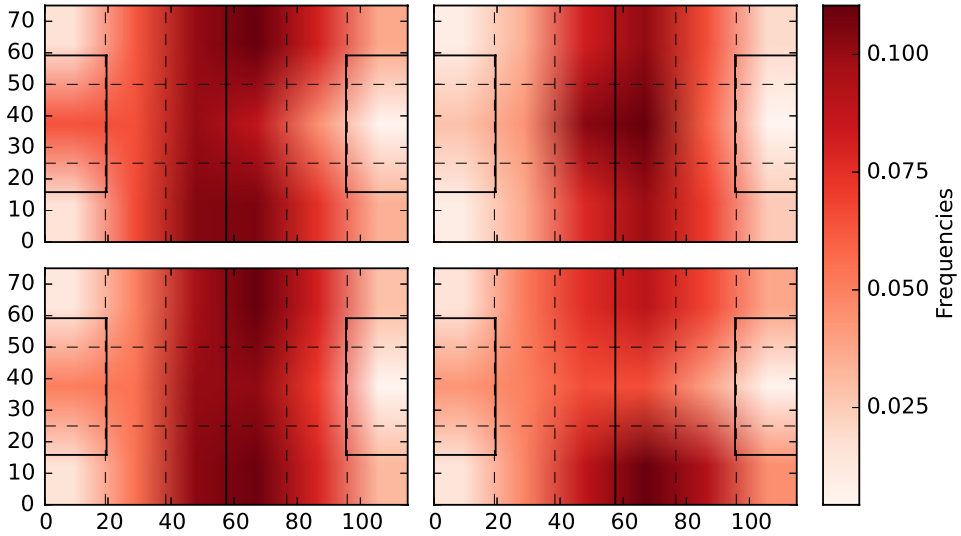

Fig. 3 Each figure is a 3×6 heatmap showing the locations of the origins of a set of passes. Starting from the top left and moving in a clockwise direction, the heatmaps represent (a) the set of all of the passes attempted during the 2012–2013 La Liga Season by every team, (b) the set of all of the passes attempted by Barcelona, (c) the set of all of the passes attempted by Real Madrid, and (d) the set of all of the passes attempted by Atletico Madrid. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

We only included the origins of passes, since we believed that this was the best representation of a team's intended style. We did not include the destinations of passes since passes could be intercepted or badly executed, which would not be representative of a team's intended style. We anticipate that future work will include the analysis of pass destinations; indeed, the second experiment highlighted in this paper includes pass destination.

In Fig. 3, we have plotted several heatmaps. Each heatmap plots the origins of all the passes that were attempted during the 2012–2013 La Liga season either by every team in the league combined or by an individual team.

We use the heatmaps as the basis for the visualization and classification experiments we present in this section.

4.2. Visualizing Style Similarity

In Fig. 4, we present a visualization of the similarity of the heatmaps for different teams to investigate how their passing styles are related.

To investigate the variance in passing style for a single team across a season, we first split the set of passes attempted by a team into smaller subsets. For each team in the dataset, we randomly assign each pass attempted during the season to one of 10 smaller subsets, creating 10 randomly constructed subsets of passes per team. Each subset is then used to construct a heatmap, i.e. there are 10 heatmaps per team.

We then calculate the average distance between the different heatmaps of all pairs of teams to construct a distance matrix M. In more detail, each i, j entry, where $i \neq j$, is equal to the average distance between all the different heatmaps of team i and team j:

$$M_{i,j} = \frac{1}{|H_i||H_j|} \sum_{(h_x,h_y)\in(H_i,H_j)} \|h_x - h_y\|$$

where H_i is the set of all the heatmaps constructed for a team i. Recall that we partitioned the set of all passes attempted during a season for a single team into 10 random subsets, i.e. $|H_i| = 10, \forall i$. Each i, j entry, where $i = j$, is the average distance between each of the 10 heatmaps constructed for a single team i. The entries of the final distance matrix are averages of 10 repetitions of splitting the set of passes into random subsets and calculating a new M each time.

After creating the distance matrix M, we plot the visualization that is presented in Fig. 4. It is a 20×20 distance heatmap that plots the inverse of the entries of M, i.e. a darker color indicates a smaller distance between the heatmaps of two teams. The distances are also scaled to a range $[0, 1]$ using the min and max entries of M. The teams are ordered by their final position in the league. Because $M_{i,j} = M_{j,i}$, the heatmap is symmetrical.

In Fig. 4, we observe that the diagonal is of a darker shade than the rest of the heatmap. This is not surprising; we expect the distances between heatmaps of the same team to be relatively small. This suggests that dividing the set of all passes for a single team over the course of a season into 10 random subsets results in heatmaps that are similar and are characteristic of a team. The heatmap also indicates

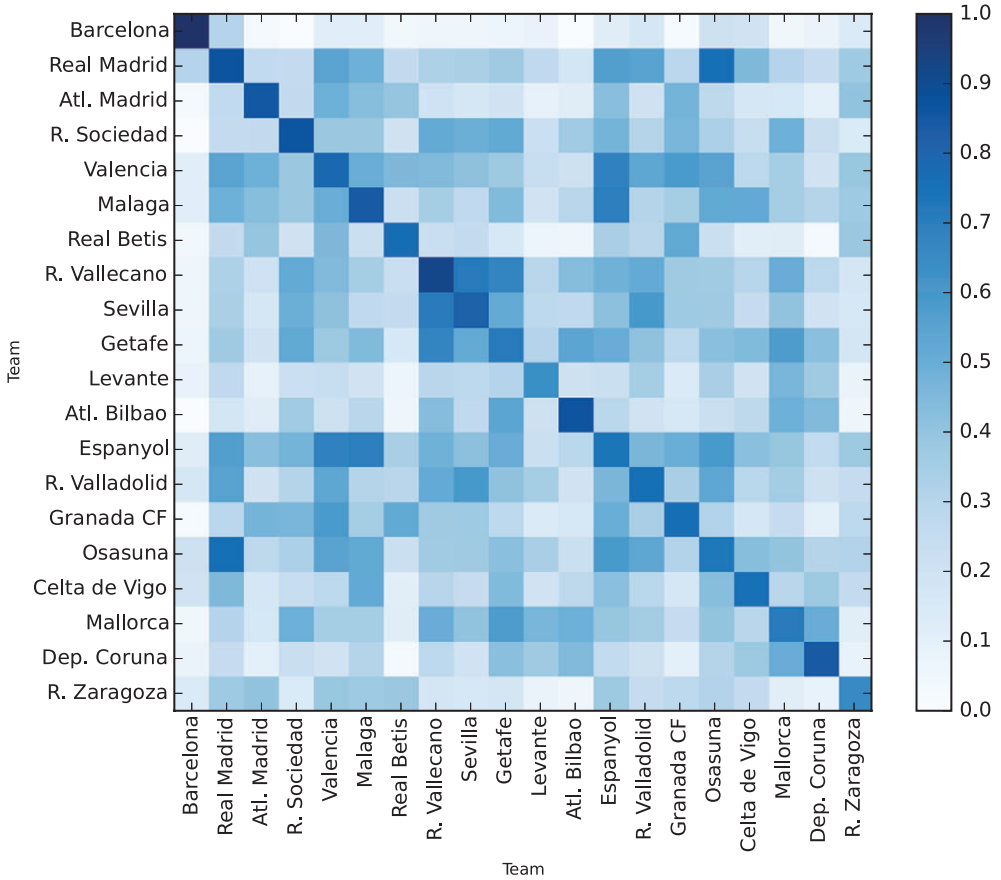

Fig. 4 Visualization of the Euclidean distances between pass heatmaps of teams. A 20×20 distance heatmap that plots the inverse of the average distance between the heatmaps of teams. The distances are scaled to the range $[0, 1]$. A darker shade indicates a smaller distance. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

that some teams are more consistent in their style of play that others.

We also observe that the distances between Barcelona's heatmaps and those of the rest of the teams in the La Liga are relatively large. This suggests that Barcelona's passing style is the most distinctive, which is a popular comment among soccer observers. This result is further supported by the results found in the classification task (presented in Section 4.3). Barcelona attains a perfect F-score during classification, which suggests that it is very easy to classify heatmaps of Barcelona because they are distinctive. It is interesting to note that the distinctiveness of a team is not correlated with a team's success. Recall that the teams are ordered by their final position in the league. However apart from Barcelona, there is no obvious relationship between success and the heatmap.

4.3. Team Classification

To determine whether the heatmaps were representative of a team's passing style, we conducted the following classification experiment: As in the visualization task, for all the teams we first split the set of all passes attempted by a team during the season into 10 randomly constructed subsets to construct 10 heatmaps per team. We chose 10 because it offered a balance between ensuring the size of a subset was still large enough to be representative of a team's passing style and providing enough examples for classification. Each heatmap is treated as an example, and its respective label is the team that attempted the passes. To use the heatmaps as features, we flatten the 3×6 matrix into a single vector of values, creating a model that has feature dimensionality of 18.

After constructing the heatmaps, we take a 70–30 training/test split stratified by team, thus ensuring that the class balance is equal in both the training set and test set. We then construct a K-nearest neighbor (K-NN) classifier to perform the classification task. We chose to use a K-NN classifier because of the interpretability it offers, and also because it showed strong performance during preliminary experiments. To choose a value for K, we performed threefold cross-validation on the training set and selected a

Table 1. Mean precision and recall values for each team. The table is ordered by the final league positions of the teams.

Team	Mean precision	Mean recall
Barcelona	1.00 ± 0.000	1.00 ± 0.000
Real Madrid	0.996 ± 0.0310	0.996 ± 0.0378
Atl. Madrid	0.988 ± 0.0551	0.981 ± 0.0766
Real Sociedad	0.921 ± 0.135	0.944 ± 0.139
Valencia CF	0.731 ± 0.215	0.818 ± 0.223
Malaga	0.959 ± 0.0972	0.958 ± 0.123
Real Betis	0.967 ± 0.0882	0.971 ± 0.0989
Rayo Vallecano	0.709 ± 0.211	0.805 ± 0.222
Sevilla	0.937 ± 0.135	0.885 ± 0.191
Getafe	0.637 ± 0.316	0.559 ± 0.300
Levante	0.954 ± 0.117	0.868 ± 0.203
Atl. Bilbao	0.967 ± 0.0863	0.997 ± 0.0297
Espanyol	0.660 ± 0.305	0.587 ± 0.309
Real Vall.	0.975 ± 0.0777	0.992 ± 0.0505
Granada CF	0.888 ± 0.179	0.909 ± 0.234
Osasuna	0.740 ± 0.258	0.700 ± 0.276
Celta	0.956 ± 0.103	0.943 ± 0.127
Mallorca	0.871 ± 0.201	0.774 ± 0.244
Dep. Coruna	0.995 ± 0.0353	0.994 ± 0.0449
Real Zaragoza	0.968 ± 0.0958	0.878 ± 0.18

value for K from the set $\{2, 3, 4, 5, 6, 7\}$—we chose the value that resulted in the highest mean accuracy across the folds. The classifier uses a weighted voting scheme; when classifying a single example, each of the K nearest neighbors will have a weighted vote proportional to the inverse of its Euclidean distance from the example.

After choosing a value for K, we classify each example in the test set and calculate the overall accuracy, as well as the precision and recall for each class. We repeat the 70–30 training/test split 200 times, and repeat the entire experiment 10 times. The results presented in Section 4.4 are the mean values taken over all of the repetitions.

One of the alternatives to splitting the set of passes randomly would have been to split the dataset contiguously, i.e. create a subset of passes for each single game, or for a collection of sequential games. However, in this experiment we wished to explore the typical style of a team; we believed that a team's characteristic playing style is more identifiable when their play is looked at across an entire season. Teams will often change their tactics depending upon whom they are playing and the score of the game, introducing more variation. We attempted to smooth this variation by randomly sampling across the entire season when constructing each subset of passes.

4.4. Classification Results

After repeating the experiment 2000 times as described, we obtained a mean accuracy of 0.873 on the test set, with a standard deviation of 0.0384. The mean value of K for each of the 2000 classifiers was 5.78, with a standard deviation of 1.3. We calculated accuracy as the percentage of all the examples that we correctly classified. We also calculated the mean precision and recall values for each team, and present them in Table 1. The precision of a given class C is calculated as: $p = \frac{t}{t+f}$, where t is the number of true positives, and f is the number of false positives. The recall of a class is calculated as $r = \frac{t}{t+n}$, where t is still the number of true positives and n is the number of false negatives. In Fig. 5, we plot the mean F-score for each team. The F-score is the weighted average of a precision and recall for an individual label and has the range [0, 1]. A greater value suggests that a team is more distinguishable and therefore is easier to classify.

Given that there are 20 teams in La Liga, a random classifier would have an accuracy rate of 5%. Our accuracy rate of 87.3% strongly suggests that a team's passing style is highly characteristic of the team and that the heatmaps of just the pass origins are able to effectively capture the different styles. It also further suggests that a team's passing style is consistent across a season, since the heatmaps are constructed using passes from different games.

In Section 4.5, we show that even if we limit the set of passes to only those that originated in midfield, we are still able to identify teams accurately using the same features.

4.5. Midfield Passing Style

In this section we present our results from repeating the same experiment described in Section 4.1, but using only passes that originated from the six midfield zones (zones 7–12). By filtering out passes that did not occur in the midfield, we reduced the set of passes by roughly 50% (from 358,202 to 185,069).

We obtained a mean accuracy rate of 0.518 with a standard deviation of 0.0537. The mean value of K was 5.88 with a standard deviation of 1.23 across the 2000 classifiers. We present the mean F-score of each team in Fig. 6.

Although the accuracy rate is much smaller than that obtained during the initial experiment, it is still much higher than the 5% accuracy of a random classifier. This suggests that although passes in the midfield do not characterize teams to the same extent as all passes, they are still indicative of a team's style.

After comparing the results shown in Figs 5 and 6, we observed that the change in classification performance varied for different teams. For example, the F-score of Barcelona remained high in the second experiment, whereas the F-score of Real Madrid dropped enormously from 0.996 to 0.376. This is not a result of different teams making more or fewer passes in the midfield. The reduction in the number of passes for each team was fairly uniform ($\sim$50%) and the heatmaps are normalized by the number of

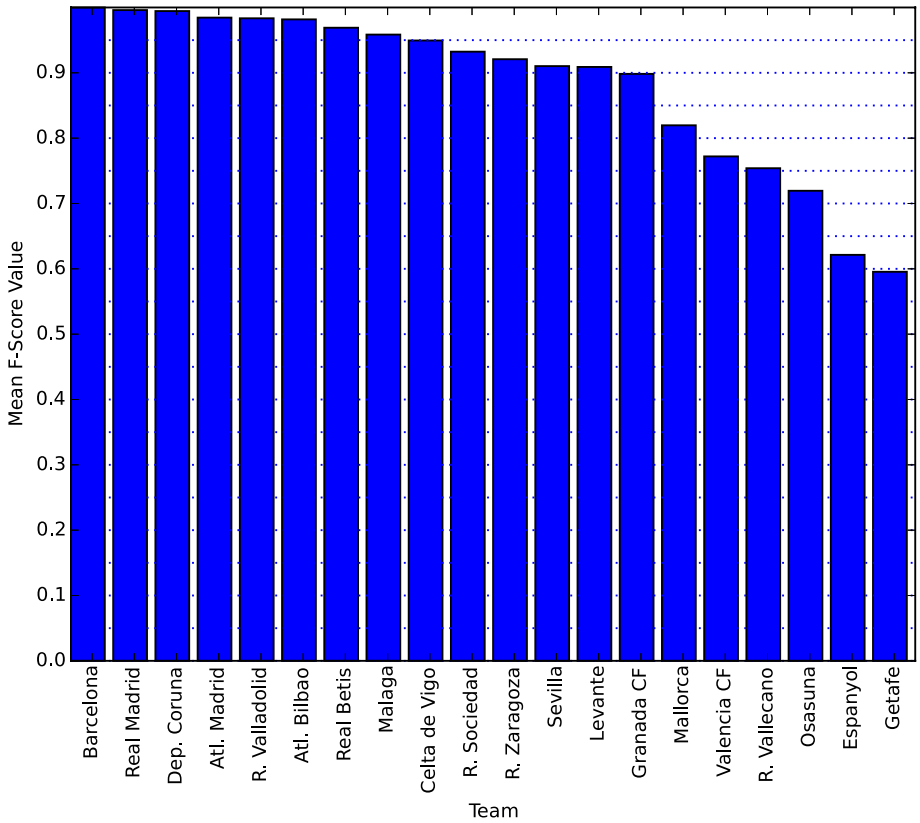

Fig. 5 **Mean F-score value for each team that participated in the 2012–2013 La Liga season.** A higher F-score value indicates better separation. The teams are ordered by their respective F-score. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

passes. The results suggest that some teams retain a strong, easily identifiable passing style in the midfield but other teams become more difficult to distinguish. A more detailed investigation of midfield passes is required to fully explore this phenomenon.

5. SHOT PREDICTION USING PASSING DATA

We now look at how passing strategy within a game relates to positive outcomes for the offensive team.

5.1. Feature Extraction

To extract the features that we used to build our predictive models, we first segment each game into a discrete sequence of observations. We chose to segment the game at the level of *possessions*. A possession in soccer is defined as a period of time that a single team retains the ball among their own players without an interruption in play or loss of the ball to the opposing team. We filtered out any possessions that did not contain at least three passes in order to remove regions of play where a team only had a few touches on the ball, since we hypothesized that these epochs of play are less likely to reveal useful strategic elements.

After segmenting each game into a discrete sequence of observations, we extracted features from each of these observations to construct feature vectors. All of the features that we utilized are based on an abstract representation of passing strategy, which we call *pass grids*.

We constructed three types of pass grids for each possession:

Origin grid: Percentage of passes in a possession that originate in a particular zone

Destination grid: Percentage of passes in a possession that have a destination in a particular zone

Origin-Destination grid: Percentage of passes in a possession from one zone to another zone

We then constructed the feature vectors by concatenating all the values from each of the three grids. The origin and destination grids have one value per zone, so each accounts for 18 features, and the origin–destination grid has a value for each origin–destination pair, and thus is $18 \times 18 = 324$

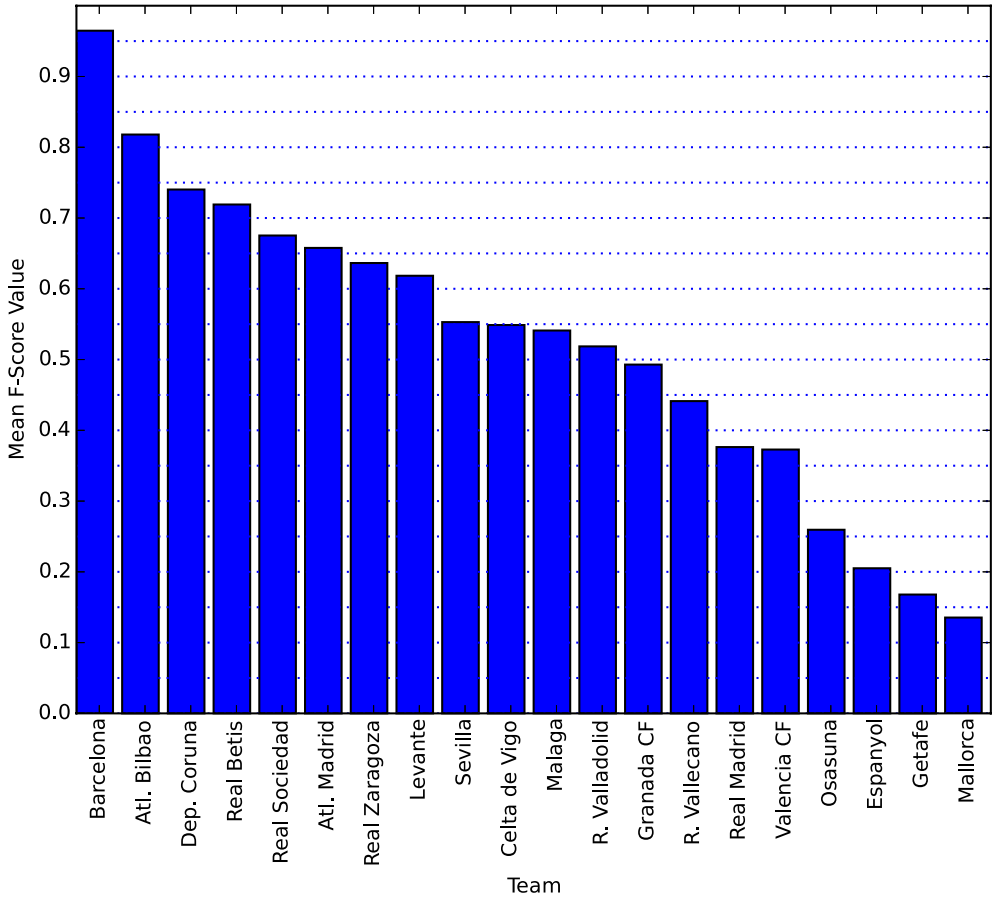

Fig. 6 **Mean F-score value for each team that participated in the 2012–2013 La Liga season, using only midfield passes.** A higher F-score indicates better separation. The teams are ordered by their respective F-score. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

features. As a result, each possession was converted into a feature vector of length 360.

Each feature vector is labeled according to how the possession ended. Possessions that ended in a shot taken by the offensive team were assigned a label of 1, and all others were assigned a label of -1. We chose shots instead of other positive offensive outcomes such as goals because they are less sparse and less influenced by factors such as luck and the skill of the goalkeeper.

5.2. Method, Experimental Design, and Testing

Upon converting each possession in a game to a fixed-length feature vector, we then used these feature vectors to train models to relate passing strategy in a possession to shots taken. We first split the data by using the first 80% of games chronologically as the training set, and setting aside the final 20% as the holdout set. We did this split to simulate the scenario of applying our models to unseen data.

Using the training set, we trained an L2-regularized support vector machine (SVM) model using the LIBLINEAR package [18]. We used class-specific cost parameters in order to account for the extreme class imbalance between positive and negative examples in the training set. We utilized a two-dimensional grid search and fivefold cross-validation to find the optimal class-specific cost parameters on the training set. We searched for cost parameters over the range $\{10^{-6}, 10^{-5}, \ldots, 10^{2}, 10^{3}\}$. The folds were constructed at the *game* level so possessions in a single game were not split across multiple folds. We chose the cost parameters that had the maximum average area under the curve (AUC) on the five test folds, and used those parameters to train the final model. We found the optimal costs to be $\{1, 0.1\}$ for the positive and negative classes, respectively. The final model was then tested on the holdout set.

5.3. Classification Results

Our model that predicts when a possession will end in a shot has an area under the receiver operating characteristic

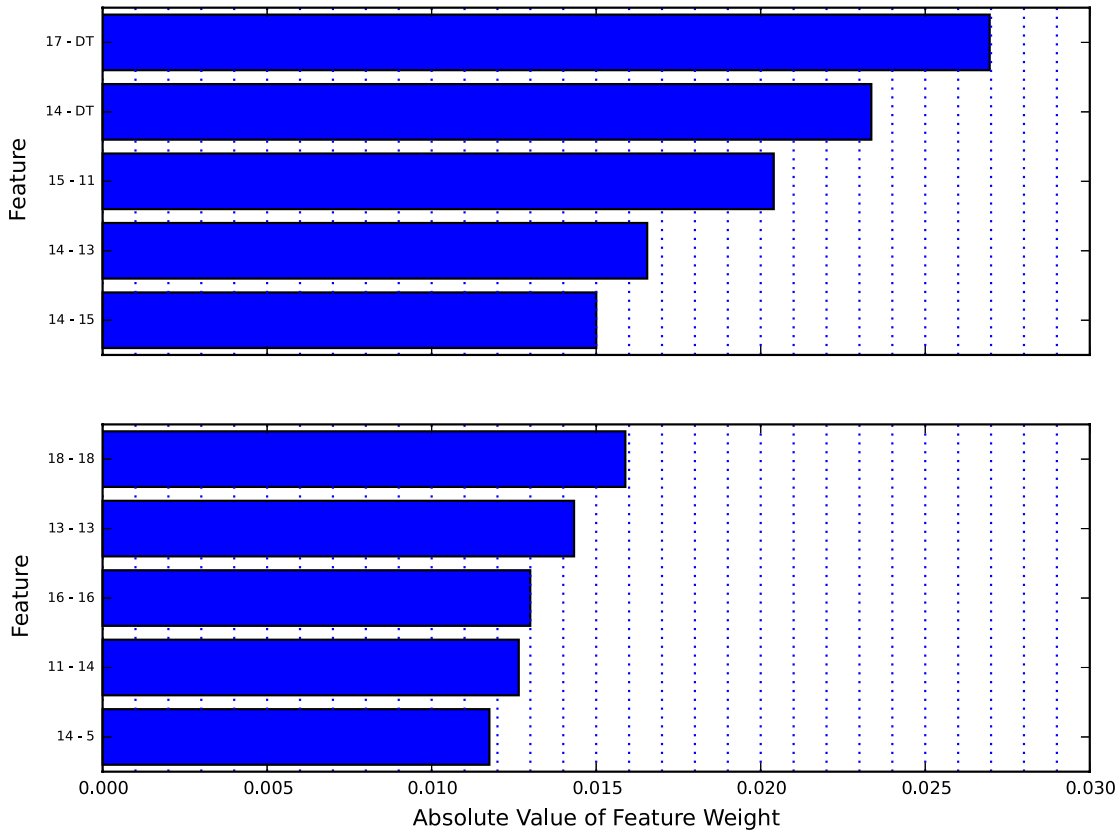

Fig. 7 Ten of the most influential features for predicting when a possession will end in a shot. The top chart includes the five features with the greatest positive weights and the bottom chart includes the five features with the most negative weights. All features are normalized by the total sum of absolute weights. Each number corresponds to a zone. 'OR' designates pass origin features, 'DT' designates pass destination features, and features labeled with two zones, 'Z_1-Z_2,' designate an origin–destination feature to Z_1 from Z_2. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

(AUROC) of 0.785 and an F-score of 0.311. We used this model to investigate the relationship between our features and shots by looking at the relative importance of each zone. We show the feature weights for the top five positively and negatively associated features in Fig. 7. The weights presented in the figure are normalized by the total sum of the absolute value of all the weights to better understand a given feature's relative importance. We can see that no single feature dominates and that there are both positive and negative features with relatively high model weights. This provides further evidence that our model captures a tradeoff between where passes are more likely or less likely to lead to shot opportunities in the future, and is not based solely on simple rules such as 'shots happen when there are passes near the opposing team's goal.'

Looking at the relative feature importance provides insight into what passing strategies generally lead to shot opportunities. The top two most important features are both pass destination features to zones 17 and 14, respectively. These are the 'critical zones' usually identified as being strongly associated with positive offensive outcomes [17]. This represents a simple relationship between shot opportunities and passes within a possession: simply get the ball into the critical zones. The fact that getting the ball into these zones leads to shots is not surprising. However, other features in this set of 10 provide less obvious insights. The next two most important features are both origin–destination pair features: a pass from zone 11 to zone 15 and a pass from zone 13 to zone 14. The first represents a pass from the center of the midfield to the left side of the attacking third. It suggests that moving the ball from the center of the pitch to the wingers on the outside is associated with an eventual shot opportunity. The next most important feature is a pass from the right side of the attacking third to zone 14, one of the 'critical zones' right in front of the opposing team's penalty box. This could be representative of passes, such as crosses, from the outside into the dangerous areas, which is one of the most identifiable offensive strategies in soccer.

Two of the three most negatively associated features are a pass from a player in zone 18 to a target also in zone 18, and a pass from zone 16 to zone 16. This could be representative

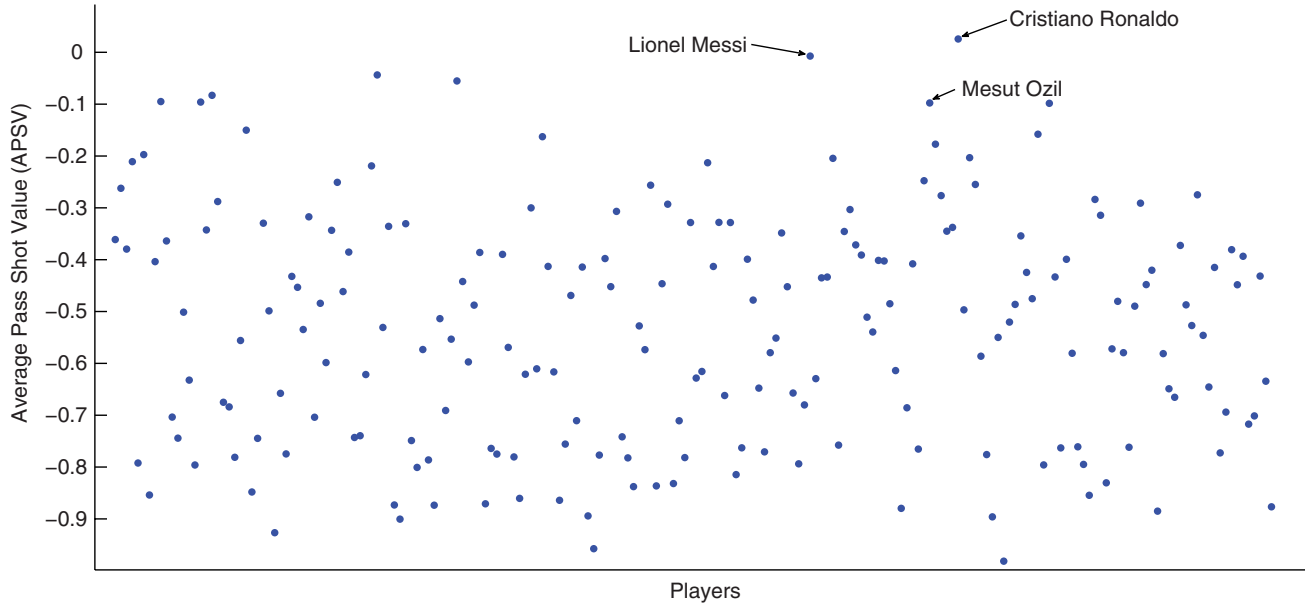

Fig. 8 Average pass shot value (APSV) for all players with more than 200 passes in the 2012–13 La Liga season. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

of a team getting trapped by the defense in the corner, unable to further advance it toward the defending team's goal. It is interesting to note that passes from zone 14 to 18, and vice versa, are both are positively associated with shot opportunities. Therefore, in general, it is advantageous to get the ball to and from the corner offensive zones, but it seems that it is detrimental to linger there.

Finally, while most of the top features involving the 'critical zone' 14 are positively associated with shot opportunities, there are a few for which the opposite holds true. For example, passes from zones 2, 5, or 7 to zone 14 are all negatively associated with shot opportunities. These could be representative of goal kicks, punts, or other long-balls that are sent from defensive areas deep into the offensive area. This suggests that while zone 14 is generally a 'critical zone' with a strong positive association with shots, it is not beneficial to advance the ball from defensive areas to zone 14 without a coordinated offensive attack.

5.4. Player Rankings by Shot Prediction Models

In the previous section, we described how we trained a model relating a possession to the outcome of the possession ending in a shot. As a result, this model has a feature weight associated with a pass origin and destination for each zone on the pitch, as well as a weight for each origin–destination pair. This provides a map of the pitch that suggests which passes are most likely to lead to a shot opportunity later in the possession. We can use this map to rate a given pass by its association with shot opportunities using our model.

Table 2. Top 10 players in the 2012–2013 La Liga season by average pass shot value (APSV).

Rank	Player
1	Cristiano Ronaldo
2	Lionel Messi
3	Iago Aspas
4	Sergio Garcia de la Fuente
5	Giovani dos Santos
6	Alvaro Cejudo
7	Carlos Reina Aranda
8	Mesut Ozil
9	Karim Benzema
10	Gonzalo Higuain

We took every completed pass in the La Liga 2012–13, and using our model computed an estimate of its relative importance for generating a shot. This importance, called the pass shot value (PSV), is computed for a pass p as

$$PSV(p) = wo_p + wd_p + wod_p$$

where wo_p and wd_p are the model weights for the origin and destination of p, respectively, and wod_p, is the model weight for the pair of the origin and destination of p. For example, a pass from zone 3 to zone 4 would have a PSV of the sum of the model weight for an origin in zone 3, the weight for the destination in zone 4, and the weight of the pair of having an origin of 3 and a destination of 4. We then computed the average pass shot value (APSV) for all players in La Liga who had over 200 completed passes in the 2012–13 season. Since this metric is heavily

biased toward players who are positioned in the offensive ends of the field, we excluded all backs and goalkeepers from this analysis. We plot the APSV for these players in Fig. 8, which shows how our model would rank each player by their average tendency to complete passes that are conducive to leading to a shot. It is not surprising that this value is almost always negative. Most possessions do not end in a shot, and thus most of the model's features are negatively associated with a shot opportunity being generated. Therefore, players make passes with a negative PSV the vast majority of the time. In spite of this, Cristiano Ronaldo has a positive APSV, and as such is the highest ranked player by APSV in La Liga for the 2012–2013 season. This suggests that generally his passes were rated by the model to be positively associated with shot opportunities later in the possession. We present the top ten players for the 2012–13 season of La Liga in Table 2. The top two, Ronaldo and Messi, were universally considered among the best players in the world at that time, seen by the fact that Ronaldo won the 2013 FIFA *Ballon d'Or* trophy and Messi was the runner-up. They were also the number 2 and number 1 scorers, respectively, in La Liga that season, and they both finished the season in the top 10 for assists as well. Others in Table 2 were also successful that season, including offensive-oriented midfielders, such as Mesut Ozil. Although APSV is derived only by examining passes, there seems to be a strong relationship between this metric and overall offensive performance.

6. CONCLUSION

In this paper we presented two approaches to deriving insights into soccer by analyzing characteristics of passing. We first presented a method for characterizing the passing style of a team. We used the locations of the origins of passes to create heatmaps for each team in La Liga 2012–2013 over the course of the season. This distribution of passes creates a 'fingerprint' that can be used to identify teams with a high-degree of accuracy. Using a KNN model, we were able to identify teams from the heatmaps of their pass-origin locations with 87% accuracy in a 20-way classification task. We also showed that a 'fingerprint' was still evident even when we restricted the set of passes to those that originated from the midfield. These results imply that most teams have a characteristic passing style that is consistent throughout a season.

We also showed that the locations of the origins and destinations of passes in a possession relate strongly to whether that possession will end in a shot. Using supervised machine learning techniques, we built a model for predicting whether a possession will end in a shot. The model had an AUROC of 0.785. The features of this model provide a map to understand the relative importance for generating shot opportunities of passing from one location to another. We also used this map to build a simple data-driven ranking of players by weighing a pass by its relative importance for generating a shot later in the possession. When we ranked all offensive players in La Liga 2012–2013 with more than 200 passes with this metric, Cristiano Ronaldo and Lionel Messi, winner and runner-up of the 2013 FIFA Ballon d'Or trophy, came out on top. This ranking also seems to correlate well with standard offensive box score metrics such as goals and assists, even though neither was directly used in its computation. We believe this warrants further investigation into its utility as a player comparison tool.

We believe that our results show that appropriate analyses of pass-event data in soccer can provide interesting and sometimes nonobvious insights. However, soccer is a complicated sport with constantly changing game situations. Incorporating temporal information in any analysis would provide more situation-specific insights. Also, utilizing player-tracking data as a source dataset would better allow investigation into the strategic aspects of the game that are not directly involved with the ball. Expanding our features to include sequential information could give a more detailed understanding of how passing strategy relates to outcomes. Lastly, if a team had a large collection of event data from their own games, they could build team-specific models that would perhaps provide a better analysis of which strategies are most promising in their system. Further investigation will better reveal how useful this type of analysis can be for gaining a deeper understanding of the world's most popular game.

7. ACKNOWLEDGMENTS

We would like to acknowledge the Qatar Computing Research Institute (QCRI) for their funding support and provision of the dataset.

REFERENCES

[1] And the silver goes to The Economist, September 2011.
[2] Soccer Analytics-Presented by Prozone — MIT Sloan Sports Analytics Conference, http://www.sloansportsconference.com/?p=9740.
[3] SlamTracker, http://www.wimbledon.com/en_GB/slamtracker/.
[4] S. S. Intille and A. F. Bobick, A framework for recognizing multi-agent action from visual evidence, In Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence, AAAI '99/IAAI'99, Menlo Park, CA, American Association for Artificial Intelligence, 1999, 518–525.

[5] M. Kates, Player Motion Analysis: Automatically Classifying NBA Plays. Master's thesis; Massachusetts Institute of Technology, 2014.
[6] M. Perše, M. Kristan, S. Kovai, G. Vukovi, and J. Perš, A trajectory-based analysis of coordinated team activity in a basketball game, Comput Vis Image Underst 113(5) (2009), 612–621.
[7] C. Reep and B. Benjamin, Skill and chance in association football, J R Stat Soc Ser A Gen 131(4) (1968), 581–585.
[8] C. Collet, The possession game? A comparative analysis of ball retention and team success in European and international football, 2007-2010, J Sports Sci 31(2) (2013), 123–136.
[9] A. C. Constantinou, N. E. Fenton, and M. Neil, pi-football: A Bayesian network model for forecasting Association Football match outcomes, Knowl Based Syst 36 (2012), 322–339.
[10] M. Kerr, Applying Machine Learning to Event Data in Soccer. Master's thesis; Massachusetts Institute of Technology, 2015.
[11] A. Redwood-Brown, Passing patterns before and after goal scoring in FA premier league soccer, Int J Perform Anal Sport 8(3) (2008), 172–182.
[12] J. Bloomfield, G. K. Jonsson, R. Polman, K. Houlahan, and P. O'Donoghue, Temporal pattern analysis and its applicability in soccer, 2005. In: *The Hidden Structure of Interaction: From Neurons to Culture Patterns* (eds L. Anolli, S. Duncan Jr., M.S. Magnusson & G. Riva), IOS Press, Amsterdam, The Netherlands.
[13] L. Gyarmati, H. Kwak, and P. Rodriguez, Searching for a unique style in soccer, September 2014, arXiv: 1409.0308 [physics].
[14] 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.
[15] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews, Identifying team style in soccer using formations learned from spatiotemporal tracking data, In Data Mining Workshop (ICDMW), 2014 IEEE International Conference on, IEEE, 2014, 9–14.
[16] OptaPro, http://www.optasportspro.com/.
[17] A. Coghlan, The secret of zone 14, New Scientist, 1999.
[18] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, Liblinear: a library for large linear classification, J Mach Learn Res 9 (2008), 1871–1874.