9TH ANNUAL
MIT SLOAN
SPORTS ANALYTICS CONFERENCE
FEBRUARY 27 - 28, 2015 BOSTON CONVENTION AND EXHIBITION CENTER

"Quality vs Quantity": Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data

Patrick Lucey, Alina Bialkowski, Mathew Monfort, Peter Carr and Iain Matthews
Disney Research
Pittsburgh, PA, USA, 15232
Email: patrick.lucey@disneyresearch.com

Abstract

In this paper, we present a method which accurately estimates the likelihood of chances in soccer using strategic features from an entire season of player and ball tracking data taken from a professional league. From the data, we analyzed the spatiotemporal patterns of the ten-second window of play before a shot for nearly 10,000 shots. From our analysis, we found that not only is the game phase important (i.e., corner, free-kick, open-play, counter attack etc.), the strategic features such as defender proximity, interaction of surrounding players, speed of play, coupled with the shot location play an impact on determining the likelihood of a team scoring a goal. Using our spatiotemporal strategic features, we can accurately measure the likelihood of each shot. We use this analysis to quantify the efficiency of each team and their strategy.

1 Introduction

In the 2014 FIFA World Cup in Brazil, arguably the most memorable match was when Germany blitzed Brazil in the semi-final 7-1. However, when analyzing the shooting statistics for this game Brazil actually had more shots and shots on target (18 vs 14 and 13 vs 12 respectively) which does not reflect the sheer dominance that Germany had [1]. In soccer, it is well known that not all shots are created equally, but in this paper we ask the question *"how can we quantify the value of a shot directly from player tracking data?"* An obvious starting point to consider is the proximity of the shot location to the goal – the closer the shot to the goal the more likely it will result in a goal (see Figure 1). However, additional contextual features such as "space" (i.e., the distance from the defender), and number of defenders between the shot and goal play an important role. The position of other attackers their motion paths also give important cues on the quality of shot (as well as uncover how teams get open shots). These features can only be derived from fine-grained player tracking data.

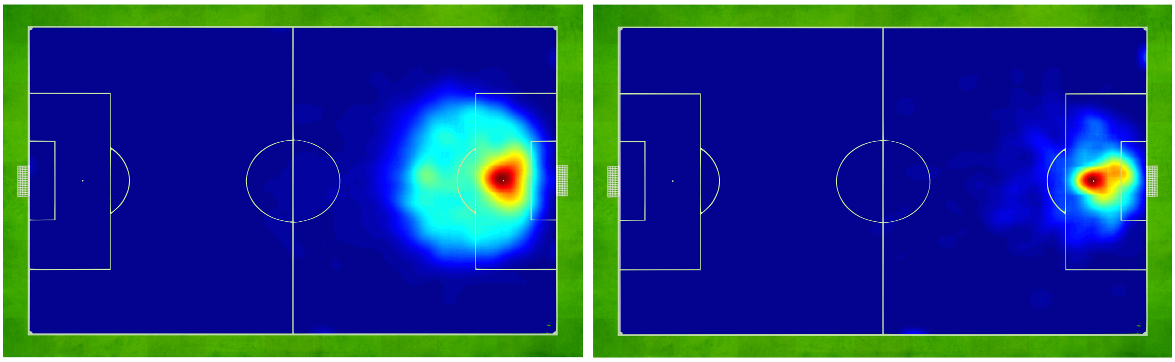

Figure 1. (Left) Shots the probability distribution of all shot locations, (Right) Shows the probability distribution of shot locations then resulted in a goal

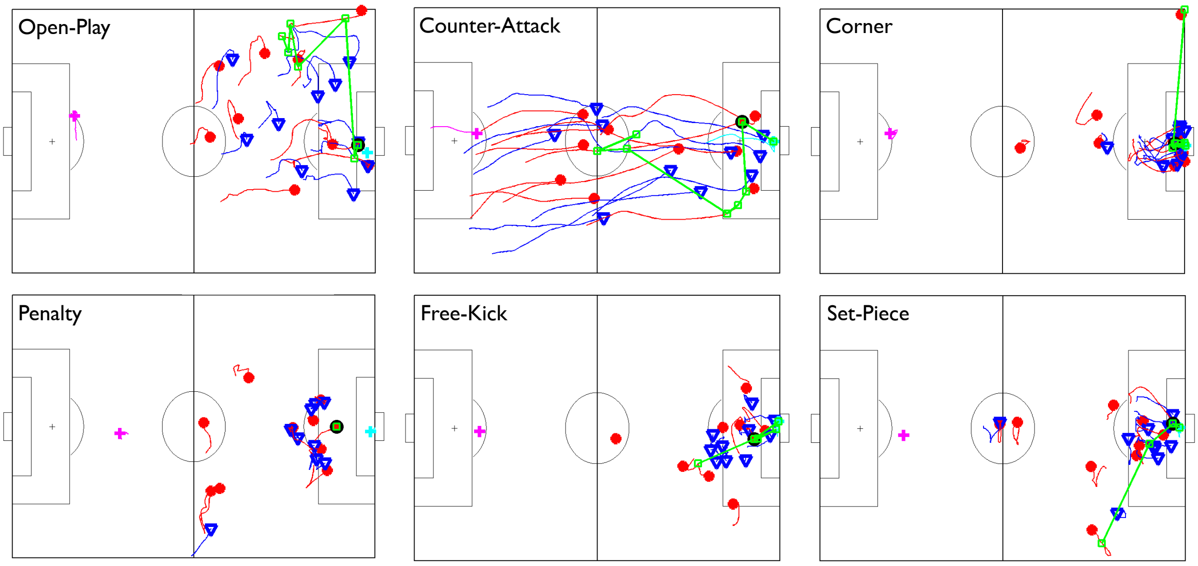

Figure 2. Example plays which represent the different match contexts that shots occur (red team has possession and is attacking left to right).

Match-context also plays an important factor in determining the likelihood of a goal. For this work, we partitioned the shots into six different match-context: i) open-play (possession in the forward third), ii) counter-attack (players break quickly from one-end to the other), iii) corners, iv) penalties, v) free-kick (shot on goal from a free-kick), and vi) set-pieces (a cross that comes into the box from a free-kick) – visual examples are shown in Figure 2. In Table 1 it can be seen that a team is more likely to score on a counter-attack compared to a normal possession (and of course, a penalty). Additionally, a team is more likely to score from a normal possession than a free-kick. In terms of corners, the shot/goal ratio of around 9% appears to represent a reasonable chance, but considering that only a small portion of corners result in a shot, corners tend to be rather inefficient which backs up previous work [2].

Game Context:	Open-Play	Counter Attack	Corners	Penalties	Free-Kick	Set-Pieces
Number (Goal)	6467 (534)	1116 (166)	1115 (100)	94 (67)	539 (26)	388 (39)
Average Shot/Goal	8.26%	14.87%	8.97%	71.3%	4.82%	10.05%

Table 1. Shows the number of shots and goals for the various shot contexts.

In this paper, we present a method which can accurately estimate the likelihood of chances in soccer using strategic features from a seasons worth of player and ball tracking data from a professional league from Prozone [3]. The league we analyzed had 20 teams and played each team home and away. Due to sensitivities of the league, we anonymized the identity of the league and teams - as such we labeled them A-T. From the data, we analyzed the spatiotemporal patterns of the ten-second window before a shot of 9732 shots. In this data, the spatial location of players are given at 10 frames per second, and the spatial location and time-stamp of ball events are given. In the season we analyzed, we used 353 games (27 games were omitted). Due to the constant changing of player role, our recent work of aligning multi-agent player trajectories [4-7] enabled us to craft strategic features which capture fine-grained team spatiotemporal dynamics, which we then fed to a Conditional Random Field (CRF) [8] to estimate the likelihood of a team scoring from a given chance. As soccer is ultimately decided by shots and goals, our approach analyzes teams as a function of both the *quantity and quality of chances*. Similar approaches have been applied to basketball [9-11], however, our approach uses strategic features which incorporates team tendencies instead of individual attributes.

2 Quantifying Goal Likelihood

On the season we analyzed, on average a team will score approximately 9.6% of all their shots. A naïve method would be to assign this estimate for all shots which would lead to a large average error of 0.1745. However, knowing the match-context in which the shot was taken (see Table 1), we can form a better estimate which reduces the average prediction error down to 0.1662. Clearly, this is not satisfactory either as there are many other features that should be incorporated to give a better estimate of goal likelihood. As can be seen in Figure 1, the shot location also plays an important role in estimating if a goal is going to occur so if we further condition the likelihood of the spatial location we can reduce this further. For anyone who has watched soccer, however, these are still very coarse measures and are devoid of fine-grain context. For example, in a counter-attack, if a player is one-on-one with the goal-keeper, his/her chances are increased. Or given space on top of the 18 yard box, a player is more likely to score given space from the defenders. The only means of getting such features is from using fine-grained player tracking data and crafting features from this information. Using this data, not only can we obtain important spatial and action data, we can also include strategic elements such as the motion of surrounding players and the structure of the defending team. In the following subsections, we describe how we captured these semantic and strategic elements.

2.1 Defender Proximity

Having a defender in close proximity effects the decision that will be made as well as the execution. As the major goal of a defender is to protect the goal, the orientation of their position relative to their goal needs to be captured. The way we determined defender proximity is by first checking if any defending players were in the area between the shot and the goal (see Figure 3). If they were, we calculated the Euclidean distance between the shot location and defender. If a defender was not within this area, we gave this distance a negative value. In open-play, when a defender was not goal-side when a shot occurred (i.e., not within the shaded space) the likelihood of a goal increased to 11.59%, compared to 7.49% ($p<0.00001$) where 23.18% of shots were "open". Similarly for counter-attacks, the likelihood of a goal increased to 18.44% compared to 12.46% ($p<0.01$) where 40.32% of shots were "open". Additionally, for counter-attacks getting an open shot occurred more often which makes sense as more space is created in a counter attack. Of course, this also depends on the distance from goal. We also used the number of defenders in this area as a feature as well. To determine if a defender was within this area, we defined the vertices of the triangle and then used standard point-in-polygon calculation.

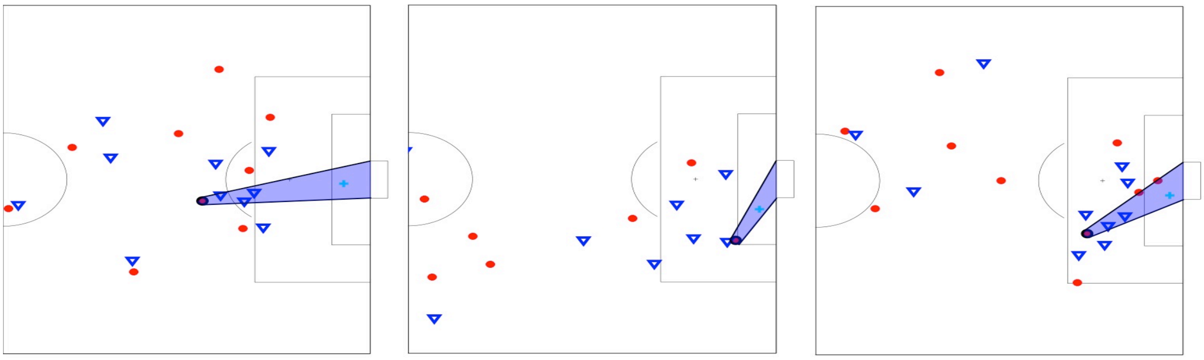

Figure 3. To capture how much pressure the shooter is under, we devise a defender proximity descriptor which first counts how many defenders are in the space between the goal and the shooter and then we get the distance between the shooter and the defenders (red are the attacking players and blue are the defenders).

2.2 Defensive Formation/Structure

The shape and defensive structure plays an important part in the likelihood of a goal. Quantifying team structure is difficult however, but our recent work in this area has allowed us to craft features to measure such behavior [4-7]. Vital to this estimation is to determine the "role" of each player within the formation. This is done by finding the permutation of the raw location points which minimizes the distance to the base template. Once this cost matrix is calculated, we use the Hungarian algorithm [12] to make the assignment of role to each player. Once this is determined, we crafted the following features: i) the distance between the defensive line, ii) the distance between the back-line and the midfield line, iii) the number of defensive role-swaps, and iv) the number of attackers in-front of the defensive center.

2.3 Attacking Features

In this subsection, we describe some of the attacking features we extracted from the data. Important factors which we wanted to extract where, was it a long pass, cross, dribbling and taking on players, or pressing (causing turnovers) which lead to the shot on goal. Additionally, the space of the player who gave the incoming pass/cross also plays an important role as it suggests the quality could be potentially higher. The pace of the players moving and how the attacking team moves relative to the opposition was also captured in our feature set.

2.4 Expected Goal Value using Strategic Features for each Game-Context

Given the game-context and the various spatiotemporal features, we can estimate the likelihood of each shot using logistic regression. We call this estimate **Expected Goal Value (EGV),** which is similar to other approaches used in basketball [9-11]. Approaches such as these have also been used in soccer, but have not included player tracking data – just ball-event data. As the game-context is important, we first partition the examples into distinct game-context clusters and learn an individual regressor for each of these 6 game-context. To avoid over-fitting, we used regularization and we divided the examples into a train/test set. Using the features we show in Table 2 how the average error of our prediction lowers.

Factor	**Average Likelihood**	**Shot-Context**	**Context + Location**	**Context + Location + Defending**	**Context + Location + Defending + Attacking**
Average Error	0.1745	0.1662	0.1554	0.1545	0.1439

Table 2. Showing the residual when we use different methods to estimate the likelihood of a shot resulting in a goal.

3 Team and Game Analysis

3.1 Team Efficiency Ratings (Season wide analysis)

Having the ability to better estimate the likelihood of a goal, allows us to do deeper analysis which may help unlock characteristics or traits of teams. First of all, we can evaluate the efficiency of a team's performance in terms of offense and defense and compare them to the rest of the teams. Using this as a starting point, we can drill down further to check how efficient each team is in terms of different match context. Let us first analyze the attacking performance across the season (due to some missing matches, the overall statistics here may not match the complete season). The performance is shown in Table 3.

ID	SHOTS	GOALS	EGV	Average Error
A	514	58	51.91	7.89
B	434	46	39.47	5.85
C	594	68	63.62	9.57
D	562	46	50.4	6.95
E	440	42	42.57	6.21
F	694	65	65.85	9.28
G	593	59	62.85	9.65
H	**514**	**71**	**57.21**	**10.35**
I	474	41	40.16	5.33
J	416	39	41.45	5.78
K	**447**	**26**	**35.95**	**3.93**
L	364	33	34.3	4.66
M	464	42	44.12	6.01
N	338	29	34.72	4.54
O	416	39	38.93	5.38
P	467	45	44.95	6.86
Q	611	57	50.13	6.98
R	458	43	46.39	6.40
S	479	40	43.34	5.86
T	458	44	38.44	5.28

ID	SHOTS	GOALS	EGV	Average Error
A	371	34	35.64	4.79
B	620	62	59.31	8.47
C	443	35	38.42	5.12
D	415	37	41.68	5.87
E	604	58	60.01	9.09
F	407	38	37.45	5.01
G	353	26	31.19	4.12
H	451	38	35.16	4.46
I	**458**	**59**	**47.13**	**7.96**
J	533	56	51.63	7.76
K	547	48	47.61	6.53
L	614	62	59.02	8.73
M	389	50	44.87	8.02
N	447	35	40.2	4.76
O	**592**	**50**	**61.33**	**8.16**
P	523	49	53.13	7.29
Q	344	41	36.3	6.13
R	529	46	52.38	7.08
S	576	45	45.41	5.76
T	**521**	**64**	**48.85**	**7.68**

Table 3. All Shots: (Left) The offensive shooting statistics, and (Right) defensive statistics for every team in the league. Columns 5 and 6 give the expected goal value and the average error per prediction. The rows highlighted in bold highlight teams where their goals is significantly different than their EGV.

In terms of offense (left), taking into account the error of the estimate, most teams scored within their expected range with the exception of two teams. Team H were very efficient, scoring 71 goals (but with an expected goal value of approximately 57±10 goals). Even with the maximum error, a difference of 4 goals is quite significant. On the other-hand, Team K only scored 26 goals, which was different from their EGV of 36±4. As both teams finished at either ends of the table, the quality of strikers may suggest the difference between actual goals scored and their EGV. In terms of defense, similar patterns emerge with Team I conceding 59 goals when their EGV was around 47±8. Team T also game up more goals then expected, with a EGV of 49±8, when they actually gave up 64 goals. Poor goal-keeping, excellent strikes by the opposition or a combination of the two could have caused this. Team O on the other hand, were expected to give up 61±8 goals, but only conceded 50 goals which maybe due to the inverse of the previous example.

Performance may vary for various match contexts too. In Table 4, we show the EGV for the various teams based on shots just from open-play. When we focus on this particular match-context, Teams H and K still have a big difference between actual goals and EGV, but three other teams do as well. Team Q scored more goals then expected (this could be due to the fact they had a player who score some incredible goals that season). Teams N and E underachieved though, which may suggest the lack of quality for those teams. In terms of defense, Team N only conceded 17 goals in open-play with their EGV being higher at 25.5±2.8. Team P also conceded much less than expected. Teams I and T conceded much more than expect for shots in open-play.

ID	SHOTS	GOALS	EGV	Average Error
A	358	35	31.01	4.66
B	281	26	23.38	3.43
C	401	37	36.16	5.35
D	390	31	33.25	4.66
E	**297**	**20**	**26.19**	**3.57**
F	468	38	37.85	5.12
G	371	30	32.87	4.70
H	**332**	**39**	**27.52**	**4.27**
I	326	26	24.42	3.19
J	276	20	23.31	3.04
K	**290**	**14**	**19.49**	**1.87**
L	225	17	18.67	2.57
M	328	28	27.34	3.67
N	**214**	**14**	**18.29**	**2.28**
O	249	19	18.77	2.42
P	333	35	30	4.63
Q	**411**	**37**	**29.94**	**4.14**
R	287	19	21.63	2.60
S	334	26	26.72	3.60
T	296	23	21.78	2.73

ID	SHOTS	GOALS	EGV	Average Error
A	223	16	17.18	2.08
B	427	37	34.05	4.80
C	309	22	26.19	3.33
D	258	18	22.08	2.91
E	420	33	33.28	4.56
F	267	18	21.38	2.66
G	210	16	16.21	2.05
H	297	23	20.44	2.51
I	**294**	**38**	**26.21**	**4.53**
J	368	34	30.07	4.33
K	352	30	25.99	3.46
L	420	40	35.60	5.12
M	245	25	23.51	3.86
N	**315**	**17**	**25.50**	**2.78**
O	**375**	**26**	**33.10**	**4.43**
P	**353**	**23**	**30.52**	**3.91**
Q	230	25	21.73	3.55
R	347	25	29.77	3.97
S	420	33	30.45	3.94
T	**337**	**35**	**25.32**	**3.71**

Table 4. Open Play: (Left) The offensive shooting statistics, and (Right) defensive statistics for every team in the league. Columns 5 and 6 give the expected goal value and the average error per prediction. The rows highlighted in bold highlight teams where their goals is significantly different than their EGV.

ID	SHOTS	GOALS	EGV	Average Error
A	63	10	9.53	2.13
B	57	10	9.92	2.27
C	56	8	8.96	1.79
D	34	5	4.83	0.97
E	47	8	9.21	2.19
F	88	15	15.31	3.46
G	88	13	13.08	2.73
H	82	16	13.85	3.59
I	59	9	7.18	1.48
J	38	4	6.24	1.19
K	52	6	7.38	1.54
L	35	3	4.06	0.67
M	44	5	5.47	0.84
N	35	5	5.12	1.06
O	59	8	7.04	1.49
P	51	6	8.4	1.82
Q	77	11	13.56	2.99
R	57	10	9.85	2.11
S	35	4	5.34	1.00
T	59	10	8.01	1.70

ID	SHOTS	GOALS	EGV	Average Error
A	56	7	8.07	1.57
B	47	8	8.1	1.88
C	51	7	7.3	1.64
D	53	10	8.53	1.98
E	70	11	11.27	2.69
F	36	5	6.14	1.25
G	60	5	7.62	1.29
H	59	8	6.95	1.31
I	57	9	9.87	2.39
J	**58**	**14**	**9.25**	**2.59**
K	53	6	8.19	1.57
L	88	8	11.52	2.01
M	42	11	7.62	2.09
N	38	4	5.06	0.93
O	61	11	12.5	2.93
P	51	8	8.88	1.71
Q	38	4	7.84	2.05
R	57	9	9.23	1.81
S	64	5	7.29	1.08
T	**77**	**16**	**11.11**	**2.61**

Table 5. Counter Attack: (Left) The offensive shooting statistics, and (Right) defensive statistics for every team in the league. Columns 5 and 6 give the expected goal value and the average error per prediction. The rows highlighted in bold highlight teams where their goals is significantly different than their EGV.

In Table 5, we show the difference between actual goals scored and conceded for counter-attacks. There is no enormous gaps between the actual goals and EGV apart from Team J defense which game up 14 goals, when they should have only gave up 9.2±2.6 and Team T who gave up 16 goals when they were expected to only give up 11.1±2.6.

3.2 Individual Game Analysis

Circling back to our original example of Brazil vs Germany where the statistics do not tell the full story of a match, in this subsection we show that our analysis can also be used to give a better indication on whether a team was "dominant" or "lucky". What we mean by that is soccer is still rather random by nature due to the fact that goals are sparsely occurring events and outliers can occur such as a goal-keeper having a bad day, or all shots for a particular team being successful. We show some examples in Table 6. In the top 3 examples, we show three matches where teams with significantly less shots won (the first two by large margins), but our EGV measure gave a better approximation of dominance. In the remaining 6 examples, we show matches where the dominant team did not win despite having the better chances. Over the season these tend to cancel each other out, but in terms of individual data points this can give a better indication of how the match was played.

Example1		
Teams	M	S
Shots	17	11
Goals	0	3
EGV	1.50	2.89

Example 2		
Teams	K	P
Shots	22	14
Goals	0	5
EGV	1.39	2.02

Example 3		
Teams	I	S
Shots	18	12
Goals	0	1
EGV	0.83	1.54

Example 4		
Teams	I	O
Shots	17	9
Goals	0	3
EGV	1.37	0.74

Example 5		
Teams	C	M
Shots	19	7
Goals	2	2
EGV	2.14	0.66

Example 6		
Teams	O	L
Shots	15	19
Goals	3	0
EGV	1.65	1.65

Example7		
Teams	F	B
Shots	29	10
Goals	1	3
EGV	2.66	0.75

Example 8		
Teams	F	R
Shots	28	5
Goals	0	2
EGV	2.87	0.53

Example 9		
Teams	F	N
Shots	18	5
Goals	0	0
EGV	2.25	0.07

Table 6. Examples of matches using the EGV measure to given a better idea of how the match was played and which team dominated.

4 Quantifying Chances: Examples

It is one thing to have a reasonable model, but if the predictions do not look "reasonable" then there is a good chance something is going wrong. In this section, we visualize some of our predictions to show that it passes the "eye-test". Examples are shown in Figure 4. In the top left, a play which has the left-winger controlling down the left uncontested and then slotting the ball between the back four to a player in the six-yard box results in a chance of 70.59%. In the second example, a similar break occurred with the ball being crossed to the striker with a defender in close proximity which reduced the goal likelihood. In the third example, a free-kick which was taken and was parried by the goal-keeper had a chance of around 50% ending up as a goal (i.e., for every 2 times you see that occur, one will go in). The fourth example shows a corner which results in a shot in the six yard-box giving a likelihood of 46.10%. However as this would normally be taken by a goal-keeper the likelihood of getting a shot in this instance is low. The remaining examples show low percentage shots often occur when the location is outside the box and a defender is in the way. We didn't show penalties kicks, as these have little variance in terms of strategic factors.

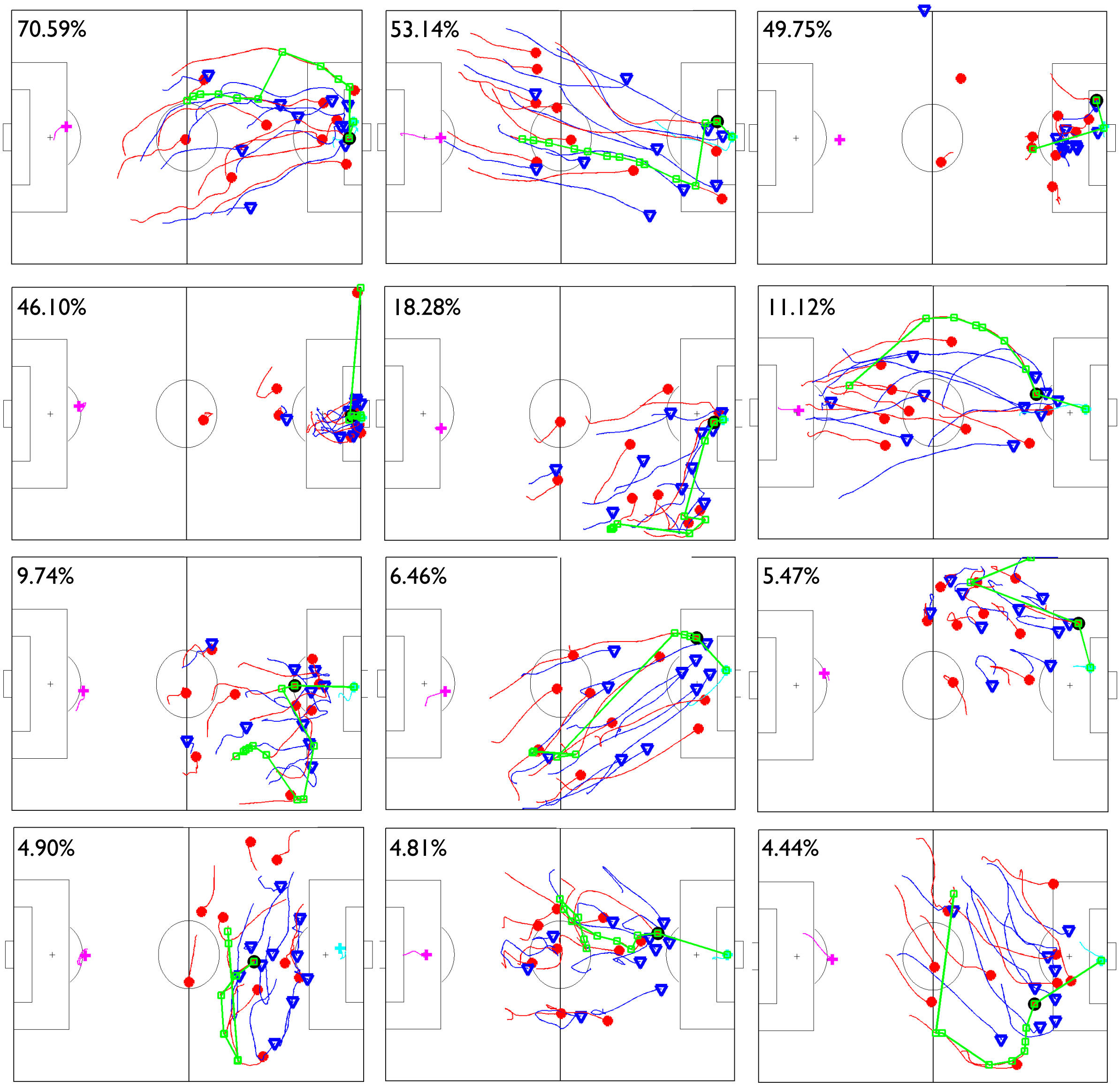

Figure 4. Examples showing the goal likelihood from various examples (red team has possession and is attacking left to right).

5 Summary

In this paper, we presented a method which accurately estimates the likelihood of chances in soccer using strategic features from an entire season of player and ball tracking data taken from a professional league. From the data, we analyzed the spatiotemporal patterns of the ten-second window of play before a shot for nearly 10,000 shots. From our analysis, we found that not only is the game phase important (i.e., corner, free-kick, open-play, counter attack etc.), the strategic features such as defender proximity, interaction of surrounding players, speed of play, coupled with the shot location play an impact on determining the likelihood of a team scoring a goal.

REFERENCES

[1] 2014 Fifa World Cup Semi-Final Match, Brazil vs Germany Match Statistics, http://www.fifa.com/worldcup/matches/round=255955/match=300186474/statistics.html.
[2] C. Anderson and D. Sally, "The Numbers Game: Why Everything You Know About Soccer is Wrong", Penguin Books, 2013.
[3] Prozone. http://www.prozonesports.com
[4] A. Bialkowski, P. Lucey, P. Carr, Y. Yue and I. Matthews, "Win at Home and Draw Away: Automatic Formation Analysis Highlighting the Differences in Home and Away Team Behaviors", in *MIT Sloan Sports Analytics Conference (MITSSAC)*, 2014.
[5] P. Lucey, A. Bialkowski, P. Carr, S. Morgan, Y. Sheikh and I. Matthews, "Representing and Discovering Adversarial Team Behaviors using Player Roles", in the *International Conference of Computer Vision and Pattern Recognition (CVPR),* 2013.
[6] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan and I. Matthews, "Large-Scale Analysis of Soccer Matches using Spatiotemporal Tracking Data", in the *International Conference of Data Mining (ICDM)*, 2014.
[7] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan and I. Matthews, "Identifying Team Style in Soccer using Formations Learnt from Spatiotemporal Tracking Data", in the *ICDM Workshop on Spatial and Spatiotemporal Data Mining (SSTDM)*, 2014.
[8] J. Lafferty, A. McCallum and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", in *International Conference on Machine Learning (ICML)*, 2014.
[9] Y. Yue, P. Lucey, P. Carr, A. Bialkowski and I. Matthews, "Learning Fine-Grained Spatial Models for Dynamic Sports Play Prediction", in the *International Conference of Data Mining (ICDM)*, 2014.
[10] A. Miller, L. Bornn, R. Adams, and K. Goldsberry, "Factorized point process intensities: A spatial analysis of professional basketball," in *International Conference on Machine Learning (ICML)*, 2014.
[11] D. Cervone, A. D'Amour, L. Bornn and K. Goldsberry, "POINTWISE: Predicting Points and Valuing Decisions in Real Time with NBA Optical Tracking Data", in *MIT Sloan Sports Analytics Conference (MITSSAC)*, 2014.
[12] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.