MIT SLOAN SPORTS ANALYTICS CONFERENCE
presented by ESPN

Quantifying the Value of Transitions in Soccer via Spatiotemporal Trajectory Clustering

Jennifer Hobbs, Paul Power, Long Sha, Hector Ruiz, Patrick Lucey
STATS, AI Group
Email: jhobbs@stats.com, paul.power@stats.com, lsha@stats.com, hruiz@stats.com, plucey@stats.com

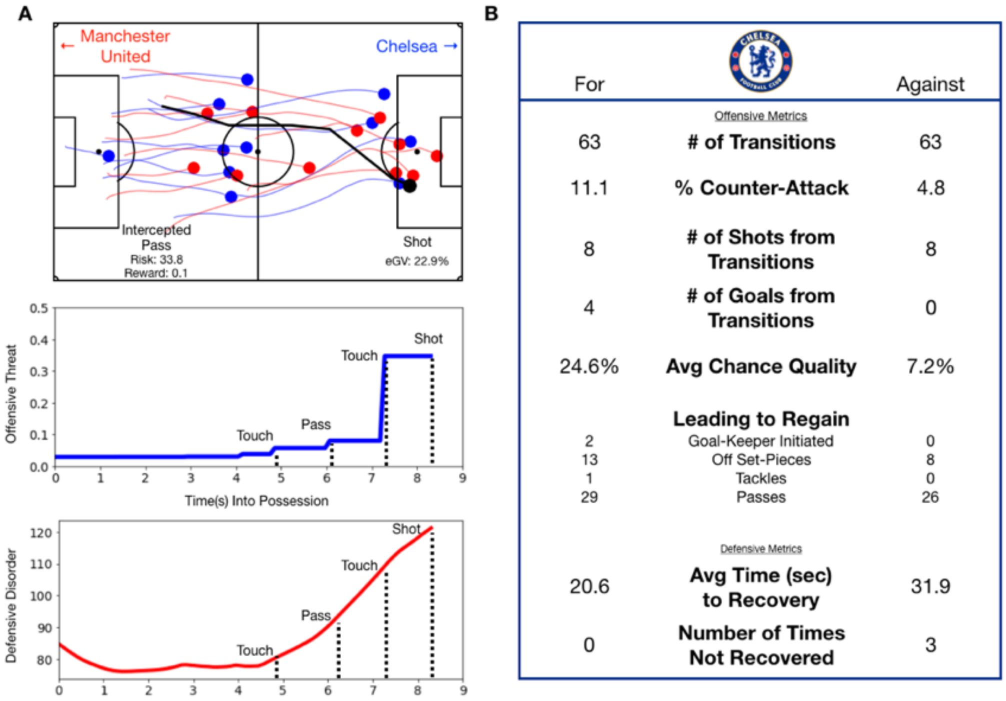
1. Introduction

In recent years, the *"counter-press"*, *"geggenpress"* and *"counter-attack"*, have been in vogue due to their ability to create good scoring chances by effectively overloading the team that has just lost the ball. These *"transitions"* capture how effective a team is as they transit from defense to offense (and vice-versa) without any stoppage in play. *Counter-attacks*, a sub-category of transitions which are especially fast-paced, aggressive, and direct, provide some of the best opportunities for scoring and are a potent strategy when executed effectively [1]. Specifically, if a team counters immediately (creating a chance less than 15sec after regaining the ball), the likelihood of scoring is **12.4% vs 8.0%** (p-value<0.01), when they look to maintain possession. Additionally, the likelihood of a team obtaining a shot is **35.1% vs 10.6%** (p-value<0.01).

The value and importance of transition play is highlighted in Figure 1. In this game between Chelsea and Manchester United, both teams had the same number transitions (63) and the same number of shots (8) off of these transitions. However, these numbers do not capture the full story: Chelsea counter-attacked on their transitions almost 2.5 times more often than Manchester United, leaving Manchester United defensively disordered, resulting in superior scoring opportunities, and ultimately, four goals off of the counter-attack.

Identifying types of transition play, such as counter-attacks, is crucial to this type of analysis. Despite the importance around transitions in soccer, in terms of analytics, no quantitative measures have emerged. There are two prime reasons for this: i) obtaining the precise onset and offset time-stamp of a counter-attack is extremely challenging as the task is subjective and fine-grained (i.e., human annotators are not reliable at this task), and ii) measuring the structural patterns and movements of a team is equally subjective and challenging for a human to annotate.

In this paper, we do this automatically and objectively using machine learning techniques. Our approach is a hybrid between supervised and unsupervised learning using player tracking data. First, we hierarchically cluster the multi-agent player trajectories to determine a codebook of commonly run plays, or *playbook*. For each playbook entry we are able to determine a game-state specific formational template for each team; from this we are able to quantify the "*defensive disorder*" of a team as they transition from offense to defense. Taking inspiration from work done in semantic trajectory analysis, we incorporate event sequence and team information to create a measure of *"offensive threat"* which we use to further cluster each playbook entry into threatening and non-threatening plays which use can use to identify counter-attacks. From this analysis we are able to detect counter-attacks directly from the player-tracking data, without any human labels, and then use this to quantify the value and impact of execution on the counter-attack, both offensively and defensively.



For		Against
	Offensive Metrics	
63	# of Transitions	63
11.1	% Counter-Attack	4.8
8	# of Shots from Transitions	8
4	# of Goals from Transitions	0
24.6%	Avg Chance Quality	7.2%
	Leading to Regain	
2	Goal-Keeper Initiated	0
13	Off Set-Pieces	8
1	Tackles	0
29	Passes	26
	Defensive Metrics	
20.6	Avg Time (sec) to Recovery	31.9
0	Number of Times Not Recovered	3

Figure 1: (a) *Top*: Chelsea regains possession of the ball off of a risky pass made by Manchester United and proceeds to move the ball with speed, culminating in a high quality shot 8 seconds later. *Middle*: As the ball is moved up the field and into open space, the offensive threat of the possession grows. *Bottom*: Examination of the defensive formation shows that as the possession evolves, Manchester United becomes increasingly positionally disordered. (b) How teams capitalize on transition opportunities is key to their success in a match. Offensively, Chelsea was more aggressive after regaining the ball, resulting in higher quality chances and number of goals scored. Defensively, they managed to recover more quickly and effectively after losing possession of the ball.

2. Learning the Playbook of Transitions

2.1 Hierarchical Template Learning for Player Alignment and Codebook Formulation

Our end goal is to identify segments of game play, such as transitions, which lead to good scoring opportunities and to analyze the positioning and movements of the players which created those opportunities. To accomplish this, we must begin with an accurate and robust player-role alignment, which is critical for the analysis of multi-agent spatiotemporal analysis as has been shown within the domain of sport [2-6] as well as the broader machine learning and pattern recognition community [7].

Formational analysis, alignment (i.e., assignment of each player to a given role in that formation), and game-state discovery (i.e., playbook formulation) are intrinsically interwoven and cannot be solved independently. Without appropriate alignment, formational patterns are lost as players swap positions and game-state identification is limited. Without a template that is descriptive of the present game-state, robust alignment is challenging. And without identification of the current game-state, an accurate template cannot be constructed. Many of these issues could be addressed via manual labeling of formation, alignment, or game-state. However, we seek to solve these key challenges in a purely unsupervised manner by exploiting the spatiotemporal structure in the data to avoid the inherent subjectivity and time-consuming nature of manual annotation.

Here we utilize a tree-based alignment process based on hierarchical clustering and linear assignment techniques to align each player to his appropriate role-assignment in a given game-state. This tree-based alignment process is described in detail in [5] and is summarized as follows: first, a global template is constructed which corresponds to the average play; the player centroids of this first "cluster" then serves as our global template to which each player is initially aligned. In the next and all subsequent iterations, the data is clustered via k-means to partition the previous node into sub-clusters where the number of clusters is determined via the Silhouette score.

After clustering occurs, the template centroids (i.e. the average trajectories for each player-assignment in that cluster) are aligned to the parent node to ensure consistency of alignment. Next, each player in that sub-cluster is re-aligned, now to the template of the child-node; this allows the alignment to be performed on a template that best captures the current game-state context. This process of align-cluster-align is repeated until a maximum tree depth is reached or the number of plays in each leaf-node falls below a minimum value. The output of this process is a tree where each leaf-node corresponds to an entry in our playbook and the trajectories of the players are assigned to their appropriate role in that playbook entry.

2.2 Constructing the Playbook

We constructed our playbook using player and ball tracking data from the 2016-17 English Premier League. In-play segments of each game were cut into 10-second segments, offset in 1-second increments, to generate plays of interest for training.

The maximal tree-depth was found to be 9 with 218 total leaf-nodes or *playbook entries*. As the focus of this paper is on transitions, we have a shown a portion of the playbook run on plays immediately following possession regain (i.e. a change in possession with no stoppage of play) in Figure 2. The transition-portion of the playbook has 14 leaf-nodes which occur at a depth between four and five.

Other works have used unsupervised learning to derive a playbook based on player trajectories. However, in contrast with approaches such as [8], which use a bottom-up approach to characterize individual player trajectories, our approach is inherently top-down, allowing for the coordinated movements of players to drive the codebook formulation. Furthermore, our alignment enables us to directly capture the natural structure between the players instead of combining them in an unordered fashion via techniques such as bag-of-words.

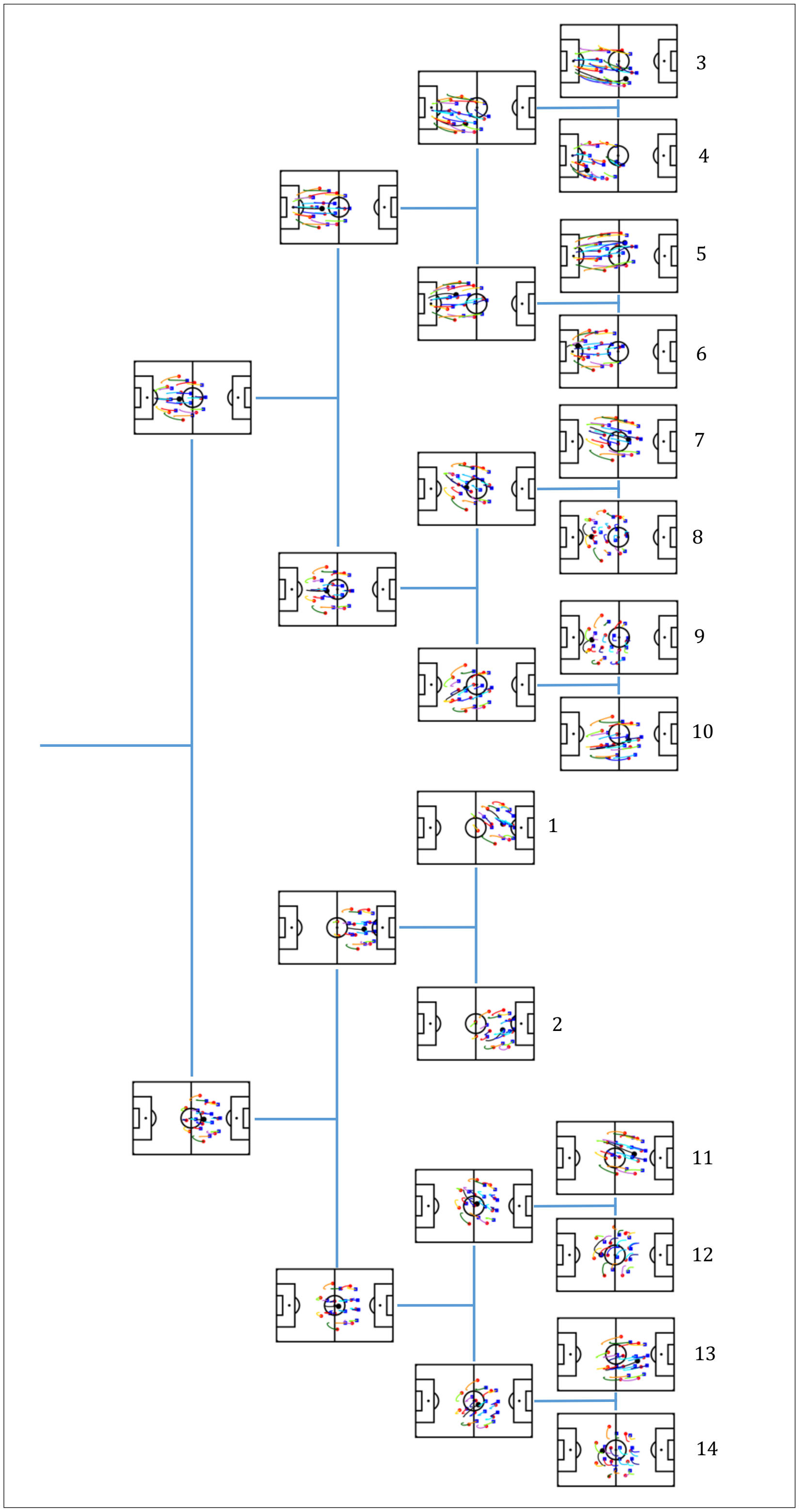


Figure 2: A portion of the playbook learned via tree-based alignment. At each step in the tree the data is clustered, and the players aligned within that cluster. Note that although there are only 2 child-nodes at every split, up to 6 nodes were allowed with the optimal number chosen by examining the Silhouette score. In each subplot the attacking team is indicated by red circles while the defending team is indicated by blue squared; the data is normalized so that the attacking team is always attacking to the right. The color of the trajectory corresponds to the role assigned to that player. For the analysis that follows, we focus on these 10-second plays which correspond to plays immediately following a possession regain. The 14 lead nodes shown have been given a condensed label which we will reference throughout the analysis.

Our current approach has several advantages. First, having multiple templates allows us to perform better alignment as the play-specific template better captures the average positioning in that cluster. Second, a trajectory versus single-frame approach enables us to capture both the spatial and temporal elements of a play and thereby captures a spatio-temporal template, and not just a purely spatial one. Additionally, we simultaneously obtain both the desired player alignment as well as a codebook of 10-second trajectories (i.e. our playbook) which describes different states of the game.

2.3 "*Defensive Disorder*"

Quantifying the performance of a defense is a challenge across sports, although recent work has made tremendous strides in this area [4]. Of particular relevance to transitions is the notion of disorder; it is a widely held belief that as a team transitions from offense to defense, they are prone to deviate from their preferred formation and therefore susceptible to the counter-attack.

Using this notion as a starting point, we leverage the playbook templates learned during the alignment process to identity the preferred formation. Then we are able to simply find the total displacement or cost associated with moving each player back to his preferred location in the formation. Note that it is critical that we use our multiple game-state specific templates for this analysis; a single template would result in an over-inflated disorder score for formations that are perfectly normal for a given state.

Referring back to Figure 1 we see how our measure of *defensive disorder* can shed light on a team's transition performance: as play evolves, we are able to calculate their disorder in each frame. From this we can determine if and when a team returns to its stead-state, or whether they become increasingly disordered, as seen in Figure 1(a).

2.4 Identification of Game-States

From Figure 2 we are able to quickly identify a number of these playbook entries as relevant, well-known modes of play. Firstly, nodes 1 and 2 correspond to regains which occur near the team's attacking goal. It is important to note that since these are the first ten-seconds following the regain, the team must have been able to regain the ball in this region and then proceeded to attack immediately. Therefore we can identity these two clusters as corresponding to the counter-press or "geggenpress"*.*

We also wish to identify plays that correspond to a counter-attack. This proves to remain a difficult challenge at this point; clusters 3, 5, and potentially 7, 10, 11, and 13 show some direct, relatively fast-paced attacks toward the goal. However, how aggressive and over what time and/or distance must an attack be for it to be considered a counter-attack? As we do not have labels, and since humans are notoriously poor at labeling such tasks anyway, we must rely on additional data-driven cues to perform this identification. Identifying counter-attacks within these transitional play states is our focus in Section 3.

3. Identification of Counter-Attacks via Semantic Trajectory Analysis

3.1 *"Offensive Threat"*

The hierarchical clustering approach described above provides us with a playbook based purely on the raw tracking data which captures the spatio-temporal similarities of play. However, this clustering step is agnostic to how it relates to goal-scoring opportunities in soccer. Here we enhance our trajectories with an "*offensive threat*" metric which corresponds to the likelihood of a shot occurring within the next ten seconds. To generate this metric, ball position, binary event labels (corresponding to events such as passes, touches, tackles, crosses, etc.), and attacking team information are fed into a logistic regressor.

We apply this model to each play in our training set. Specifically, we generate an offensive threat curve for every example in each of the leaf-nodes generated by our hierarchical clustering. Next, we use k-means to sub-cluster the leaf-nodes based on their offensive threat curve. We examined n=1 to n=8 sub-clusters and in every case, n=2 or n=1 (i.e. no further clustering) was supported by Silhouette analysis. For this, we were able to further segment each playbook leaf node as either "non-threatening" and potentially a "threatening". This approach is called semantic trajectory analysis and it has been used in various other domains outside of sport [9]. It avoids the need for human annotations as they are time consuming, and subjective. Our approach derives these "annotations" directly from the data.

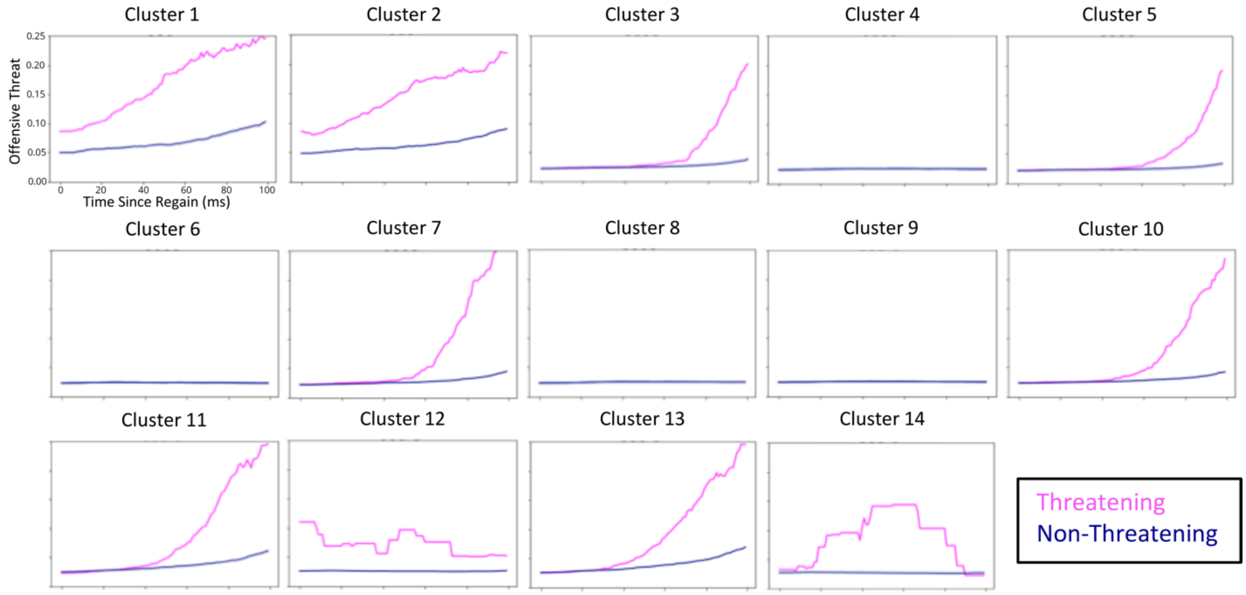


Figure 3: We create an "offensive threat" metric which measures the likelihood of a shot being taken within the next 10-seconds. This value is calculated at every frame for every example used in the hierarchical clustering discussed in Section 2. We then cluster each of the leaf-nodes based on the traces of this metric to create a "threatening" and a "non-threatening" sub-cluster; note that some leaf-nodes do not have a threatening sub-cluster.

3.2 Identification and Analysis of Counter-Attacks

The creation of the playbook and semantic sub-clustering now provide us with the ability to identity counter-attacks as those plays in clusters 1,2, 3, 5, 7, 10, and 13 which correspond to the "threatening" sub-clusters. This corresponds to 9.8% of the transition playbook, which is consistent with the frequency of successful counter-attacks which domain knowledge.

With the ability to objectively identify counter-attacks, we are now in the position to quantify their importance. The value of counter-attacks is seen through a comparison to other, non-counter attack transitions as seen in Table 1. First, we see that counter-attacks generate more shots. In fact, not only are shots more likely to be generated during the transition period (i.e. the first 10 seconds after a regain), but at any time during the ensuing possession. This suggests that even if a counter-attack does not directly lead to a shot, it may result in keeping the defense unstable for a period of time, resulting in subsequent chances. Additionally, the shots generated from a counter-attack are of higher quality, i.e. have a higher expected goal value [1, 16], than shots during non-counter-attack possession.

We also examine the passing involved during transition play. Here we leverage the passing risk (i.e. probability that a pass is incomplete) and reward (i.e. probability a pass leads to a shot) models from [17]. ***Consistent with the discussion above, counter-attacks involve passes with a higher reward value. However, they are no riskier on average than passes made during other transitions***.

	Counter-Attack	Other Transitions
% Possessions with Shot (any time)	35.1%	10.6%
% Possessions with Shots in First 10sec	11.8%	0.04%
Time to First Shot	12.9 sec	28.3 sec
Expected Goal Value	12.4%	8.0%
Average Passing Risk	13.0%	13.1%
Average Passing Reward	18.5%	3.34%

Table 1: Comparison of counter-attacks to all other transitions. Possessions which begin with a counter-attack are more likely to have shots (both within the first 10seconds as well as at any time during that possession) and have higher quality shots (i.e. higher expected goal value). Examining the passing involved during transitions suggests that while counter-attacks contain passes with a higher reward value (i.e. more likely to generate a shot), they are no riskier than passes made during other transitions. All differences, except passing risk, are significant ($p < 0.01$).

4. Analyzing Transitions: English Premier League, 2016-2017

4.1 Transition Style

An immediate benefit of our hierarchical alignment and codebook method is that it provides us with a spatiotemporal descriptor of each team. Prior work such as [10-12] has described the style of a team through a combination of traditional match statistics, ball possession, and formational features. Here we use an approach similar to [13] in which we construct our style descriptor directly from the team's playbook, that is, the frequency with which each leaf-node entry is observed during a game. Recall that [14] compared teams based on the locations of the field in which they tended to regain or lose possession of the ball. Our approach further captures the temporal element of loss/regain as well the incorporation of all the player's trajectories in addition to the ball.

As we are interested in transitions for this particular analysis, we specifically looked at each team's "transition style", that is, the types of plays run immediate after regaining the ball. An example of this style descriptor is shown in Figure 4(a) for Manchester City. Note the significant frequency with which regains from clusters 1 and 2 occur and recall from Section 2.4 that these clusters are associated with the counter-press, which Pep Guardiola is known to employ aggressively.

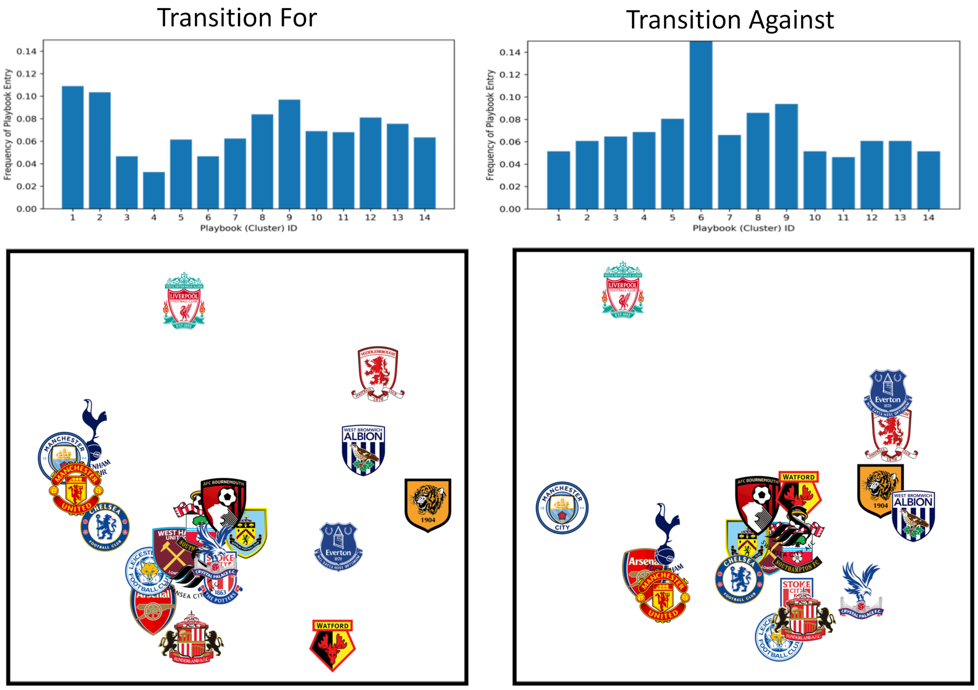


Figure 4: Analyzing team's transition style. (a) A style vector can be created for a given team (Manchester City shown here) based on the frequency with which they run a play immediately after regaining the ball (left) and another style vector for the frequency of plays which are run against them immediately after they lose the ball (right). The cluster IDs correspond to those shown in Figure 2. Notice the high frequency with which Manchester City's transitions begin from clusters 1 and 2 which are associated with counter-pressing. (b) The difference between each team's style vectors can be compared to illustrate the similarity among teams.

To get an overview of the similarity among teams, we can compare the style vectors between teams. As our hierarchical clustering method has already accomplished the challenging task of linearizing the data for us, we simply take the L2-distance between the style vectors of each team. Isomap [15] a non-linear embedding technique, is then used to reduce this 14-dimensional distance matrix down to 2-dimensions for visualization.

Figure 4 shows the similarity among teams during the 2016-17 EPL season based on which plays they run immediately after regaining the ball (left) and on which plays are run against them after loosing possession of the ball (right). Notice teams which are known for some form of counter-pressing such as Tottenham, Manchester City, and Manchester United are located to the far left of Figure 4 (bottom-left).

4.3 Counter-Attacks During the 2016-2017 EPL Season

Since counter-attacks generate both a higher quantity and quality of shots, it would be expected that execution of counter-attacks correspond to success during a season. Table 2 shows the frequency with which each team transitions via a counter-attack during the 2016-17 season. Not surprisingly, top teams like Chelsea, Manchester City, and Manchester United are near the top in both engaging in counter-attacks and limiting the counter-attacks of their opponents. While Arsenal, who finished fifth overall, often was successful in their counter-attacks, they were near the bottom in being counter-attacked against. In contrast, Liverpool, who finished fourth, successfully limited their opponents, but failed to capitalize on their own counter-attacks.

% Counter-Attacks For		% Counter-Attacks Against	
Chelsea	14.90%	Hull City	6.64%
Manchester City	11.93%	Chelsea	7.27%
Bournemouth	11.58%	West Ham United	8.08%
Arsenal	10.90%	Liverpool	8.19%
Manchester United	10.81%	Manchester City	8.59%
Sunderland	10.23%	Manchester United	8.65%
Southampton	9.96%	Southampton	8.84%
Tottenham Hotspur	9.93%	Leicester City	9.25%
Leicester City	9.80%	Burnley	9.49%
Stoke City	9.42%	Everton	10.06%
West Ham United	8.96%	Tottenham Hotspur	10.14%
West Bromwich Albion	7.85%	Stoke City	10.19%
Swansea City	7.55%	West Bromwich Albion	10.72%
Crystal Palace	7.44%	Arsenal	10.77%
Burnley	7.29%	Sunderland	10.98%
Hull City	7.17%	Swansea City	11.68%
Everton	7.12%	Middlesbrough	12.06%
Watford	6.90%	Bournemouth	12.63%
Liverpool	6.15%	Watford	13.07%
Middlesbrough	3.42%	Crystal Palace	14.20%

Table 2: The percent of transitions leading to a counter-attack both for and against each team during the 2016-2017 English Premier League. Many of the top-performers during the season (Chelsea, Manchester City, Manchester United) were successful in both capitalizing on the counter-attack as well as limiting the counter-attack of their opponents.

5. Summary

In this work we have shown how we are able to objectively and automatically identify counter-attacks and counter-pressing without requiring unreliable human annotations. In the processes, we constructed a playbook of identifiable and relevant game-states through a tree-based alignment and clustering algorithm. We then used the templates provided by this playbook to quantify the "defensive disorder" of a team as they transition from offense to defense. Next, we further refined our playbook to identify sub-clusters of plays which were likely to produce goal-scoring opportunities through a measure of "offensive threat".

With this approach we were able to identify counter-attacks directly from the data and subsequently assess their value and impact. Consistent with expectations, counter-attacks generate more, high-quality shots. Furthermore, execution on both the defensive and offensive end of a transition is seen to be reflective of overall performance during a season.

References

[1] P. Lucey, A. Bialkowski, M. Monfort, P. Carr and I. Matthews, "Quality vs Quantity: Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data", in *MIT Sloan Sports Analytics Conference*, 2015.

[2] 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 SSAC*, 2014.

[3] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan and I. Matthews. "Large-scale analysis of soccer matches using spatiotemporal tracking data", in *2014 IEEE International Conference on Data Mining (ICDM)*, 2014

[4] H. Le, P. Carr, Y. Yue and P. Lucey, Data-Driven Ghosting using Deep Imitation Learning, *in MIT SSAC*, 2016.

[5] L. Sha, P. Lucey, Y. Yue, P. Carr, C. Rohlf and I. Matthews, "Chalkboarding: A New Spatiotemporal Query Paradigm for Sports Play Retrieval", in IUI (Sonoma, USA), 2016.

[6] S. Intille and A. Bobick, "Recognizing Planned, Multi-Person Action," *CVIU*, vol. 81, pp. 414–445, 2001.

[7] M.A. Mattar, A.R. Hanson, E.G. Learned-Miller, "Unsupervised Joint Alignment and Clustering using Bayesian Nonparametrics", *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2012.

[8] A.C. Miller, L. Bornn, "Possession Sketches: Mapping NBA Strategies", in *MIT SSAC,* 2016.

[9] C. Parent , S. Spaccapietra , C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M.L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, Y. Theodoridis. "Semantic trajectories modeling and analysis" in *ACM Computing Surveys (CSUR)*. 2013 Aug 1; 45(4):42.

[10] P. Lucey, A. Bialkowski, P. Carr, S. Morgan, I. Matthews, and Y. Sheikh, "Representing and Discovering Adversarial Team Behaviors using Player Roles," in *CVPR*, 2013.

[11] P. Lucey, D. Oliver, P. Carr, J. Roth, and I. Matthews, "Assessing team strategy using spatiotemporal data," in *ACM SIGKDD*, 2013.

[12] 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* (pp. 9-14). IEEE.

[13] X. Wei, P. Lucey, S. Morgan, M. Reid and S. Sridharan, "The Thin Edge of the Wedge": Accurately Predicting Shot Outcomes in Tennis using Style and Context Priors", in *MIT SSAC*, 2016.

[14] I. Bojinov and L. Bornn, "The Pressing Game: Optimal Defensive Disruption" in *MIT SSAC*, 2016.

[15] J. Tenenbaum, V. De Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction". Science 290 (5500).

[16] H. Ruiz, P. Power, X. Wei, and P. Lucey. "The Leicester City Fairytale?: Utilizing New Soccer Analytics Tools to Compare Performance in the 15/16 & 16/17 Seasons", in *KDD*, 2017.

[17] P. Power, H. Ruiz, X. Wei and P. Lucey, "Not All Passes Are Created Equal:" Objectively Measuring the Risk and Reward of Passes in Soccer from Tracking Data", in KDD 2017.