SoccerMix: Representing Soccer Actions with Mixture Models

Tom Decroos ✉, Maaike Van Roy, and Jesse Davis[0000−0002−3748−9263]
{firstname.lastname}@cs.kuleuven.be

Department of Computer Science & Leuven.AI, KU Leuven, Belgium

Abstract. Analyzing playing style is a recurring task within soccer analytics that plays a crucial role in club activities such as player scouting and match preparation. It involves identifying and summarizing prototypical behaviors of teams and players that reoccur both within and across matches. Current techniques for analyzing playing style are often hindered by the sparsity of event stream data (i.e., the same player rarely performs the same action in the same location more than once). This paper proposes SoccerMix, a soft clustering technique based on mixture models that enables a novel probabilistic representation for soccer actions. SoccerMix overcomes the sparsity of event stream data by probabilistically grouping together similar actions in a data-driven manner. We show empirically how SoccerMix can capture the playing style of both teams and players and present an alternative view of a team's style that focuses not on the team's own actions, but rather on how the team forces its opponents to deviate from their usual playing style.

1 Introduction

Style of play, which refers to the behavior on the field of the teams and players during a game, is an important concept in soccer. There is substantial value in gaining a better understanding of playing style as this can be leveraged in areas such as player scouting and match preparation. Because simple descriptive statistics such as pass completion percentage or shot count are usually insufficient to capture playing style, media and fans have traditionally assessed playing style via manual video analysis. However, the advent of novel data sources such as optical tracking and event stream data have motivated an explosion of interest in applying automated techniques to try to glean insights into both player and team behaviors [2, 5, 7–10, 16, 18].

Because it is much more widely accessible than optical tracking data, most techniques focus on analyzing event stream data which describes all on-the-ball actions performed by players during a match. Vendors such as WyScout, StatsBomb, and Opta collect this data using human annotators. While watching video feeds of soccer match, annotators record attributes such as the timestamp, location, type (e.g., pass, dribble, shot), involved player, etc. per on-the-ball action. Depending on the type of the action, the annotator also collects additional information such as the end location of a pass or the outcome of a tackle.

Analyzing the playing style of a team or player based on event stream data often involves constructing a so-called *fingerprint* of that team or player which summarizes their actions and captures distinguishing behaviors such as where on the field they tend to perform certain actions. This is often done by dividing actions into groups of similar actions and counting how often players or teams perform actions within each group. However, assessing similarity is difficult because actions are described by various attributes (e.g., type, location) which lay in different domains (e.g., discrete, continuous).

One approach is to lay a grid over the field and proclaim two actions to be similar when they are of the same type and fall in the same grid cell [5, 7, 16, 17]. However, this approach has three downsides. First, the somewhat arbitrary and abrupt boundaries between grid cells can make certain spatially close actions appear dissimilar. Second, choosing the best resolution for the grid is non-trivial as a coarse grid ignores important differences between locations, while a more fine-grained grid will drastically increase the sparsity of the data as a smaller number of actions will fall in a single grid cell. Third, ideally we would like to group actions on additional attributes such as ball direction, but considering more attributes makes each action more unique, which increases the sparsity in the data. Hence, most approaches only include one or two attributes in their analysis [2, 7, 14]. Rarely do approaches consider three or more attributes [8, 17].

In this paper, we make three contributions. Our first contribution is SoccerMix, a novel mixture-model approach for analyzing on-the-ball soccer actions that addresses the shortcomings of grid-based approaches. On the one hand, it alleviates the problem of sparsity by grouping actions in a data-driven manner. On the other hand, SoccerMix's probabilistic nature alleviates the issues of the arbitrary and abrupt boundaries imposed by grid cells. More uniquely, SoccerMix also considers the direction that actions tend to move the ball in, which is an important property for capturing style of play that has received little attention thus far. For example, it allows distinguishing among players or teams that play probing forward passes versus those that play safer lateral passes in a specific zone of the pitch. Intuitively, the action groups produced by SoccerMix can be thought of as describing *prototypical* actions of a certain type, location, and direction. Our second contribution is that we provide a number of use cases that illustrate how SoccerMix can aid in scouting and match analysis by capturing the playing styles of both teams and players. In contrast to existing approaches which solely focus on the offensive style of a team, SoccerMix can also yield insights into a team's defensive style. Specifically, we model how a team can force its opponent to deviate from its typical style of play. Our third contribution is that we provide a publicly available implementation of SoccerMix.[1]

2 Methodology

Our goal is to capture the playing style of either a player or a team. As in past papers [7, 10, 16, 17], our intuition is that playing style is tied to where on the

[1] `https://github.com/ML-KULeuven/soccermix`

pitch a player (or team) tends to carry out certain types of actions. Most playing style analysis techniques follow the same two-step approach:

Step 1: Partition all on-the-ball actions into groups of similar actions and represent each action by its membership to one or more of these groups.

Step 2: Transform the group membership counts of a player's or team's actions into a human-interpretable summary of playing style.

Traditionally, most research has focused on the second step [5, 7, 8, 16]. However, picking sub-optimal groups in the first step can introduce significant problems such as sparsity down the line. In fact, many sophisticated data aggregation methods such as pattern mining [8] and matrix factorizaton [7] are often only used in step two to combat the problems introduced by the sub-optimal groups established in the first step.

In this paper, we attempt to tackle the first step in a more intelligent manner than before in order to greatly simplify the second step. More specifically, we aim to find groups of similar actions such that players' or teams' group membership counts are already human-interpretable and informative of playing style. This way, no additional sophisticated transformation is needed in step two. Finding these groups of similar actions involves answering four questions:

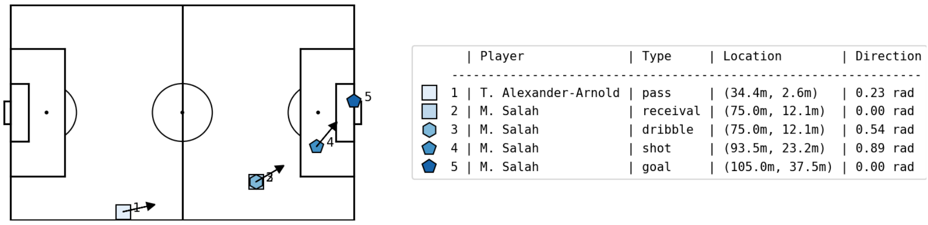
1. Which properties of actions are relevant for capturing playing style?
2. How can we group actions based on both discrete and continuous properties?
3. How can we prevent sparsity (many groups with little or no actions in them)?
4. How can we group actions based on properties with different notions of similarity (e.g., linear data vs. circular data)?

2.1 Describing Actions

Various companies provide event stream data and each one uses a different format, has varying definitions of events, and records different sets of events. Moreover, the data also contains extraneous information such as changes in weather that are not crucial for analysis. The SPADL representation [6] addresses these concerns by converting event streams to a uniform representation designed to facilitate analysis.[2] Hence, we first transform our data into this format.

Typically, playing style analysis focuses on action types and locations. One piece of data that is important for style of play that has received little attention is the direction of actions. For example, it is important to differentiate among players who tend to play probing forward passes versus those that tend to play safer, lateral passes. Unfortunately, the direction of the ball is only implicitly present in the SPADL representation through the start and end locations of actions. Therefore, in this paper, we post process SPADL's output and represent each action as a tuple (t, x, y, θ) where t is the type of the action (e.g., shot, tackle, pass, receival), $x \in [0, 105]$ (meters) and $y \in [0, 68]$ (meters) denote the location on the field where the action happened, and $\theta \in [-\pi, \pi]$ (radians) denotes the direction the ball travels in following the action (Fig. 1).

[2] `https://github.com/ML-KULeuven/socceraction`



```
     | Player              | Type     | Location         | Direction
  ------------------------------------------------------------------
□  1 | T. Alexander-Arnold | pass     | (34.4m, 2.6m)    | 0.23 rad
□  2 | M. Salah            | receival | (75.0m, 12.1m)   | 0.00 rad
⬡  3 | M. Salah            | dribble  | (75.0m, 12.1m)   | 0.54 rad
⬟  4 | M. Salah            | shot     | (93.5m, 23.2m)   | 0.89 rad
⬟  5 | M. Salah            | goal     | (105.0m, 37.5m)  | 0.00 rad
```

Fig. 1. This phase of Liverpool scoring a goal illustrates the event stream data used in this paper. Actions are described by their type t, location (x, y), and direction θ.

2.2 Grouping Actions with Mixture Models

Grouping actions on multiple attributes is non-trivial as it requires fusing together both discrete attributes (i.e., the action type) and continuous attributes (i.e., the location and direction). Past work has mostly ignored direction and focused on fusing action type and location. The most common approach is to lay a grid over the field and for each action type count the number of times it occurs in each zone [7, 16, 17]. However, this approach has two significant problems. First, this approach ignores the fact that some actions only ever occur in certain areas of the pitch (e.g., throw-ins only occur on the outer edges of the field, shots typically only occur on the attacking half of the field). Second, the boundaries between grid cells are arbitrary and abrupt, which can disrupt the spatial coherence. This can make some actions that occurred in nearby locations appear dissimilar because they fall in different location groups.

This paper takes a different approach and uses mixture models to cluster actions. Mixture models are probabilistic models that assume that all the data points are generated from a mixture of a finite number of distributions with unknown parameters [12]. Formally, a mixture model calculates the probability of generating observation x as:

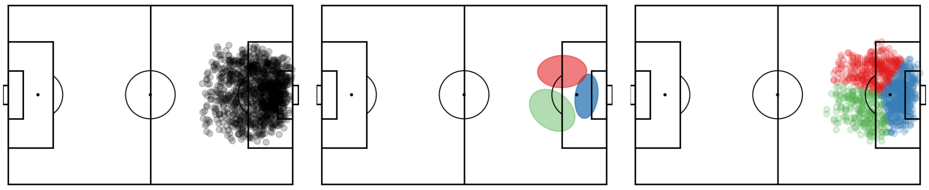
$$p(x) = \sum_{j=1}^{k} \alpha_j \cdot F_j(x|\Theta_j) \tag{1}$$

where k is the number of components in the mixture model, α_j is the probability of the j^{th} component, and F_j is a probability distribution or density parameterized by Θ_j for the j^{th} component. Intuitively, mixture models can be thought of as a soft clustering variant of k-means clustering. Mixture models address all the drawbacks of the grid-based approach. First, they perform a more data-driven as opposed to hand-crafted partitioning of the pitch. This results in a more nuanced partitioning as the mixture model can learn a more fine-grained representation in zones where lots of actions take place and a more course-grained one in zones where actions are less frequent. Second, by performing a soft grouping each action has a probability of belonging to each cluster, which alleviates the arbitrariness of grid boundaries.

SoccerMix hierarchically groups actions with mixture models in two stages:

Stage 1 For each action type, fit a mixture model to the locations (x, y) of the actions of that type. This allows SoccerMix to model that certain action types usually occur in specific areas of the field (e.g., shots only occur close to the goal, see Fig. 2)

Stage 2 For each component of each mixture model in stage 1, fit a new mixture model to the directions θ of the actions in that component. This allows SoccerMix to model that the direction that a specific action tends to move the ball in, depends on the location where the action occurred (e.g., passes in central midfield are usually lateral or backwards, rarely forwards, see Fig. 3).

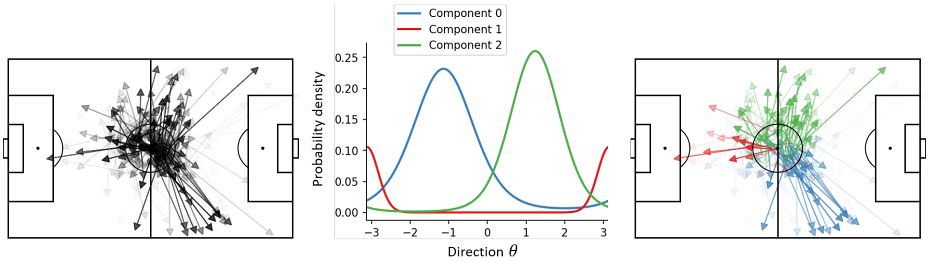


(a) Shot locations

(b) The 2D Gaussian distributions learned on shot locations

(c) Shots colored according to the distributions that most likely generated them.

Fig. 2. Stage 1 of SoccerMix: a mixture model with three 2D Gaussian distributions is fitted to shot locations.



(a) Passes that start in the central midfield

(b) The Von Mises distributions learned on pass directions

(c) The passes colored according to the distribution that most likely generated them.

Fig. 3. Stage 2 of SoccerMix: a mixture model with three Von Mises distributions is fitted to a group of passes that start in the central midfield. In Fig. 3b, component 1 (red) illustrates how a single Von Mises distribution can be fitted to observations close to $-\pi$ and π and is thus essential for describing backwards passes.

2.3 Distributions of Locations and Directions

The next question to consider is which distributions to use as the components of the mixture models. Locations and directions require a different notion of similarity. In the spatial domain, nearby locations are similar, which we can naturally model using a 2D Gaussian distribution (Fig. 2) [15]:

$$pdf(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^2|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{2}$$

where $\boldsymbol{\mu}$ is the mean and $\boldsymbol{\Sigma}$ is the covariance matrix of the distribution.

When viewed as directions, $-\pi+\epsilon_1$ and $\pi-\epsilon_2$ are similar because directions can be seen as values on a circle rather than on a line. However, a Gaussian distribution would not consider these directions to be similar. Therefore, we model the directions using a Von Mises distribution which arises in the directional statistics literature [3, 11]. Unlike a Gaussian, Von Mises distributions allow for the possibility that observations close to $-\pi$ and observations close to π can be generated by the same distribution (Fig. 3). The probability density function of a Von Mises distribution is:

$$pdf(\theta) = \frac{1}{2\pi I_0(\kappa)} \exp\left(\kappa\cos(\theta-\mu)\right) \tag{3}$$

where μ is the mean direction (the distribution is centered around μ) and κ is a measure of concentration ($\kappa = 0$ means that the distribution is uniform over the circle while a high value for κ means that the distribution is strongly concentrated around the angle μ). Finally, $I_0(\kappa)$ is the modified Bessel function of order 0, whose exact definition lies beyond the scope of this paper [11].

2.4 Fitting a mixture model to the data

Fitting the parameters of a mixture model to a data set is typically done using the Expectation Maximization algorithm [1]. Given n observations $\{x_1,\ldots,x_n\}$, k distributions $\{F_1,\ldots,F_k\}$, and nk latent variables r_{ij} which denote how likely it is that distribution F_j generated observation x_i, the algorithm iteratively performs the following two steps:

Expectation For each observation x_i and distribution F_j, compute the responsibility r_{ij}, i.e., how likely it is that F_j generated x_i:

$$r_{ij} = \alpha_j \cdot F_j(x_i|\Theta_j).$$

Maximization For each distribution F_j, compute its weight α_j and its parameter set Θ_j. α_j is the prior probability of selecting component j and can be computed as follows:

$$\alpha_j = \frac{\sum_{i=1}^{n} r_{ij}}{\sum_{j=1}^{k}\sum_{i=1}^{n} r_{ij}}.$$

Θ_j is the parameter set that maximizes the likelihood of distribution F_j having generated each observation x_i with probability r_{ij}. To update Θ_j, we employ the distribution-specific update rules detailed below.

It is straightforward to compute the maximum likelihood estimates for the Gaussian distribution's parameter set $\Theta_j = \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}$:

$$\boldsymbol{\mu}_j = \frac{1}{\sum_{i=1}^{n} r'_{ij}} \sum_{i=1}^{n} r'_{ij} \cdot x_i \tag{4}$$

$$\boldsymbol{\Sigma}_j = \frac{1}{\sum_{i=1}^{n} r'_{ij}} \sum_{i=1}^{n} r'_{ij} \cdot (x_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_j - x_i)^T \tag{5}$$

where r'_{ij} is a normalized responsibility computed as:

$$r'_{ij} = \frac{r_{ij}}{\sum_{j=1}^{k} r_{ij}}.$$

Computing the maximum likelihood estimates for the Von Mises distributions is more challenging for two reasons. First, we use the output of the learned location mixture models as input for the direction mixture models. More specifically, each observation x_i has a respective weight $w_i = \alpha_{loc} \cdot F_{loc}(x_i|\Theta_{loc})$ (where F_{loc} is the location distribution we wish to further decompose) that represents the probability of observation x_i being part of the input set of observations for the direction mixture model. These weights w_i necessitate slightly altering how the responsibilities r_{ij} are normalized. Second, learning the parameters for a Von Mises distribution is inherently harder than for Gaussians. Directly estimating κ_j is impossible as its exact equations cannot be analytically solved. Luckily, an approximation using the mean result distance R_j exists that works remarkably well for many practical purposes (Eq. 7) [11]. We first construct normalized responsibilities r''_{ij} that pretend that each observation x_i in the data set was generated by the mixture model with a probability of w_i and then update the parameter set $\Theta_j = \{\mu_j, \kappa_j\}$ as follows:

$$\mu_j = \text{atan2}\left(\mu_j^{sin}, \mu_j^{cos}\right) \tag{6}$$

$$\kappa_j \approx \frac{R_j(2 - R_j^2)}{(1 - R_j^2)} \tag{7}$$

where

$$\mu_j^{sin} = \frac{1}{\sum_{i=1}^{n} r''_{ij}} \sum_{i=1}^{n} r''_{ij} \cdot \sin x_i \qquad \mu_j^{cos} = \frac{1}{\sum_{i=1}^{n} r''_{ij}} \sum_{i=1}^{n} r''_{ij} \cdot \cos x_i$$

$$R_j = \sqrt{(\mu_j^{sin})^2 + (\mu_j^{cos})^2} \qquad r''_{ij} = w_i \cdot \frac{r_{ij}}{\sum_{j=1}^{k} r_{ij}}.$$

One of the contributions in this paper is that we publicly release our implementation of mixture models at `https://github.com/ML-KULeuven/soccermix`. This implementation supports learning a mixture of any type of distribution from a weighted input set of observations.

2.5 Practical Challenges

When applying SoccerMix to real-world event stream data, three practical challenges arise. First, the locations in event stream data are approximations. For some actions, such as goal kicks, annotators use a set of predefined start locations instead of its actual location. Therefore we add random noise to the locations and directions of actions to ensure that we do not simply recover the annotation rules for some actions. Second, the mixture models are sensitive to outliers (e.g., actions with highly irregular locations). Therefore, we preprocess the event stream data to remove outliers using the Local Outlier Factor algorithm [4]. Third, we need to select the number of components used in each mixture model. The number of components needed depends on the action type. For example, passes need more components than corners; a team can perform passes anywhere on the field, but they can take corners from only two locations (the corner flags). We select the number of components in each mixture model by formulating an integer linear programming problem where the goal is to optimize the total Bayesian Information Criterion (BIC) of the entire set of mixture models.[3]

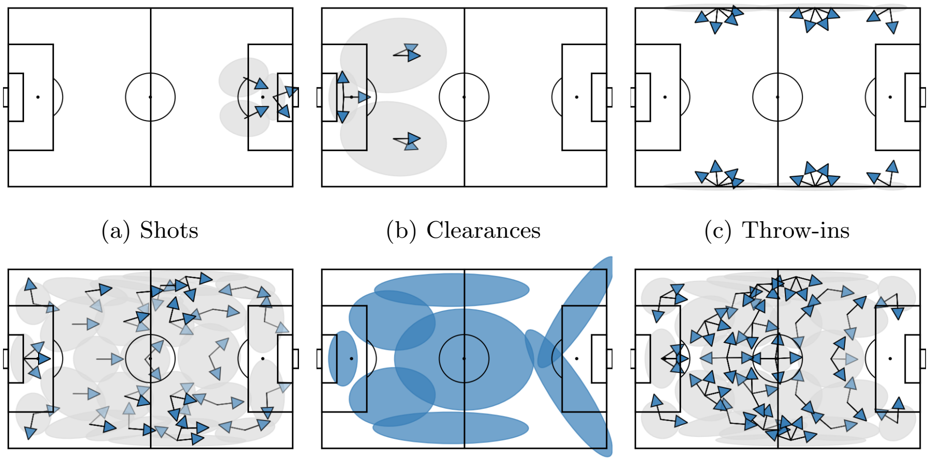
2.6 Capturing Playing Style with SoccerMix

Our goal is to construct a vector that describes a specific player's or team's style. Intuitively, SoccerMix discovers groups of similar actions, where each group describes a *prototypical* action of a certain type, location, and direction. Hence, we can use the learned mixture models to encode each action as a probability distribution over all prototypical actions and encode this in a weight vector. We can then build a style vector for a player (team) by summing the weight vectors of all actions performed by that player (team) in a specific time frame (e.g., a game or a season). In the style vector, the weight of an action group can be interpreted as how often a player (team) performed that prototypical action.

3 Experiments

In our experiments, we use event stream data provided by Statsbomb for the 2017/18 and 2018/19 seasons of the English Premier League (EPL). Using 400,000 actions sampled from the 2017/18 season, we fitted 2D Gaussian mixture models to the locations of the 23 action types to produce 147 location groups. Next, we fitted Von Mises mixture models to the directions of the actions in those groups to produce 247 groups that describe prototypical actions of a certain type, location, and direction (Fig 4). Learning all mixture models took approx. 30 minutes on a computer with 32GB RAM and an Intel i7-6700 CPU @ 3.40GHz with 8 cores. We used these mixture models to produce weight vectors for ± 2,300,000 actions in 760 games and used those to construct style vectors for 676 players and 23 teams.

[3] More details on our approach to select the number of components used in each mixture model can be found in the public implementation.



(a) Shots (b) Clearances (c) Throw-ins

(d) Dribbles (e) Interceptions (f) Passes

Fig. 4. Examples of the prototypical actions discovered by SoccerMix. Ellipses denote 2D Gaussian distributions that describe locations. Arrows denote the center of the Von Mises distributions that describe ball directions. Some action types do not directly move the ball and are thus only grouped on location (e.g., interceptions in Fig. 4e).

In this section, we first show how the style vectors produced by SoccerMix can be used to identify players based on their playing style. Next, we show how to compare the playing styles of teams and players, along with an approach for capturing the defensive style of teams. Finally, we use our style vectors to take a closer look at the game that cost Liverpool the title to Manchester City in the 2018/19 season and investigate what exactly went wrong.

3.1 De-anonymizing Players

No objective definition of playing style exists, which creates challenges. Intuitively, one would expect that in the short-term (i.e., across consecutive seasons) a player's style will not change substantially. Based on this insight, Decroos et al. [7] proposed the following evaluation setup: Given anonymized event stream data for a player, is it possible to identify the player based on his playing style in the previous season?

We perform the exact same player de-anonymization experiment as Decroos et al. and compare SoccerMix to their approach: player vectors based on non-negative matrix factorization (NMF). For both approaches, we used the actions of 193 players that played at least 900 minutes in both seasons. Then, for each player, we create a rank-ordered list of his most similar players by comparing that player's style vector constructed over the 2018/19 season to the style vectors of all players constructed over the 2017/18 season. Table 1 shows how SoccerMix is more successful than the NMF-based player vectors on nearly all ranking metrics. In 48.2% of the cases, SoccerMix correctly identifies a player's style for the current season as being most similar to his previous season's style, which is a 33% relative improvement over the NMF-based approach. Moreover, SoccerMix has a substantially better mean reciprocal rank than the prior approach for this task, which suggests that the style vectors of SoccerMix offer a more complete and accurate view of players' playing style.

Table 1. The top-k results (i.e., the percentage of players whose 2017/18 style vectors are one of the k most similar to their 2018/19 style vectors) and the mean reciprocal rank (MRR) when retrieving 193 players in the English Premier League from anonymized (season 2018/19) and labeled (season 2017/18) event stream data.

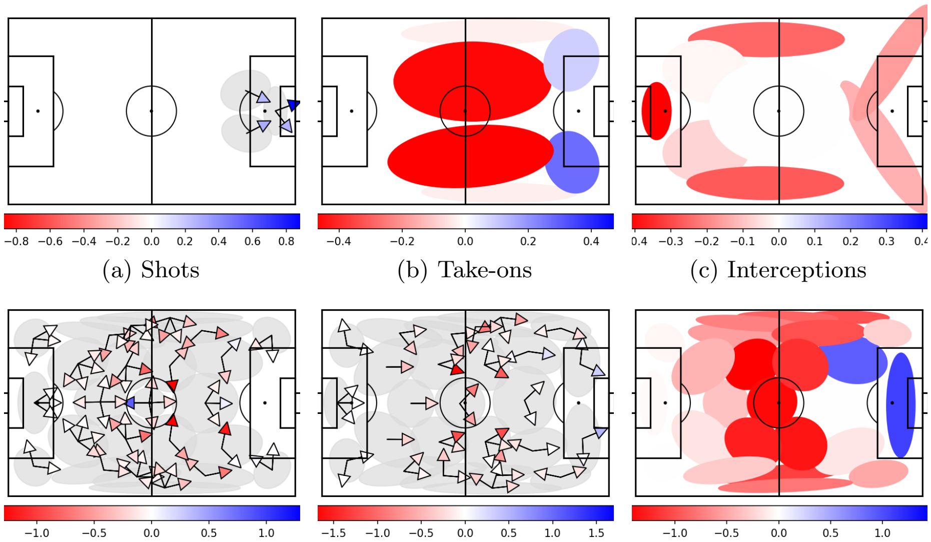
Method	Top-1	Top-3	Top-5	Top-10	MRR
SoccerMix	**48.2%**	**62.7%**	**71.5%**	80.8%	**0.589**
Player Vectors (NMF)	36.5%	53.2%	66.5%	**83.2%**	0.505

3.2 Comparing the Playing Style of Players

The style vectors produced by SoccerMix can be used to illustrate the differences in playing style between two players. As an illustrative use case, consider comparing the playing style of Manchester City forward Sergio Agüero and Liverpool forward Roberto Firmino who are both world-class center forwards playing for top teams. Figure 5 illustrates the differences in their style vector for shots, take-ons, interceptions, passes, dribbles, and receivals during the 2018/19 EPL season. Spatially, Agüero is more active in the penalty box as he performs more take-ons, dribbles, and ball receivals in that area than Firmino. In contrast, Firmino performs these actions more in the midfield. Finally, the interception map shows that while Agüero does not completely neglect his defensive duties, Firmino plays a more expansive role that sees him also intercept the ball on the flanks and near the penalty box. These insights correspond to Agüero's reputation of being an out-and-out striker who camps out near the opponent's penalty box whereas Firmino often drops deep to facilitate for his attacking partners Mohammed Salah and Sadio Mané. SoccerMix allows generating such figures for any two players which has the potential to aid clubs in player scouting as they can identify players whose style fits how they wish to play.

3.3 Comparing the Playing Style of Teams

SoccerMix's style vectors can also be used to compare the playing style of teams. To illustrate this use case, we compare the playing styles of Manchester City and Liverpool, who both completely dominated the 2018/19 English Premier League, finishing at the top of the table with 98 and 97 points respectively with a large 25-point gap to distant third contender Chelsea.



(a) Shots (b) Take-ons (c) Interceptions

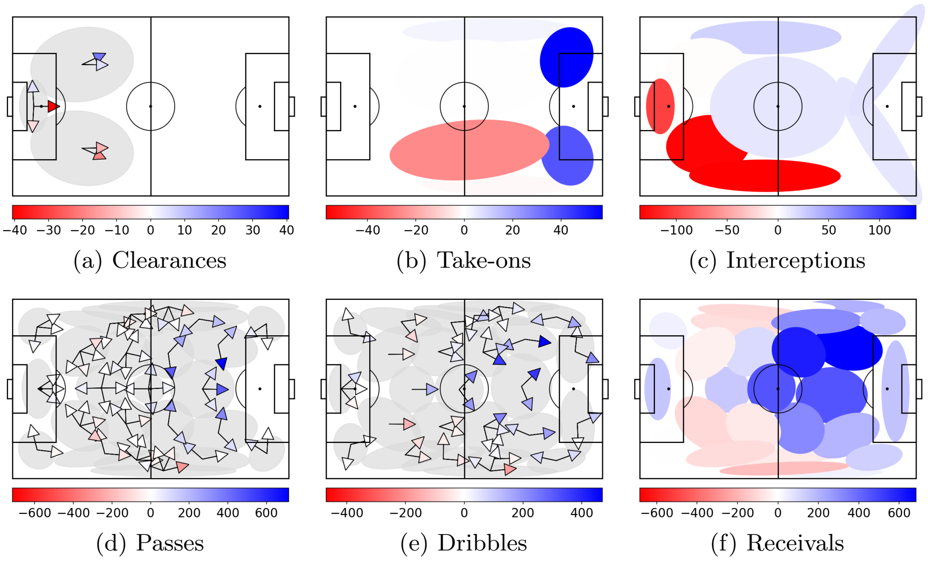
(d) Passes (e) Dribbles (f) Receivals

Fig. 5. Differences in playing style between Manchester City forward Sergio Agüero and Liverpool forward Roberto Firmino during the 2018/19 EPL season. Blue (red) actions indicate that Agüero (Firmino) performed more of these actions than the other. Both players are shown as playing left to right (→). Agüero is more active in the penalty box, while Firmino's actions are more spread out over the midfield.

Figure 6 shows how Manchester City performs noticeably more take-ons, passes, dribbles, and receivals in the heart of the opponent's half compared to Liverpool. This illustrates how the coaches of both teams have shaped their team's playing style to their own soccer philosophy. Under Jürgen Klopp, Liverpool have perfected the art of frequent counter-pressing and speedy counter-attacks. Under Pep Guardiola, Manchester City at times mimics the possession-based, tiki-taka style of its coach's ex-club (FC Barcelona), passing and moving the ball high up on the field.

Additionally, Liverpool seems to funnel the play towards their right side, performing noticeably more clearances, take-ons, and interceptions on their right flank. The most likely source of this uptick is Trent Alexander-Arnold, a right-back at Liverpool who is widely regarded as one of the best attacking full-backs in professional soccer and is a spearhead of Liverpool's transitional, counter-attacking style of play.[4]

[4] `https://sport.optus.com.au/articles/os6422/trent-alexander-arnold-is-changing-the-full-back-position`



(a) Clearances (b) Take-ons (c) Interceptions

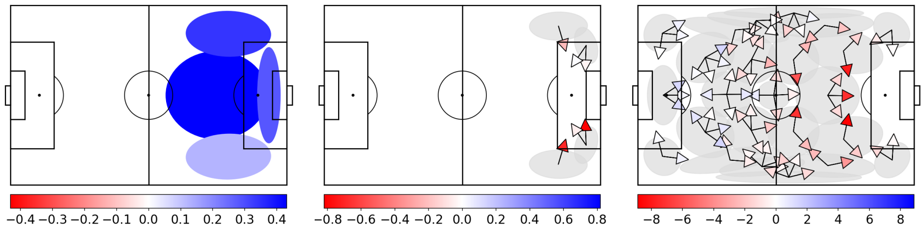
(d) Passes (e) Dribbles (f) Receivals

Fig. 6. Differences in playing style between Manchester City and Liverpool during the 2018/19 EPL season based on the prototypical action groups obtained with SoccerMix. Blue (red) actions indicate that Manchester City (Liverpool) performed more of these actions than the other team. Both teams are shown as playing left to right (→). Liverpool funnels play towards their right side, while Manchester City generally plays higher up the field.

3.4 Capturing the Defensive Playing Style of Teams

Approaches that capture playing style usually focus on offensive playing style, i.e., what does a team do when in possession of the ball? Analyzing defensive style is much harder as it involves off-the-ball actions such as correct positioning and putting pressure on attackers, which are not recorded in event streams. Our insight is that these off-the-ball actions are often performed with the intention of *preventing certain actions from occurring.* This suggests that we can gain a partial understanding of defensive style by measuring the effects that a team's off-the-ball actions have on what on-the-ball actions their opponent performs. More precisely, we analyze how a team forces its opponents to deviate from their usual playing style.

To illustrate this, we measure the mean difference between teams' style vectors constructed using (1) only the matches against Liverpool and (2) all other matches (i.e., those not involving Liverpool). Figure 7 shows how Liverpool causes their opponents, playing left to right, to be flagged more for offside than is typical. This indicates a well-synchronized line of defense that employs a very effective offside trap. The crosses show that, although Liverpool limits the number of crosses its opponents perform, this restriction is not symmetric: they allow fewer crosses from the left of defense (the offense's right) than the right. Lastly, as a combination of both offensive and defensive playing style, Liverpool generally forces the other teams to play more on their own half than on Liverpool's half.



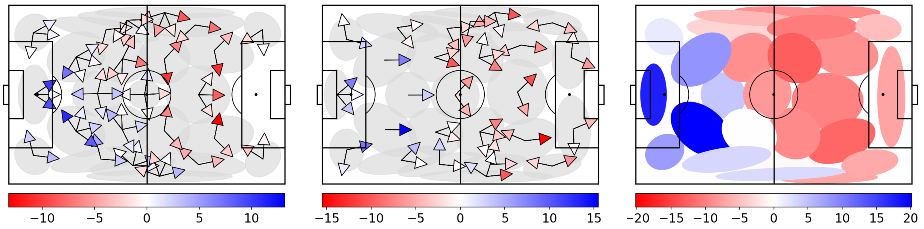
(a) Opponents' offsides (b) Opponents' crosses (c) Opponents' passes

Fig. 7. Illustrations of how Liverpool (a) employs a good offside trap, (b) has a weaker defense at their right flank when it comes to preventing their opponents from crossing the ball, and (c) forces other teams to play more on their own half. Blue (red) indicates that teams perform more (fewer) of these actions when playing against Liverpool.

3.5 Case Study: How Liverpool Lost the Title to Manchester City in a Single Game

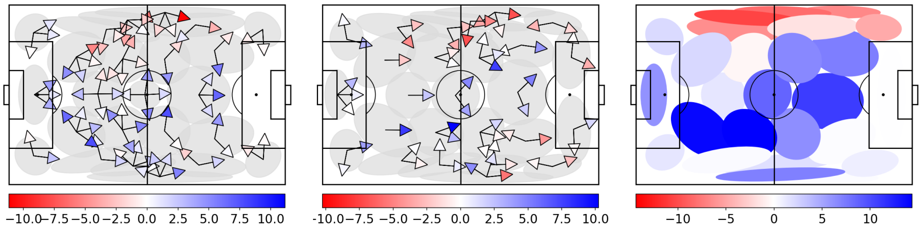
On January 3rd, 2019, Liverpool held a 6 point lead atop the EPL table when they traveled to play Manchester City in a highly anticipated match. Alas, in their only league loss of the season, Liverpool fell 2-1 and ended up missing out on the title to Manchester City by a single point. It is not a stretch to say that this was the game that cost them the title. Using the concept of style difference vectors from the previous section, Figure 8 illustrates how Liverpool's playing style in this game drastically deviated from how they played against other teams. In short, Manchester City maintained their typical high defensive line and forced Liverpool to remain on their own side of the field. This is apparent in both the higher number of passes, dribbles, and receivals Liverpool had to perform deep in their own half as well as the fact that they performed significantly fewer actions than normal in Manchester City's half.

While interesting, it is not completely surprising that Liverpool's offensive output suffered against its only decent rival that season, Manchester City. To dig deeper, we adjust for the level of the opponent and compare Liverpool's playing style in their away game (loss) and home game (draw) against Manchester City in 2018/19 (Fig. 9). In its away game, Liverpool made noticeably less use of its left flank, performing fewer passes, dribbles, and receivals in that area. This suggests that Liverpool's left flank players were not functioning very well that game, which is further evidenced by midfielder James Milner and winger Sadio Mané on Liverpool's left flank being substituted out in the 57th and 77th minute of the game.



(a) Passes (b) Dribbles (c) Receivals

Fig. 8. Differences in Liverpool's playing style during their lost away game against Manchester City compared to their style when playing against all other teams in the 2018/19 EPL. Blue (red) indicates Liverpool performing more (fewer) of these actions in their away game against Manchester City. The direction of play is left to right (→).



(a) Passes (b) Dribbles (c) Receivals

Fig. 9. Differences in Liverpool's playing style between their away game and their home game against Manchester City in the 2018/19 EPL. Red (blue) indicates fewer (more) of these actions in the away game than in the home game. Liverpool's left flank players were having a bad day in the away game, as evidenced by the fewer passes, dribbles, and receivals in that area than normal.

4 Related work

Many approaches group actions by overlaying a grid on the field [7, 8, 16]. How they differ is in how they combat the challenges associated with this grid. Decroos et al. [8] avoid the sparsity issues of a fine-grained grid by dividing the field into only four zones (left-flank, midfield, right-flank, and penalty box), as the performance of their pattern mining algorithm rapidly declined when using a more fine-grained grid. However, their patterns can then only describe ball movements between these four zones and are thus too broad and simple to be able to identify unique characteristics related to playing style. Van Haaren et al. [16] attempted to combine the advantages of both coarse and fine-grained grids by encoding action locations on multiple granularity levels. However, they found that this multi-level representation of actions blew up the search space of their inductive logic programming approach and led to heavy computational costs.

Decroos and Davis [7] apply a post-processing step to the counts of a fine-grained grid. More specifically, the count of each grid cell is replaced by a weighted mean of itself and its neighboring grid cells, which promotes spatial coherence between grid cells and combats issues such as sparsity and abrupt boundaries. However, there are two downsides to this approach. First, a new technique with its own parameters (that are non-trivial to tune) is added to the analysis pipeline. Second, this approach encourages dividing actions into a number of groups that is excessive for representing the characteristics of the data, which makes it difficult for automated systems to numerically process the new data representation and for humans to interpret the end results. For example, Decroos and Davis use 50×50 grid cells to represent shot behavior of players (of which most will be empty), while SoccerMix only needs 3 location groups to represent shot behavior.

5 Conclusion

Capturing the playing style of teams and players in soccer can be leveraged in areas such as player scouting and match preparation. In this paper we introduced SoccerMix: an approach to intelligently partition player actions into groups of similar actions. Intuitively, each group describes a prototypical action with a specific type, location, and direction. We have shown how SoccerMix can be used to capture the playing style of both teams and players. Additionally, we introduced a new way to capture the defensive playing style of a team by using deviations in the actions of that team's opponents. Finally, we have publicly released SoccerMix's implementation at `https://github.com/ML-KULeuven/soccermix`.

Acknowledgements

Tom Decroos is supported by the Research Foundation-Flanders (FWO-Vlaanderen). Maaike Van Roy is supported by the Research Foundation-Flanders under EOS No. 30992574. Jesse Davis is partially supported by KU Leuven Research Fund (C14/17/07), Research Foundation - Flanders (EOS No. 30992574, G0D8819N). Thanks to StatsBomb for providing the data used in this paper.

STATSBOMB
INSIDE EVERY PASS, EVERY SHOT, EVERY MOVE

References

1. Bailey, T.L., Elkan, C., et al.: Fitting a mixture model by expectation maximization to discover motifs in bipolymers (1994)
2. Bekkers, J., Dabadghao, S.: Flow motifs in soccer: What can passing behavior tell us? Journal of Sports Analytics (Preprint), 1–13 (2017)
3. Best, D., Fisher, N.I.: Efficient simulation of the von mises distribution. Journal of the Royal Statistical Society: Series C (Applied Statistics) **28**(2), 152–157 (1979)
4. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD international conference on Management of data. pp. 93–104 (2000)

5. Cintia, P., Rinzivillo, S., Pappalardo, L.: A network-based approach to evaluate the performance of football teams. In: Machine learning and data mining for sports analytics workshop, Porto, Portugal (2015)
6. Decroos, T., Bransen, L., Van Haaren, J., Davis, J.: Actions speak louder than goals: Valuing player actions in soccer. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1851–1861. KDD '19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3292500.3330758
7. Decroos, T., Davis, J.: Player vectors: Characterizing soccer players' playing style from match event streams. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer (2019)
8. Decroos, T., Van Haaren, J., Davis, J.: Automatic discovery of tactics in spatio-temporal soccer match data. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 223–232 (2018)
9. Gyarmati, L., Hefeeda, M.: Analyzing in-game movements of soccer players at scale. arXiv preprint arXiv:1603.05583 (2016)
10. Gyarmati, L., Kwak, H., Rodriguez, P.: Searching for a unique style in soccer. arXiv preprint arXiv:1409.0308 (2014)
11. Mardia, K.V., Jupp, P.E.: Directional statistics, vol. 494. John Wiley & Sons (2009)
12. McLachlan, G.J., Basford, K.E.: Mixture models: Inference and applications to clustering, vol. 38. M. Dekker New York (1988)
13. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
14. Pena, J.L.: A Markovian model for association football possession and its outcomes. arXiv preprint arXiv:1403.7993 (2014)
15. Reynolds, D.A.: Gaussian mixture models. Encyclopedia of biometrics **741** (2009)
16. Van Haaren, J., Dzyuba, V., Hannosset, S., Davis, J.: Automatically discovering offensive patterns in soccer match data. In: International Symposium on Intelligent Data Analysis. pp. 286–297. Springer (2015)
17. Van Haaren, J., Hannosset, S., Davis, J.: Strategy discovery in professional soccer match data. In: Proceedings of the KDD-16 Workshop on Large-Scale Sports Analytics. pp. 1–4 (2016)
18. Wang, Q., Zhu, H., Hu, W., Shen, Z., Yao, Y.: Discerning tactical patterns for professional soccer teams: an enhanced topic model with applications. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 2197–2206 (2015)