Player Vectors: Characterizing Soccer Players' Playing Style from Match Event Streams

Tom Decroos and Jesse Davis

KU Leuven, Department of Computer Science

Abstract. Transfer fees for soccer players are at an all-time high. To make the most of their budget, soccer clubs need to understand the type of players they have and the type of players that are on the market. Current insights in the playing style of players are mostly based on the opinions of human soccer experts such as trainers and scouts. Unfortunately, their opinions are inherently subjective and thus prone to faults. In this paper, we characterize the playing style of a player in a more rigorous, objective and data-driven manner. We characterize the playing style of a player using a so-called 'player vector' that can be interpreted both by human experts and machine learning systems. We demonstrate the validity of our approach by retrieving player identities from anonymized event stream data and present a number of use cases related to scouting and monitoring player development in top European competitions.

1 Introduction

Data analysis is becoming increasingly important in many professional sports [19]. Sports clubs are analyzing huge amounts of data in order to gain a competitive advantage over their opposition. Soccer has been a relative late comer to this trend. The classic statistics about a soccer match (e.g., that appear in boxscores or are often reported on television) tend to be raw counts or fractions, such as the ball possession percentage, number of shots on target or pass success percentage. While interesting, these statistics do not give a complete picture of the match. Moreover they can sometimes obscure important information. For example, the raw number of shots a player took does not tell us about the relative difficulty or quality of the attempts. Recently, research has focused on approaches that allow for a deeper and more insightful analysis of soccer players [4]. One well-known example is the expected goals statistic [9], which is now discussed on the popular soccer talk show *Match of the Day* on BBC One.

A reoccurring concept when discussing soccer is a player's style of play. While both Messi and Ronaldo are great players, each one approaches the game in a different way. Fans often form preferences for players based on his perceived style. From a practical point of view, characterizing playing style is important for professional clubs for three following reasons:

Scouting Soccer clubs can search the market more intelligently if they know the type of player they are looking for and how well prospective targets match that type. Transfers are expensive, and clubs are always looking for bargains and ways to mitigate risks in player recruitment.

Monitoring player development The coach can inspect the playing style of a player in a human-interpretable player vector. If the player vector matches the expectations of the coach, then the coach can monitor that this player vector remains stable and unchanged. If the player vector does not match the expectations of the coach, then he can give his player some pointers and afterwards monitor how well the player is implementing the advice.

Match preparation Understanding the playing style of your opponent can offer certain tactical advantages. The defenders of a team will wish to know what type of attackers they are up against. Similarly, the attackers will be interested in the playing style of the defenders they need to score against.

In this paper, we attempt to characterize a player's playing style in an objective and data-driven manner based on analyzing event stream or play-by-play match data. While playing style is a somewhat subjective concept, our working definition is that a playing style manifests itself as where on the pitch a player tends to perform specific actions with the ball. Our goal is to summarize this playing style in a fixed-length player vector. Characterizing playing style from event stream data is challenging as we have to reason about spatial locations, discrete actions, and a variable number of events. We cope with these challenges by overlaying a grid on the pitch and counting how often each player performs a specific action in a given location. Then, to reduce the dimensionality we perform non-negative matrix factorization. We repeat this for several types of actions. Finally, we concatenate together a player's compressed vectors for each action type to construct his player vector. To evaluate the quality of our player vectors, we propose a retrieval task. Given anonymous event data for a player, we show that we can accurately predict the player's true identify. Moreover, we show how to interpret player vectors and present several qualitative use-cases related to scouting and monitoring player development.

2 Data and Challenges

In this section, we first describe our data set and then highlight a number of data science challenges encountered when analyzing this data.

2.1 Event Stream Data

Event stream data annotates the time and locations of specific events (e.g., passes, shots, dribbles) that occur in a match. This type of data is commercially available from providers such as Opta, WyScout, and STATS. Our data set consists of event stream data from 9155 matches in the five major soccer competitions in Europe: the English Premier League, the German Bundesliga, the Spanish Primera Division, the Italian Serie A and the French Ligue Un. Our data spans almost all matches between the 2012/2013 and 2016/2017 seasons.

Our match event stream data is encoded in the SPADL format [7], which is a format for soccer match event stream data specifically designed to enable automatic data analysis. Some of its benefits are that (1) it focuses exclusively on physical on-the-ball actions (e.g., events such as yellow cards and weather changes are ignored), (2) it works with fixed attributes rather than optional information snippets (which are notoriously difficult to deal with in an automatic analysis pipeline), and (3) it unifies event stream data from different providers into a common format. Each match is represented as a sequence of roughly 1650 player actions. Each action contains (among others) the following five attributes:

Type: the type of action (e.g., shot, tackle, pass),
Player: the player who performed the action,
Team: the player's team,
StartLoc: the (x,y) location where the action started,
EndLoc: the (x,y) location where the action ended.

2.2 Data Science Challenges

Analyzing event stream data from soccer matches is challenging as soccer is a highly dynamic game with many movements and interactions among players across time and space. Concretely, the challenges include the following:

Challenge 1: Characterizing playing style involves coping with the spatial component of the data, as the location where an action happens is important. However, there is very little exact repetition in event stream data. That is, the same player rarely performs the same action in the same location. Characterizing playing style therefore requires us to intelligently generalize over the location of actions.

Challenge 2: Actions have both discrete (e.g., Type, Player, Team) and continuous (e.g., StartLoc, EndLoc) attributes. Most machine learning and data mining techniques prefer to work on either discrete or continuous features exclusively and rarely work well on a mix of both.

Challenge 3: Most standard machine learning techniques require fixed-size feature vectors as their input and cannot natively handle a sequence or set of data points of varying size.

Challenge 4: An important aspect of a player's playing style is how he behaves off the ball (e.g., how much does he run and work to regain the ball?). However, this aspect is almost impossible to measure as action sequence data only describes on-the-ball actions.

3 How to Define and Evaluate Playing Style

Characterizing a soccer player's playing style requires reaching a consensus must be reached on what constitutes a playing style . While this is an inherently subjective concept, our hypothesis is that a player's playing style arises from the interplay between his skills and the tactics employed by the team. Hence, a style of play will manifest itself in the player's behavior during the game.

Definition 1 (Playing style). *A player's playing style can be characterized by his preferred area(s) on the field to occupy and which actions he tends to perform in each of these locations.*

In our work, we also make the following two assumptions.

Assumption 1: Most players exhibit differences in playing style and can be differentiated on this. While it is possible that two players exhibit such a similar playing style that they cannot be discerned from each other, this is not the case for most pairs of players.

Assumption 2: A player's playing style will not drastically change in a short period of time. That is, in a sequence of consecutive games in a season, each player will exhibit the same playing style . This seems justifiable for two reasons. First, while players' skills and playing style evolve over the course of their career, these changes occur gradually rather than abruptly. Second, while the tenure of managers, who influence tactics, do not tend to be overly long in professional soccer, the majority of teams in a league do not change manager mid season.

Based on this definition and these assumptions, any system that successfully characterizes playing style from match event stream data can be used to retrieve players from anonymized event stream data. This player retrieval task can be more formally defined as follows:

Given: Anonymized event stream data describing actions of player p
Retrieve: The identity of player p

The quality of a system that characterizes playing style can be evaluated by its performance on this player retrieval task, as this task measures how well a system can recognize players and differentiate between them purely based on their actions on the field.

In the next section, we describe our system for solving this task. In addition to characterizing each player's playing style , our system also allows human analysts to interpret our representation of playing style and to automatically compare players' playing style on their similarity.

4 Building Player Vectors

In this section we address the following task:

Given: Event stream data describing actions of player p.
Build: A fixed-size player vector that characterizes player p's playing style and can be interpreted both by human analysts and machine learning systems.

At a high-level, our approach works as follows. First, we select relevant action types for characterizing playing style .

Second, for each player p and relevant action type t, we overlay a grid on the field and count how many times player p performed action t in each grid cell. This transform helps address the first three challenges listed in Section 2 because it (1) captures the spatial component, (2) fuses discrete (action type) with continuous (location) features in a unified representation, and (3) converts a variable length set of actions into to a fixed size. We end up with one matrix per player per action type.

Third, we reshape each matrix into a vector and group it together with all other vectors of the same action type to form new, bigger matrices per action type detailing all players' playing style for that specific action type. We then perform non-negative matrix (NMF) factorization to reduce the dimensionality of these matrices. NMF automatically clusters together similar grid cells into a coherent group, which is more informative and intuitive (e.g., for scouting) than looking at individual grid cells where a player operates.

Finally, we construct a player vector for each player by concatenating his compressed vectors of each action type. We also show how to compute the similarity of player vectors (to be used in machine learning algorithms such as clustering and nearest neighbors).

4.1 Selecting Relevant Action Types

Our hypothesis is that the type and location of the actions a player performs are informative of that player's playing style . Our event data contains 19 different types of actions. However, we only consider offensive actions that are performed during open play for two reasons.

First, defense is primarily about positioning, and often this involves picking a position to prevent certain actions from occuring. Hence, by definition, characterizing defensive style requires off-the-ball location data, which we do not have access to. Furthermore, most on-the-ball defensive actions (e.g. tackles, clearances) are usually performed out of necessity rather than because they are indicative of a player's playing style . One effect of this criterion is that all keeper-specific actions are automatically ignored.

Second, the open play criteria means we exclude set piece actions (e.g., freekicks and throw-ins) from our analysis. Teams typically have set-piece specialists (e.g., for free kicks). Similarly, actions like throw-ins are often performed by a pre-defined position (e.g., fullbacks or wingers), so this is more an artefact of position than style. Moreover, these actions can serve as quasi "primary keys" in the player retrieval task, making the task a less effective proxy for characterizing playing style . While we believe analyzing set pieces is extremely interesting and important, a proper study of these actions would require a different type of analysis than we perform in this paper.

When applying these two criteria, the remaining relevant action types are passes, dribbles, crosses, and shots (Table 1).

Table 1. Each action type must fit two criteria to be considered relevant for characterizing playing style : it must be offensive and it must occur during open play. The relevant action type are passes, dribbles, crosses, and shots.

Action type	Frequency	Offensive	Open play
pass	53.1%	✓	✓
dribble	25.2%	✓	✓
clearance	3.8%		✓
throw_in	2.8%	✓	
interception	2.6%		✓
tackle	2.3%		✓
cross	1.8%	✓	✓
shot	1.5%	✓	✓
bad_touch	1.4%		✓
foul	1.3%		✓
freekick_short	1.3%	✓	
keeper_pick_up	0.8%		✓
keeper_save	0.8%		✓
corner_crossed	0.6%	✓	
freekick_crossed	0.2%	✓	
keeper_claim	0.2%		✓
corner_short	0.1%	✓	
shot_freekick	0.1%	✓	
keeper_punch	0.1%		✓

4.2 Constructing Heatmaps

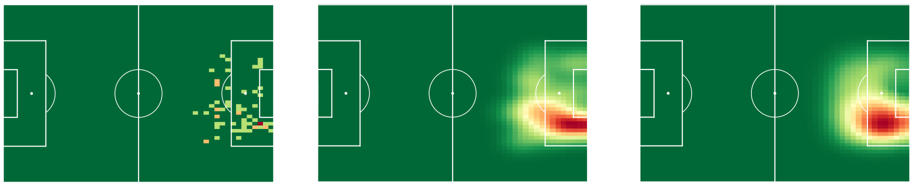
A heatmap is a summary of the locations where player p performs actions of type t. For each player and action type, we execute the following three steps.

1) Counting We overlay a grid of size $m \times n$ on the soccer field. Next, we select all of player p's actions of type t in our data set. Per grid cell X_{ij}, we count the number of actions that started in that cell. Hence, we have transformed a variable-size set of actions to a fixed-size matrix $X \in \mathbb{N}^{m \times n}$ containing the raw counts per cell.

2) Normalizing Two players p_1 and p_2 can have an identical playing style , but if player p_1 played more minutes than player p_2, then player p_1's matrix X will contain higher raw counts than the matrix of player p_2. To combat this, we normalize X such that each cell contains its count if p had played 90 minutes (1 game). For example, if player p played 1600 minutes in total in our data set, then we construct the normalized matrix $X' = \frac{90}{1600}X$.

3) Smoothing We would expect some spatial coherence, or smoothness, in the locations where the actions were performed. However, this coherence can be disrupted by laying a high granularity grid (i.e., high values for parameters m and n) over the pitch as the boundaries between grid cells are abrupt and somewhat arbitrary. Hence, the counts for nearby cells may exhibit more variance than they should. To promote smoothness in the counts of nearby cells, a Gaussian blur is applied to matrix X'. A Gaussian blur is a standard image processing technique [21] that involves convolving X' with a Gaussian function. Specifically, the value of each cell in X' is replaced by a weighted average of itself and its neighborhood, leading to the blurred matrix $X'' \in \mathbb{R}_{+}^{m \times n}$.

X'' is the heatmap detailing where player p performs actions of type t (Figure 1b). For some action types, e.g., passes, we are not just interested in their start locations, but also in their end locations. For these action types, we construct separate heatmaps X''_{start} and X''_{end} using respectively the start and end locations of the actions in the counting step.



(a) The raw shot heatmap obtained by overlaying a grid and counting shot locations

(b) The smoothed shot heatmap obtained after normalizing and smoothing the raw shot heatmap.

(c) The same shot heatmap reconstructed from a length-4 feature vector.

Fig. 1. Example of a heatmap detailing the shot playing style of Riyad Mahrez, winger at Leicester City in the 2016/2017 season.

4.3 Compressing Heatmaps to Vectors

The goal is to capture the information available in a heatmap (i.e., the locations where a player p performs actions of type t) in a small vector. We detail our approach for compressing heatmaps to vectors per action type t.

First, we reshape each heatmap X'' to a 1-dimensional vector x of length mn. In the case of action types where we are interested in both the start and end location, we reshape the heatmaps X''_{start} and X''_{end} to vectors x'_{start} and x'_{end} and concatenate them in a single 1-dimensional vector x of length $2mn$. More generally, let $s = 1$ if we are only interested in the start location of an action type and $s = 2$ if we are interested in both the start and end location of an action type. The length of x is then smn.

We then construct the matrix $M = [x_0 x_1 \dots x_l]$ that contains as columns the reshaped heatmaps of all l players in our data set for action type t. Next, we compress matrix M by applying non-negative matrix factorization (NMF), which is a form of principal component analysis where the resulting components contain only positive numbers. This results in two matrices W and H such that:

$$M \approx WH, \tag{1}$$

where $M \in \mathbb{R}_+^{smn \times l}$, $W \in \mathbb{R}_+^{smn \times k}$ and $H \in \mathbb{R}_+^{k \times l}$. Here, k is a user-defined parameter that refers to the number of principal components for action type t.

The columns of W are the principal components that represent basic spatial groups of action type t. These principal components can be visualized as heatmaps (Figure 2). The rows of H are the small vectors that are the compressed versions of the heatmaps in M. In other words, if the reshaped heatmap x was the i-th column in matrix M, then the i-th row of H is its compressed vector. Each compressed vector can be visualized by multiplying it with the principal component matrix W. The result of this multiplication is a heatmap similar to the original, but reconstructed from only k features (Figure 1). In addition, each feature in a compressed vector is interpretable in the sense that its numeric value quantifies how often the player executes actions of type t with locations in the spatial group of a specific principal component.

4.4 Assembling Player Vectors

The player vector v of a player p is the concatenation of his compressed vectors for the relevant action types: passes, dribbles, crosses, and shots. The total length of a player vector v is equal to $k_{pass} + k_{dribble} + k_{cross} + k_{shot}$ where k_t is the number of principal components chosen to compress heatmaps of action type t. In this paper, we set k_t as the minimal number of components needed to explain 70% of the variance in the heatmaps of action type t. This parameter setting was empirically found to work well because of the high variability of players' positions in their actions (see Challenge 1 in Section 2). Ignoring 30% of the variance allows us to summarize a player's playstyle only by his dominant regions on the field rather than model every position on the field he ever occupied. This design choice lead us to use 4 shot components, 4 cross components, 5 dribble components, and 5 pass components, adding up to form length-18 player vectors.

We can now quantify two player's playing style similarity by computing the Manhattan distance between their player vectors. Manhattan distance works well because the value of each feature in each player vector is a meaningful quantity. The Manhattan distance does not alter this meaning and simply computes the sum of the absolute differences per feature, unlike Euclidean distance which tends to unfairly penalize large differences in a few features. We also empirically confirm that the Manhattan distance works best in Section 5.4.

5 Experiments

Evaluating our method is challenging as no objective ground truth exists for characterizing playing style. Therefore our experiments address three main questions: (1) providing intuitions into what information our player vectors capture,

(2) demonstrating how our approach could be used for scouting and monitoring player development by substantiating a number of claims in popular media about professional soccer players, and (3) measuring our performance at the player retrieval task, which we argue in Section 3 is an effective proxy for how well our approach characterizes playing style .

5.1 Intuition

Figure 2 illustrates all 18 components (4 shots, 4 crosses, 5 dribbles, 5 passes) corresponding to the weights in our length-18 player vectors. The shot, cross, and dribble components only describe where groups of actions start, while the pass components descibe where groups of passes start and end. This is because the end locations of shots, crosses, and dribbles are not informative of a player's playing style . Shots and crosses all end in roughly the same location, while dribbles are usually short and vary in direction such that there is no noticeable difference between their start and end heatmaps.

Figure 3 shows the player vectors of four archetypical players in their 2016/2017 season.

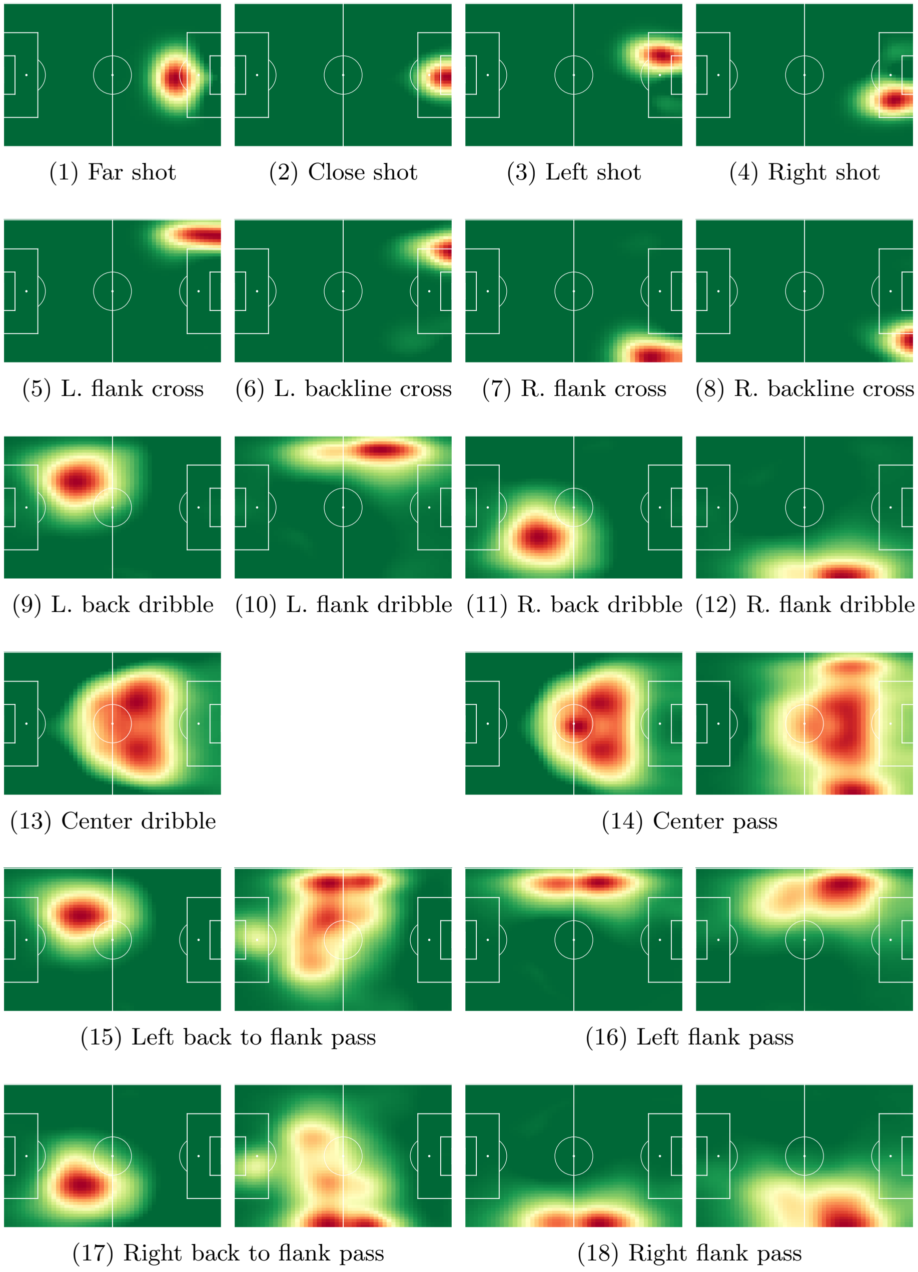
Robert Lewandowski: Striker at Bayern Munich. He shows high weights for three components: *C2: Close shot*, *C13: Center dribble*, and *C14: Center pass*. These are the actions central strikers are expected to focus on.

Jesus Navas: Winger at Manchester City. He shows high weights for three components that are typical of an offensive right winger: *C8: Right backline cross*, *C12: Right flank dribble*, *C18: Right flank pass*.

Kevin De Bruyne Midfielder at Manchester City. De Bruyne's player vector seems less pronounced than the others at first glance, but is actually very informative. First, we can deduce that De Bruyne plays mostly on the opponent's half due to the non-existent weights for components *C9/C11: Left/Right back dribble* and *C15/C17: Left/Right back to flank pass.* Second, his player vector shows similar values for (almost) all mirroring components (e.g., *C16/C18: Left/Right flank pass*). The exception is shots: his weight for *C4: Right shot* is high, while almost non-existent for *C3: Left shot*. De Bruyne's player vector suggests that he is an offensive central midfielder with no preference towards the left or the right when it comes to passing, dribbling or crossing, but attempts to score only from the right.

Sergio Ramos Defender at Real Madrid. Two of his components stand out: *C9: Left back dribble* and *C15: Left back to flank pass*. While less notable than his defensive components, Ramos shows an unusually high weight for *C2: Close shot*. This is because Ramos often attempts to head the ball in the goal at corner kicks, as proven by his nine goals in the 2016/2017 season headed in from corner kicks. Ramos is a left-most central defender with a very defensive playing style , except when it comes to corners.

Player vectors can characterize playing style in an intuitive manner that can make sense to domain experts (e.g., scouts and coaches), yet the interpretable components upon which the player vectors are built are constructed in a completely data-driven manner.



(1) Far shot (2) Close shot (3) Left shot (4) Right shot

(5) L. flank cross (6) L. backline cross (7) R. flank cross (8) R. backline cross

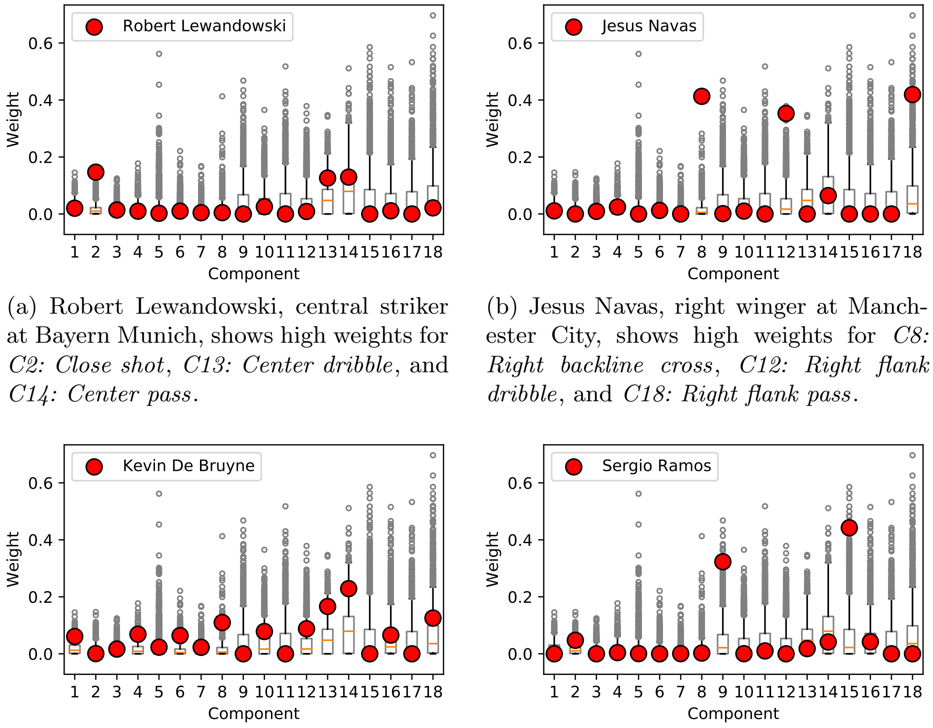
(9) L. back dribble (10) L. flank dribble (11) R. back dribble (12) R. flank dribble

(13) Center dribble (14) Center pass

(15) Left back to flank pass (16) Left flank pass

(17) Right back to flank pass (18) Right flank pass

Fig. 2. The 18 components of our player vectors constructed by compressing heatmaps of shots (1-4), crosses (5-8), dribbles (9-13), and passes (14-18) with non-negative matrix factorization.



(a) Robert Lewandowski, central striker at Bayern Munich, shows high weights for *C2: Close shot*, *C13: Center dribble*, and *C14: Center pass*.

(b) Jesus Navas, right winger at Manchester City, shows high weights for *C8: Right backline cross*, *C12: Right flank dribble*, and *C18: Right flank pass*.

(c) Kevin De Bruyne, central offensive midfielder at Manchester City, shows high weights for all offensive components, favoring neither left nor right

(d) Sergio Ramos, left-most central defender at Real Madrid, shows high weights for *C9: Left back dribble* and *C15: Left back to flank pass*

Fig. 3. Visualized player vectors for an archetypical (a) striker, (b) winger, (c) midfielder, and (d) defender in the 2016/2017 season. The boxplots in the background show the distribution of the weights per component.

5.2 Scouting

We investigate three claims in popular media about similar players. We computed and compared player vectors for all 1480 players who played at least 900 minutes in the 2016/2017 season of the five major soccer competitions in Europe. Lionel Messi is regarded by many as the best soccer player in the world. One player who has been deemed to play similarly to Messi is Paulo Dybala, a fellow Argentinian attacker [23, 13]. When ranked using our player vectors, Dybala is the 2^{nd} most similar player (out of 1479) to Lionel Messi. Idrissa Gueye (midfielder at Everton FC) is often hailed as the new N'golo Kante (midfielder at Chelsea FC) by many journalists [1, 15, 3]. Gueye is the 2^{nd} most similar player to Kante in our data set. Aymeric Laporte is a 24-year-old defender playing for Manchester City FC, who was deemed to be the long-term replacement for 33-year-old Real Madrid defender Sergio Ramos [18, 5], who was named best defender in the world in 2017 by UEFA.[1] Laporte is the 29^{th} most similar player to Ramos using our player vectors. While 29^{th} out of 1479 is not bad, this example does illustrate that our approach is better at characterizing offensive playing style than defensive playing style , as defensive playing style is often more about positioning than on-the-ball actions (see Challenge 4 in Section 2).

5.3 Monitoring Player Development

Journalists agree that Cristiano Ronaldo (ex-Real Madrid) evolved from his role as a left winger to a role as a central striker [22, 20]. Our player vectors capture this transition (Figure 4). In the 2012/2013 season, Ronaldo's most common shot types were *C1: Far shot* and *C3: Left shot*. In the 2016/2017 season however, his shot playing style is completely different with *C2: Close shot* as his most common shot type and no significant difference in output between *C3: Left shot* and *C4: Right shot*. Ronaldo also executed fewer crosses, dribbles and passes in the 2016/2017 season (see the drops in components 5-18), focusing more on finishing scoring chances than setting them up.

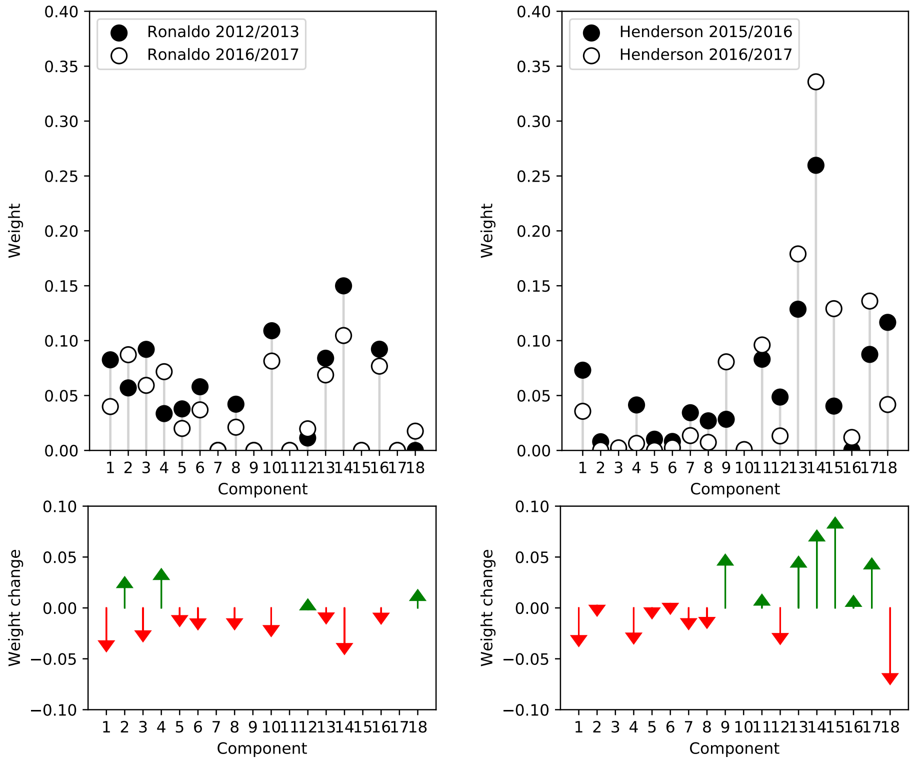
Jordan Henderson is a midfielder at Liverpool. In the 2016/2017 season, coach Jürgen Klopp instructed Henderson to play more defensively, transitioning his playing style from a box-to-box midfielder to a defensive midfielder [25, 17]. When comparing Henderson's 2015/2016 player vector to his 2016/2017 player vector (Figure 4), we notice that his output in terms of passes and dribbles (components 9-18) has significantly increased, while his output in terms of shots and crosses has completely disappeared (components 1-8).

5.4 Player Retrieval from Anonymized Match Event Stream Data

Our approach has many parameters: (a) the size of the grid to construct the heatmaps (50×50), (b) the algorithm to smooth the heatmaps (Gaussian blur), (c) the algorithm to compress the heatmaps (non-negative matrix factorization), (d) the number of components to use (4 shots, 4 crosses, 5 dribbles, and 5 passes), and (e) the distance function to compare the player vectors (Manhattan). Normally we would have no experimental way to tune these parameters as playing style is a subjective concept with no ground truth. However, as explained in Section 3, we can use player retrieval from anonymized match event stream data as a proxy for characterizing playing style .

We solve the player retrieval task as follows. First, we construct a set of labeled player vectors V using a training event stream data set that has not been anonymized. Second, we obtain a set of anonymous actions performed by a target player p_t and construct a player vector v_t based on these actions. Third, we compare v_t to all $v \in V$ and construct a rank-ordered list of the most similar players to p_t. The quality of this ranking is then the position of the unknown player in the ranking. In other words, if most players appear at the top of their

[1] http://www.uefa.com/insideuefa/awards/previous-winners/newsid=2495000.html



(a) Ronaldo evolved from a left winger in the 2012/2013 season to a central striker in the 2016/2017 season. Note the drop of *C1: Far shot* and *C3: Left shot* and the rise of *C2: Close shot* and *C4: Right shot*.

(b) Henderson transitioned to a more defensive playing style after the 2015/2016 season. Note the almost complete disappearance of shots and crosses (components 1-8) and the rise of passes and dribbles (components 9-18).

Fig. 4. Player vectors illustrating the development of (a) Cristiano Ronaldo, former striker at Real Madrid, and (b) Jordan Henderson, midfielder at Liverpool.

own rankings, then we have successfully characterized playing style. If most players do not appear near the top of their own rankings, then we have failed.

To illustrate this idea, we provide the results of an experiment to test whether Manhattan distance or Euclidean distance is the best distance function for comparing player vectors in Table 2. In our experiment, our training data was labeled event stream data from season 2015/2016 of the five top soccer competitions in Europe and the test data was anonymized event stream data from season 2016/2017 of the same competitions. We only considered players that have played 900 minutes in the same team in both seasons. This left us with 741 anonymized players in the test data which we de-anonymized using 741 labeled players in the training data.

Table 2. Top-k results and mean reciprocal rank (MRR) when trying to retrieve 741 players from anonymized event stream data of season 2016/2017 using labeled event stream data from season 2015/2016.

Distance function	Top-1	Top-3	Top-5	Top-10	MRR
Manhattan distance	**38.2%**	**49.8%**	**54.9%**	**64.4%**	**0.469**
Euclidean distance	33.0%	47.0%	52.9%	61.8%	0.429

The Manhattan distance outperforms the Euclidean distance at retrieving players from anonymized event stream data. We can successfully retrieve 38.2% of all players with only one attempt and retrieve 64.4% of all players in the top-10 of our rankings. Hence, we conclude that Manhattan distance is the better distance function to use to compare players' playing style .

6 Related Work

Danneels et al. [6] predict a player's position (i.e., attacker, midfielder, defender) based on their actions. While similar to our research, our goal and approach is more broad and ambitious, as our player vectors are much more detailed than only three distinct labels. Gyarmati et al. [14] construct movement vectors to characterize a player by his movement on the field. Van Gool et al. [24] analyze the playing style of teams instead of players. Their approach is different from ours, but their goal is similar as they also try to capture a subjective concept like playing style in a more objective and data-driven way. STATS introduced *STATS Playing Styles* [11], which are eight different styles (e.g., fast tempo, direct play, counter attack) teams use to create shooting opportunities. Fernandez et al. [10] also categorize different styles of play for teams in professional soccer.

Pappalardo et al. [16] and Decroos et al. [7] evaluate a player's quality of performance. Bransen et al. [2] measure players' resilience to mental pressure. The biggest difference between these works and ours is that our paper aims to characterize a player's playing style , with less emphasis on the player's quality of play. One way we could improve our approach is to expand our player vectors with features that capture the tactics a player is involved in (e.g. [8]).

In other sports, Franks et al. [12] used spatial information to categorize shots in professional basketball. In this work, data from the NBA was collected and analyzed using non-negative matrix factorization (NMF). This paper was a huge influence on our work, as our approach on soccer event data is largely inspired by their approach on basketball event data.

7 Conclusion

Objectively characterizing the playing style of professional soccer players has important applications in scouting, player development monitoring, and match preparation. We showed how to construct player vectors by transforming sets of actions from match event stream data to fixed-size players vectors using non-negative matrix factorization. These player vectors offer a complete view of a player's playing style (within the limits of the data source), are constructed in a purely data-driven manner, are human-interpretable and can be used in machine learning systems such as clustering and nearest neighbor analysis.

Acknowledgements

Tom Decroos is supported by the Research Foundation-Flanders (FWO-Vlaanderen). Jesse Davis is partially supported by KU Leuven Research Fund (C14/17/07, C32/17/036), Research Foundation - Flanders (EOS No. 30992574, G0D8819N) and VLAIO-SBO grant HYMOP (150033).

References

1. Adewoye, G.: Everton boss Sam Allardyce compares Idrissa Gueye to N'Golo Kante, http://www.goal.com/en/news/everton-boss-sam-allardyce-compares-idrissa-gueye-to-ngolo/gddgazktcl3b1ayeadrva1o18
2. Bransen, L., Robberechts, P., Van Haaren, J., Davis, J.: Choke or shine? quantifying soccer players' abilities to perform under mental pressure. MIT Sloan Sports Analytics Conference (2017)
3. Callaghan, S.: Everton boss was spot-on with Idrissa Gueye - N'Golo Kante comparison (2018), http://www.hitc.com/en-gb/2018/04/12/everton-boss-was-spot-on-with-idrissa-gueye-ngolo-kante-comparis/
4. Coles, J.: The Rise of Data Analytics in Football: Expected Goals, Statistics and dam (2016), http://outsideoftheboot.com/2016/07/21/rise-of-data-analytics-in-football/
5. Collins, T.: 4 Possible Replacements Should Real Madrid Sell Sergio Ramos (2015), http://bleacherreport.com/articles/2509541-4-possible-replacements-should-real-madrid-sell-sergio-ramos#slide3
6. Danneels, G., Van Haaren, J., Op De Beck, T., Davis, J.: Identifying playing styles in professional football. KU Leuven (2014)

7. Decroos, T., Bransen, L., Van Haaren, J., Davis, J.: Actions speak louder than goals: Valuing player actions in soccer. arXiv:1802.07127 (2018)
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. Eggels, H.: Expected Goals in Soccer: Explaining Match Results using Predictive Analytics (2016), eindhoven University of Technology
10. Fernandez-Navarro, J., Fradua, L., Zubillaga, A., Ford, P.R., McRobert, A.P.: Attacking and defensive styles of play in soccer: analysis of spanish and english elite teams. Journal of sports sciences **34**(24), 2195–2204 (2016)
11. Flynn, M.: STATS Playing Styles An Introduction (2016), www.stats.com/industry-analysis-articles/stats-playing-styles-introduction
12. Franks, A., Miller, A., Bornn, L., Goldsberry, K.: Characterizing the spatial structure of defensive skill in professional basketball. Annals of Applied Statistics 2015, Vol. 9, No. 1 pp. 94–121 (2015), arXiv:1405.0231
13. Goal.com: Messi admits difficulties in Dybala partnership: He plays like me at Juve, http://www.goal.com/en/news/messi-admits-difficulties-in-dybala-partnership-he-plays-like-me-/1uq96ju5zageb1s1vez93omsi3
14. Gyarmati, L., Hefeeda, M.: Analyzing in-game movements of soccer players at scale. arXiv preprint arXiv:1603.05583 (2016)
15. Kleebauer, A.: Everton's Idrissa Gueye is the new N'Golo Kante - and here are the stats to prove it (2017), https://www.liverpoolecho.co.uk/sport/football/football-news/evertons-idrissa-gueye-new-ngolo-12965076
16. Pappalardo, L., Cintia, P., Ferragina, P., Pedreschi, D., Giannotti, F.: Playerank: Multi-dimensional and role-aware rating of soccer player performance. arXiv preprint arXiv:1802.04987 (2018)
17. Pierce, J.: Henderson: I'm learning fast in the new midfield role Klopp's given me (2016), https://www.liverpoolecho.co.uk/sport/football/football-news/henderson-im-learning-fast-new-11862193
18. Prenderville, L.: Sergio Ramos 'identifies Aymeric Laporte and Matthijs de Ligt as his long-term replacements' at Real Madrid (2017), https://www.mirror.co.uk/sport/football/transfer-news/sergio-ramos-identifies-aymeric-laporte-11710624
19. Pritchard, S.: Marginal gains: the rise of data analytics in sport (2015), https://www.theguardian.com/sport/2015/jan/22/marginal-gains-the-rise-of-data-analytics-in-sport
20. Romero, A.: Cristiano Ronaldo: the change to a 'number 9' (2016), https://en.as.com/en/2016/12/19/opinion/1482164003_264275.html
21. Shapiro, L., Stockman, G.C.: Computer vision. 2001. Ed: Prentice Hall (2001)
22. Sharma, R.: How Cristiano Ronaldo has been transformed from a winger into a deadly No 9... and why he could really play for Real Madrid into his 40s (2017), http://www.dailymail.co.uk/sport/football/article-4469198/How-Ronaldo-transformed-winger-deadly-No9.html
23. Smith, R.: Is Paulo Dybala the Next Lionel Messi? "He Can Go as High as He Likes" (2017), https://www.nytimes.com/2017/04/10/sports/soccer/paulo-dybala-juventus-lionel-messi-barcelona.html
24. Van Gool, J., Van Haaren, J., Davis, J.: The automatic analysis of the playing style of soccer teams. KU Leuven (2015)
25. Williams, G.: Jordan Henderson is relishing his new role in the Liverpool midfield (2016), https://www.liverpoolecho.co.uk/sport/football/football-news/liverpool-jordan-henderson-jurgen-klopp-12123785