Qualitative spatial reasoning for soccer pass prediction

Vincent Vercruyssen, Luc De Raedt, and Jesse Davis

KU Leuven, Department of Computer Science
Celestijnenlaan 200A, 3001 Leuven, Belgium

Abstract. Given the advances in camera-based tracking systems, many soccer teams are able to record data about the players' position during a game. Analysing these data is challenging, since they are fine-grained, contain implicit relational information between players, and contain the dynamics of the game. We propose the use of qualitative spatial reasoning techniques to address these challenges, and test our approach by learning a model for pass prediction over a real-world soccer dataset. Experimental evaluation shows that our approach is capable of learning meaningful models. Since we employ an inductive logic programming system to learn the model, it has the added benefit of producing interpretable rules.

Keywords: Sports analytics, Qualitative spatial reasoning, Pass prediction

1 INTRODUCTION

Many professional soccer teams are beginning to employ specialized camera-based tracking systems during matches that are able to record precise locations of the players and ball multiple times per second [17, 3]. The analysis of these data can provide important insights into a team's strategy as well as player's strenghts and weaknesses. Currently, the primary focus of automated data analysis, particularly within the clubs themselves, is on computing descriptive statistics. In contrast, strategic analysis and player evaluation largely rests on the time-consuming process of a manual film evaluation. Clearly, automated technicals that can capture complex interactions among players have the potential to provide valuable insights into the sport. Consequently, there has been an explosion of interest in automating the analysis of sport match data [13, 14, 21, 15, 1].

The analysis of information-rich spatio-temporal data poses a number of interesting and significant challenges from a learning and knowledge discovery point-of-view. First, the same group of players rarely performs an identical sequence of actions in the same positions within the same time span. This problem is compounded by the fact that the size of a soccer field is not fixed and can vary slightly from stadium to stadium. Second, players will act based on how they are positioned with respect to other players on the field. This means that the analyst needs to take into account the relational aspects that hold between

the individual spatio-temporal data streams of different players. Third, soccer is inherently dynamic and accurately modelling the game requires accounting for these dynamics.

In this paper, we explore the use of qualitative spatial reasoning techniques to address the previous challenges. By focusing on the qualitative relationships that hold between objects, qualitative spatial representations (QSR) provide (1) a mechanism to abstract away the quantitative aspects of a player's location, and (2) a way to combine separate data streams. However, while most QSRs excel at expressing relations between objects stationary in time (henceforth referred to as *static* information), it is not obvious how they represent the *dynamics* and *transition* effects between two distinct static states. Therefore, a contribution of this work is to explore several ways to incorporate the important dynamic information into existing QSRs.

To test our approach, we focus on the specific task of predicting to whom a player will give a pass, based on the game information available to that player up until the moment of the pass. We describe each game state using the QSR predicates and use inductive logic programming to learn a pass prediction model. We perform empirical evaluation on a real-world dataset consisting of 14 matches from a Belgian professional soccer club. We conclude that, using QSRs, we are able to extract meaningful information from the spatio-temporal data that allow us to build pass prediction models. Additionally, we find that adding dynamic and transition information to the prediction models increases their performance.

2 RELATED WORK

The observed trend towards a larger volume of available game data goes hand in hand with the increase in analytical studies/tools for soccer. A lot of these focus on providing automatic summerization and analysis on the game [6, 19]. A number of studies employ spatio-temporal player data to construct more advanced analysis tools [10]. Various techniques have been developped. For instance, [8, 9] have done work towards detecting movement patterns through the use of self-organising maps and neural networks. Research has also been done on defining the dominant region of a player on the field, which is the region he can reach before every other player on the field [20]. Lucey et al show that individual player movement can serve the purpose of assessing the overal team strategy [15]. Related to our task, yet not the same, is the classifcation of different passes into categories based how *good* they are [11]. However, the authors produce quantitative measures of performance.

3 PRELIMINARIES

We briefly review the relevant background on inductive logic programming (ILP) [16], and qualitative spatial representations (QSRs) [2].

3.1 First-order logic and inductive logic programming

This paper considers a subset of first-order logic, where the alphabet consists of only three symbols. *Constants* start with a lower-case letter and refer to specific objects (e.g., a player $\texttt{p}_\texttt{i}$). *Variables* start with an upper-case letter and range over objects (e.g., `Players`). *Predicates* represent relations between objects (e.g., a pass between two players $\texttt{pass}(\texttt{p}_\texttt{i}, \texttt{p}_\texttt{j})$). A *literal* is either $\texttt{p}(\texttt{a}_1, ..., \texttt{a}_\texttt{n})$ or $\neg\texttt{p}(\texttt{a}_1, ..., \texttt{a}_\texttt{n})$, where the $\texttt{a}_\texttt{i}$ are constants or variables. A *definite clause* is a disjunction over finite sets of literal containing exactly one positive literal. Definite clauses are often written in implication form $B \implies H$, where B is a conjunction of literals and H is a single literal.

ILP is a well-known framework for learning models, in the form of definite clauses, from relational data [5]. In this work, we use the widely-used and publicly available Aleph ILP system [18]. Aleph can be used in a variety of different ways, but at its core it learns one definite clause at a time by searching through the space of a possible definite clauses for a given target concept.

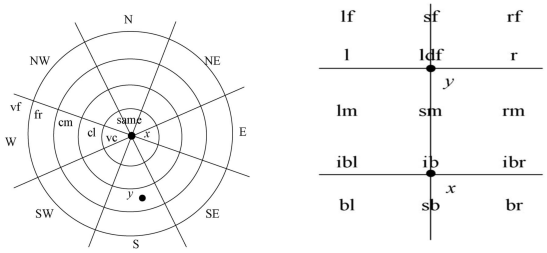
3.2 Qualitative spatial representations

QSRs are formalisms, called calculi, that define how entities in a 2D or 3D space behave relative to eachother; see Chen et al. for an extensive survey [2]. The key idea is to represent spatial relationships with qualitative statements (e.g. player x is closer to player y than to player z, or player x and y move in the same direction...). QSRs allow for relations between primitive spatial entities (points, lines, planes) or extended spatial entities (simple and complex regions). For each QSR, the set of possible relations is finite usually also a *joint exhaustive and pairwise disjoint* (JEPD) set. Relations are mostly binary. However, they can be of a higher arity. Numerous categories of QSRs exist, we are mostly interested in those formalisms related to: distance, direction, mereotopology[1], and movement.

Several calculi have been designed to define relations between objects in space based on how they are located located with respect to each other. The most common being the *cone-shaped direction calculus* (see Figure 1a) [7]. It orients directionality around a reference object by using directions such as: `north`, `east`... The *double-cross calculus* (shown in Figure 1b) provides a way to express how a player z is located with respect to points x and y [12]. Such calculi can be extended to include distance information such as: `far`, `close`, `very_close`... Most calculi require the designation of a target entity (e.g., a player), a reference entity (e.g., the passer), and a reference frame (e.g., the pitch) to construct the relations.

Expressing relations between larger spatial entities becomes possible when using the *region connected calculus* (RCC) [4]. RCC8 (shown in Figure 2a) is a reduced version of RCC that allows for eight JEPD base relations between two

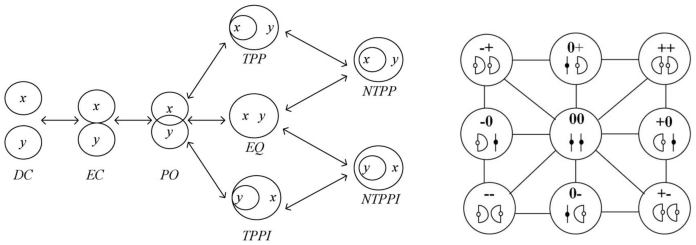
[1] Mereotopology is the integration of mereology, or the study of parts and the wholes they form, and topology.



(a) Extended cone-based. (b) Double-cross.

Fig. 1. Various formalisms exist to describe directional relations between objects.

objects. We could use it to encode whether a player is free or not by observing the area around him and relating it to other players.



(a) Region connected calculus. (b) Qualitative trajectory calculus.

Fig. 2. The RCC provides a way to describe topological relations between regions, while the QTC captures the movement of objects with regard to each other.

We use the qualitative trajectory calculus (QTC) to map relative motions between players to a number of qualitative values denoting: moving towards eachother, moving away from eachother, and maintaining the same distance (see Figure 2b) [22].

4 METHODOLOGY

Our goal is to explore the use of QSRs to extract various kinds of information from spatio-temporal data within the context of soccer pass prediction. Formally, we will define this problem as follows:

Given: a temporal sequence of the spatial configurations of players during a soccer match, and a corresponding stream of event data, up until the present time `t` when player $\mathtt{p_i}$ has the ball and is about to give a pass

Predict: to which player $\mathtt{p_j}$ the ball will be passed and $j \neq i$.[2]

4.1 Considered qualitative relationships

The key challenge is to define *predicates* that accurately model the important characteristics of the spatio-temporal and event stream data such that the ILP system can make correct inferences about who is the target of the pass. It is helpful to consider which aspects of the game a player takes into account before deciding to whom he will pass. Looking at the data (see Figure 3), it seems that a player considers the direction of movement, the position, and the speed of other players. Additionally, a player might make assumptions about the ability of his team to receive the ball, or how likely it is a player of the opposing team will intercept the pass. We define three categories of *features*, and relate them to the existing QSRs:

Static qualitative relationships. In the static setting, the data only consists of a snapshot of the game state at the moment the pass was attempted. Specifically, we employ the extended cone-based calculus to describe the relative position of players versus some reference point (chosen to be the passer and receiver). The double-cross calculus is used to describe the positioning of players with respect to each possible passline (i.e., the line connecting the passer and each possible receiver at the moment of the pass).

Dynamic relationships. In the dynamic setting, we also consider the moments leading up to the pass, which allows us to construct predicates that capture game dynamics. First, we construct a movement vector for each player. Second, we apply calculi such as the QTC to these vectors to infer predicates describing how players move across the field with respect to each other. Third, inspired by the concept of a dominant player region [20], we use the information encoded in the movement vector to estimate the region of the pitch the player can reasonably reach within a certain amount of time. While [11] uses this concept to calculate non-overlapping, player dominated-regions on the field, and infers quantitative features from this, we allow for overlapping regions and describe qualitative relations between using RCC.[3] Finally, we extract information related to the line of sight of a player, given the assumption that the latter coincides with the direction of the movement vector.

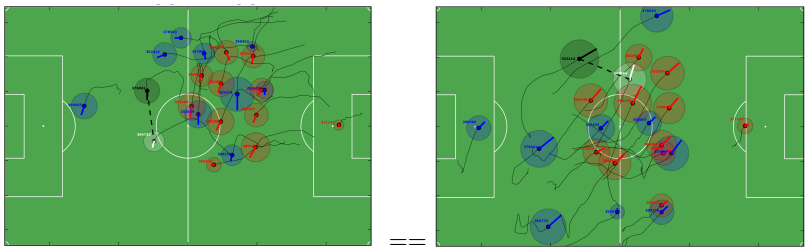
Transition features. We argue that it is interesting to capture the info embedded in the transition between two states of a spatial calculus. Concretely, if predicates $\mathtt{p_1}(\mathtt{a_1}, ..., \mathtt{a_n})$ and $\mathtt{p_2}(\mathtt{a_1}, ..., \mathtt{a_n})$ express the state of the calculus respectively at times $\mathtt{t_1}$ and $\mathtt{t_2}$, we derive a new predicate $\mathtt{p_{trans}}$ that describes this transition. For the cone-based calculus, $\mathtt{p_{trans}}$ encodes both the change

[2] Consider that the resulting model only determines to whom a pass is given, not when it is given, which is an entirely different issue.

[3] We construct these regions by defining a circle around the player with its size in accordance with the movement vector. According to [10], as long as the speed of the player is below 14.5 km per hour, this approximation is a reasonable assumption.

in directionality and distance. For the RCC, $\texttt{p}_{\texttt{trans}}$ encodes the changing mereotopology between two regions.

Background knowledge. We augment previous features with background knowledge about the role of each player during a match. We observe that the implicit discretization of the continuous feature space employed by the spatial calculi, will often be either too coarse or too fine-grained. Hence, we introduce a hierarchy of relations within each QSR. On each level, the discretizations are jointly exhaustive, while a discretization on a higher level encompasses a subset of those on the level below. For instance, in the extended cone-based calculus, we add a hierarchy to both direction and distance.



(a) Pass between defenders. (b) Pass to the front.

Fig. 3. Snapshots of two pass events. The passer and receiver are respectively represented by the black and white dot, and they are both part of the blue team. The thin line indicates the trajectory of each player in the moments leading up to the pass, while the thicker bar represents the movement vector.

4.2 Model learning

We consider the problem as a ranking problem, since in many game states, several team members might constitute an equally good option to pass to. After constructing the features of each pass instance, we use the Aleph system with the `inducemax` search strategy to learn a theory/model (which consists of a number of rules). Next, we test this theory on unseen game states. Each possible receiver is then assigned a *pass probability* based on the number of rules from the theory that cover the player. Using this probability, one can construct a ranking between players. The most likely receiver according to the model is ranked first.

5 EXPERIMENTS

The experimental set-up aims at answering following research questions:

1. Does the qualitative approach yield an advantage over the quantitative?

2. Does adding the dynamic and transition information lead to a better prediction model, as opposed to using only static information?
3. Can we use the learned model to test interesting soccer-related hypotheses on the data?

5.1 Data and experimental set-up

We possess data for 14 soccer matches from a team in the Belgian Pro League. The data consists of three major parts: event data (e.g., pass, shot on target...), the exact positions of each player on the pitch during the game, sampled with a frequency of $10Hz$, and limited background information (see Table 5.1). From this dataset, we select all successful passes for a team. We ignore unsuccessful passes as the data do not contain information about the intended target when a pass is intercepted, so we have no way of checking whether the prediction of our model is correct. Will this influence the results? It seems likely that an intercepted pass is proportionally more risky than its completed counterpart. The latter ensures that the subset of passes we work with, spans the more conservative ones. Consequently, when predicting the pass target in an unseen game situation, a model trained over this subset will be biased towards selecting the *safer* player. However, since we are careful in selecting both training and test data from the same subset, this should not influence the results. Finally, we also omit goal kicks. Each snapshot leads to one positive example (the player who received the pass) and nine negative examples (the players who did not receive the pass). We use 5-fold cross-validation. To evaluate the ranking, we use the mean reciprocal rank measure (MRR):

$$MRR = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{rank_i} \in [0,1] \tag{1}$$

where n is the total number of examples, $rank$ is the rank of the actual receiver in each example and $rank \in [1,10]$. Due to the underlying exponential function, the MRR increasingly rewards the model the higher the ranking of the actual receiver. We also report recall in the top-1, top-2, and top-3 of the ranking.[4]

Table 1. The table gives an overview of the data contained in the original database. The different tables are listed, as well as a compressed list of the attributes for each of these tables.

Tablename	Index	Attributes			
Tracking	*time*	x-position	y-position	player ID	events half
Players	*player ID*	team ID	name	position	jersey nr
Events	*event ID*	time	players	description	position

[4] This constitutes the percentage of times the real receiver is ranked accordingly.

5.2 Results and discussion

Research question 1 and 2. We learn models in six different settings: static setting with non-relational quantitative features, static setting with relational quantitative features, static setting with relational qualitative features, dynamic setting with relational qualitative features, transition setting with relational qualitative features, and the last three settings combined. The quantitative approach employs exact distances and angles between players, allowing it to learn thresholds over them. We expect that the relational approach will be better than the non-relational approach, and that the qualitative model outperforms its quantitative counterpart.

Table 5.2 contains an overview of the results. First, the relational approach is vastly superior to the non-relational. Additionally, we observe a strong improvement in terms of MRR and recall when moving from a quantitative to a qualitative model. Models learned from dynamic and transition predicates also outperform the quantitative model, but do worse than the static models. The combined model performs best and improves both recall and MRR compared to the purely static model. Almost half of the time, the combined model ranks the actual receiver in the top 3 of most likely players to receive the ball (out of 10).

Table 2. Performance of quantitative versus qualitative models for soccer pass prediction. MRR is the mean reciprocal rank, while e.g. top-3 shows the percentage of times the actual receiver is ranked accordingly by the learned model. Each model is trained and tested on pass data of one team.

		MRR	top-1	top-2	top-3	Rules
Quant.	Non-rel.	0.11	0.84	0.93	0.93	8
	Rel.	0.24	10.82	18.16	21.76	524
Qual.	Static	0.39	25.49	36.33	41.22	582
	Dynamic	0.32	15.48	26.49	34.75	687
	Transition	0.33	17.48	29.24	35.00	681
	Combined	0.42	27.87	41.59	46.70	555

Research question 3. From an analyst's perspective, the learned models can be used to test several interesting soccer-related hypotheses in a data-driven way. First, we suspect there is a difference in a team's passing behavior at home and away. To test this, we learn a model from a set of home games, and test it on both home and away games. Table 5.2 displays the results. We observe a decrease in performance when applying a model learned from home games to data from away games, suggesting a difference in passing strategy. Second, throughout the game, players get tired and might become more prone to making mistakes, or the opposing team might be winning, leading to a more aggressive style of play... These things could affect the passing behavior of a team. We hypothesize that this difference should manifest itself as a decrease in performance when using

models trained in one time segment, to predict passes in another. The results in Table 5.2 reinforce this belief; notice there is a slight decrease in MRR, and a strong downward shift in recall when mixing time frames. Lastly, passing strategy should be team specific. We test this by constructing a model from pass data of one particular team, and use it to predict passes also for different teams. The corresponding decrease in performance suggests that models of passing behavior are team-specific.

Table 3. The table displays the results of testing a qualitative model in a number of scenario's. Each model is trained using the full range of static, dynamic, and transition features.

	MRR	top-1	top-2	top-3	Rules
Train home - test home	0.42	27.87	41.59	46.70	555
Train home - test away	0.37	21.56	35.25	40.58	712
Train 1st half - test 1st half	0.42	27.87	41.59	46.70	555
Train 1st half - test 2nd half	0.38	27.15	31.95	36.03	620
Train 1 team - test multiple	0.28	13.15	22.44	30.00	591
Train multiple - test multiple	0.37	23.71	35.05	40.33	381

6 Conclusion

This paper investigates using qualitative spatial reasoning techniques to overcome a number of challenges inherent to the analysis of spatio-temporal data. We tested our approach by constructing qualitative models for soccer pass prediction. The experiments demonstrated three things. First, qualitative, relational models outperform their purely quantitative counterpart. Second, it is possible to extend the existing QSR frameworks to encode dynamic and transition information that leads to improved performance when used in conjunction with static qualitative relations. Last, our approach is capable of extracting information that captures the characteristics of a certain game states. In principal, our model can be applied to other team sports, as long as spatio-temporal data about players' positions are available.

References

1. Bialkowski, A., Lucey, P., Carr, P., Yue, Y., Sridharan, S., Matthews, I.: Identifying Team Style in Soccer Using Formations Learned from Spatiotemporal Tracking Data. In: Proceedings of the Workshop on Spatial and Spatio-Temporal Data Mining. pp. 9–14 (2014)
2. Chen, J., Cohn, A.G., Liu, D., Wang, S., Ouyang, J., Yu, Q.: A survey of qualitative spatial representations. The Knowledge Engineering Review 30(01), 106–136 (2015)
3. ChyronHego: http://www.chyronhego.com, accessed: 2016-02-10

4. Cohn, A.G., Bennett, B., Gooday, J., Gotts, N.M.: Qualitative spatial representation and reasoning with the region connection calculus. GeoInformatica 1(3), 275–316 (1997)
5. Džeroski, S., Lavrač, N.: An Introduction to Inductive Logic Programming (2001)
6. Ekin, A., Tekalp, A.M., Mehrotra, R.: Automatic soccer video analysis and summarization. IEEE Transactions on Image processing 12(7), 796–807 (2003)
7. Frank, A.U.: Qualitative spatial reasoning with cardinal directions. In: 7. Österreichische Artificial-Intelligence-Tagung/Seventh Austrian Conference on Artificial Intelligence. pp. 157–167. Springer (1991)
8. Grunz, A., Memmert, D., Perl, J.: Analysis and simulation of actions in games by means of special self-organizing maps. International Journal of Computer Science in Sport 8(1), 22–37 (2009)
9. Grunz, A., Memmert, D., Perl, J.: Tactical pattern recognition in soccer games by means of special self-organizing maps. Human movement science 31(2), 334–343 (2012)
10. Gudmundsson, J., Wolle, T.: Football analysis using spatio-temporal tools. Computers, Environment and Urban Systems 47, 16–27 (2014)
11. Horton, M., Gudmundsson, J., Chawla, S., Estephan, J.: Classification of passes in football matches using spatiotemporal data. arXiv preprint arXiv:1407.5093 (2014)
12. Isli, A., Haarslev, V., Möller, R., et al.: Combining cardinal direction relations and relative orientation relations in qualitative spatial reasoning. Univ., Bibliothek des Fachbereichs Informatik (2001)
13. Knauf, K., Brefeld, U.: Spatio-Temporal Convolution Kernels for Clustering Trajectories. In: Proceedings of the Workshop on Large-Scale Sports Analytics (2014)
14. Knauf, K., Memmert, D., Brefeld, U.: Spatio-Temporal Convolution Kernels. Machine Learning 102(2), 247–273 (2016)
15. Lucey, P., Oliver, D., Carr, P., Roth, J., Matthews, I.: Assessing team strategy using spatiotemporal data. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1366–1374. ACM (2013)
16. Muggleton, S., De Raedt, L.: Inductive logic programming: Theory and methods. The Journal of Logic Programming 19, 629–679 (1994)
17. Prozone: http://www.prozonesports.com, accessed: 2016-02-10
18. Srinivasan, A.: The Aleph Manual (2001)
19. Stensland, H.K., Gaddam, V.R., Tennøe, M., Helgedagsrud, E., Næss, M., Alstad, H.K., Mortensen, A., Langseth, R., Ljødal, S., Landsverk, Ø., et al.: Bagadus: An integrated real-time system for soccer analytics. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 10(1s), 14 (2014)
20. Taki, T., Hasegawa, J.i.: Visualization of dominant region in team games and its application to teamwork analysis. In: Computer Graphics International, 2000. Proceedings. pp. 227–235. IEEE (2000)
21. Van Haaren, J., Dzyuba, V., Hannosset, S., Davis, J.: Automatically Discovering Offensive Patterns in Soccer Match Data. In: Fromont, E., De Bie, T., van Leeuwen, M. (eds.) International Symposium on Intelligent Data Analysis. Lecture Notes in Computer Science, vol. 9385, pp. 286–297. Springer (Oct 2015), https://lirias.kuleuven.be/handle/123456789/503185
22. Van de Weghe, N., Cohn, A.G., De Maeyer, P.: A qualitative representation of trajectory pairs. In: ECAI. vol. 16, p. 1103 (2004)