Strategy Discovery in Professional Soccer Match Data

Jan Van Haaren
KU Leuven — Department of Computer Science
Celestijnenlaan 200A, 3001 Leuven, Belgium
jan.vanhaaren@cs.kuleuven.be

Siebe Hannosset
Stirr Associates — Chief Sports Officer
Koopliedenstraat 65, 1000 Brussel, Belgium
siebe@stirrassociates.com

Jesse Davis
KU Leuven — Department of Computer Science
Celestijnenlaan 200A, 3001 Leuven, Belgium
jesse.davis@cs.kuleuven.be

ABSTRACT

This paper explores the task of automatic strategy detection from event-stream data collected from professional soccer matches. Concretely, we focus on discovering interesting event sequences that lead to an attempt on goal. We describe a data-driven approach for identifying patterns of movement that account for both spatial and temporal information which represent potential strategies.

1. INTRODUCTION

Recent technological advances have enabled the collection of large amounts of data about soccer teams and players. Companies such as ChyronHego [2] and Prozone [11] record the locations of the players and the ball at a high frequency using optical tracking systems during matches. Soccer clubs aim to leverage the collected data to gain a competitive advantage over their opponents. However, soccer clubs lack computational methods that can handle the size and complexity as well as the spatial and temporal aspects of the data. As a result, there has been an explosion of interest in applying automated techniques to analyze data collected about sports matches (e.g., [8, 9, 13, 10, 1]).

This paper focuses on the task of detecting strategies from professional soccer matches based on spatio-temporal data. This problem poses a number of significant challenges from a data mining perspective. First, important patterns will involve both spatial and temporal components. Second, there will rarely be exact matches in terms of the same set of players performing the same actions in the same order in the same locations. Third, there is rich domain knowledge about soccer that can be exploited to guide the discovery process. Fourth, frequency is not necessarily the most important criteria for interestingness in strategy detection. Certain events such as goals and shots are rare, and sequences involving them are correspondingly more valuable and interesting.

We propose an approach to address the specific task of discovering event sequences that frequently lead to an attempt on goal in soccer matches. On a high level, our approach performs the following four steps. First, the algorithm automatically extracts phases leading to an attempt on goal for the team of interest. Second, we employ a data-driven approach to determine a number of spatial features about the areas occupied during the phase. We use these features to cluster together similar phases. Third, we search for frequently occurring sequences of events within each cluster. Fourth, based on domain knowledge, we developed a ranking function that orders the discovered patterns in each cluster according to their expected relevance to the user.

We evaluate our approach on data from 69 matches played by a Belgian professional soccer club, where we have access to event-stream data for all matches and detailed camera-tracking data for 13 matches. We find that our approach is capable of identifying interesting strategies related to goal-chance creation from open play as well as set pieces.

2. RELATED WORK

This paper falls within the area of work that looks at analyzing spatio-temporal sports data. Knauf et al. [8, 9] proposed a novel spatio-temporal kernel for clustering player trajectories. Another trajectory-based approach focuses on scoring opportunities [3], which cluster together different scoring chances based on hand-crafted features and trajectory data. Inductive logic programming, which allows representing rich, relational structure in a domain, has been used to characterize scoring chances [13].

Another way to analyze strategy is to build occupancy maps based on ball movements [10]. Characterizing playing style and strategy by looking at passing patterns has also received attention [6, 5]. Beyond these, other strategy analyses include recognizing team formations in soccer (e.g., [1]) or identifying specific plays in American football (e.g., [12]).

3. DATASET

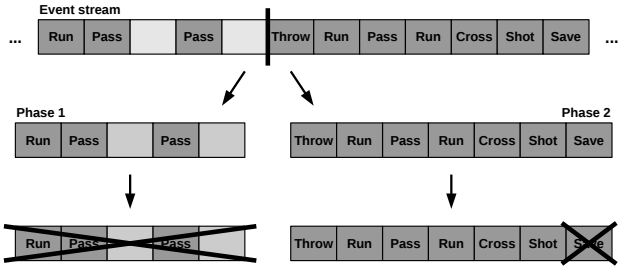
The dataset consists of match data for 69 matches from a professional Belgian soccer club collected by Prozone [11]. The dataset comprises 58 Belgian league matches, 2 Belgian cup matches, and 9 Europa League matches. For 13 matches both event and tracking data are available, while for the remaining 56 matches only event data are available.

Our dataset contains 180,981 events of 44 different types, corresponding to an average of 2,623 events per match. The most frequent event types in our dataset are passes (61,220), receptions (47,533), and runs with the ball (46,914). Our dataset also contains 18,119,283 player and ball locations in total, or 1,294,235 locations per match on average.

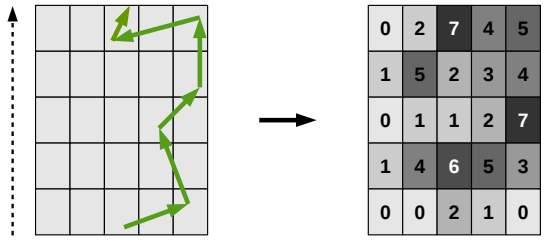
4. APPROACH

Our approach addresses the following task:

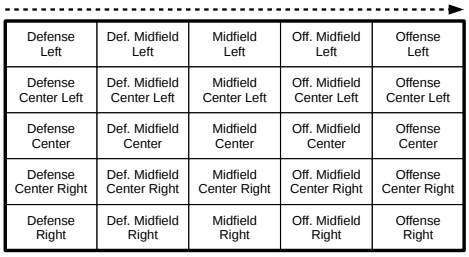
Given: A set of matches where each match is represented as an event sequence, possibly with player and ball tracking.



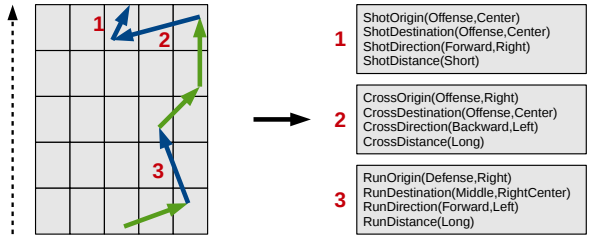
(a) A match is split into phases, which are sequences of related events. The dark-shaded cells indicate possession by the team of interest, while the light-shaded cells indicate possession by their opponent. Each time a team gains possession, a new phase starts. We discard phases that do not lead to a goal attempt as well as events in a phase that happen after a goal attempt.



(b) Each phase is mapped onto a possession map, which is a virtual grid over the pitch. A possession map essentially shows how often each area (i.e., cell in the grid) was occupied by the players and the ball during a phase. The green arrows represent the events (e.g., a pass or a cross) that happened during the phase. The direction of play is from bottom to top.



(c) The pitch is mapped onto a 5×5 grid to represent the start and end locations of the events. The direction of play is always from left to right for the team of interest.



(d) Each phase is represented as a sequence, which is a list of itemsets. Each itemset describes one event that happened during the phase. For each event, we store the origin, destination, direction, and distance. The blue arrows represent the events that are described at the right. The direction of play is from bottom to top.

Figure 1: Visualization of our approach explaining each of the performed steps.

Find: Relevant spatio-temporal patterns that characterize attacking strategies.

This is a challenging task as soccer is a highly dynamic game with many movements and interactions among players across time and space.

To tackle these challenges, we perform the following four steps. First, we divide the event stream of each match into phases and identify those phases that involve a shot. Second, we convert each phase into a possession map and then cluster the phases according to these possession maps. Third, we mine each of the obtained clusters to identify frequent sequential patterns. Fourth, we rerank the discovered patterns within each cluster according to a hand-crafted domain-specific criteria that weighs the elements in a pattern by interestingness.

4.1 Discovering relevant phases

In the first stage, our approach discovers the relevant phases in a match, where a match is an almost continuous stream of events (e.g., a pass or a run with the ball) and phases are sequences of related events (e.g., a cross followed by a shot). Formally, we represent a match as a sequence $S = \{e_1, e_2, ..., e_n\}$, where each event snapshot e_i has a corresponding label l (e.g., a pass or a shot), a timestamp t, a (x, y) position, and a set P of involved players p_i. If the tracking data are available, then there is also a list containing the (x, y) positions of the players and the ball.

Since we are interested in the offensive strategy of a particular team, we consider a phase relevant if it leads to a goal attempt for that team. In general, a phase starts when a team gains possession and ends when the team loses the ball again. While many events can lead to possession changes, they typically happen when a foul is committed, a pass is intercepted, the ball crosses the sideline or goal line, or after a duel with an opponent.

As shown in Figure 1a, our approach to discovering relevant phases proceeds in two steps. In the first step, we split the match into phases with a new phase starting each time a team gains possession. In the example in Figure 1a, the event stream is split into two phases with the second phase starting with a throw-in. In the second step, we discard all phases that do not lead to a goal attempt (e.g., phase 1 in Figure 1a) or where the opponent touched the ball more than two consecutive times. Furthermore, we discard the events from a phase that happened after a goal attempt (e.g., the save by the goalkeeper in phase 2 in Figure 1a).

4.2 Clustering relevant phases

The goal of the second stage is to identify spatially similar phases via clustering. We do this for several reasons. One, this helps reduce the space of possible patterns that we need to search in the following step. Two, a team is likely to employ multiple different attacking strategies, such as corners, attacking through the middle, down the flank, each of which will be characterized by different spatial characteristics. Clustering gives us a natural way to divide the

data along these lines. Three, the spatial features provide a convenient manner to generalize from an event's or player's specific (x, y) location on the pitch to a more general zone.

We employ a data-driven approach to describe each phase by a number of spatial characteristics. We convert each phase into a possession map M, which is a virtual $w \times h$ grid over the pitch, where each cell represents a particular area of the pitch. Figure 1b shows an example map. When building M, we consider the locations of the events as well as the locations of the players and the ball if they are available in the dataset (see Section 3).

We initialize the value of each cell in M to zero. Then, we iterate over each event snapshot e_i in the phase under consideration. For each player and the ball whose precise location falls within the grid location (x, y) at time i, we increment the value $M[x, y]$ by a user-defined value v, which depends on the entity type. We assign a higher value for the ball than the players. After incrementing a grid location's value, we also slightly increase the value of each of its neighbors to create a smoothing effect which helps overcome the rigidness imposed by the grid. Finally, we normalize the cell values to lie within 0 and 1 to account for the fact that phases can consist of an arbitrary number of events. Intuitively, a possession map shows how often each area was occupied by the players and the ball during a phase.

Next, we build a feature vector describing the possession map, where we have one feature for each cell in the map, whose value is the corresponding value from the possession map. The feature vector also contains two additional features: one whose value is the number of events in the phase, and another whose value is the number of players involved in the phase. Then, we use the Expectation-Maximization algorithm [7] to cluster the feature vectors. The algorithm assigns a probability distribution to each phase's feature vector indicating the probability that the phase belongs to each of the clusters. Finally, we convert this to a hard clustering by using a MAP assignment of the phases to the clusters.

4.3 Mining patterns

In the third stage, we search for frequent sequential patterns (i.e., sets of events) within each identified cluster. We process each cluster in turn. We represent each phase assigned to the cluster as a sequence. A sequence D is a list of itemsets $\{(d_{11}, ..., d_{1m}), ..., (d_{n1}, ..., d_{nm})\}$, where each itemset $(d_{i1}, ..., d_{im})$ describes an event that happened during a phase. For each event, we store the origin (i.e., area where the event started), the destination (i.e., area where the event ended), the direction (i.e., backward or forward and left or right), and the distance (i.e., short or long) as an item d_{jk}. To represent the origin and destination of each event, we map their precise locations to a cell f_{ij} in a 5×5-grid F, as is illustrated in Figure 1c. Distances over 20 meters (i.e., 21.9 yards) are defined as long, while all other distances are defined as short.

By applying this transformation to each phase in a cluster, we end up with a sequence database DB. Then we run the VPSM algorithm [4] to discover frequent maximal sequential patterns in each database DB. Figure 1d shows the representation of a shot (1), a cross (2), and a run with the ball (3) that happened during a phase consisting of six events in total. The direction of play is from bottom to top.

4.4 Ranking patterns

Finally, we rank the discovered frequent sequential patterns with respect to their expected relevance to a user. Typically, frequent patterns are ranked according to their support in the data. However, this evaluation function is less relevant to soccer coaches. Given that most of the action during a soccer match typically happens in the middle of the pitch, the top of the ranking is likely to be dominated by patterns describing passing sequences in that area.

We propose an alternative evaluation function that considers the types of the events appearing in a pattern to determine its relevance. More specifically, we first assign a weight to each event type and then compute the relevance of a pattern by summing the weights of the events appearing in the pattern. Higher weights indicate higher relevance. This approach allows the user to define a bias towards a particular type of patterns. Given that we are mostly interested in goal attempts, we assign high weights to shots and crosses and low weights to the other types of events in our experiments.

5. EXPERIMENTAL STUDY

The goal of our empirical evaluation is to investigate whether our approach discovers interesting patterns. We implemented the approach as discussed in Section 4. The maximum number of events per phase was set to ten. The dimension of the possession map was set to 16 cells wide and 21 cells high. The values for the location of the ball and players were set to three and one, respectively. For the VPSM algorithm, the maximum pattern length was set to ten, the maximum gap was set to one, and the minimum support was set to 0.02.

Figure 2 shows four concrete instances of patterns that our approach discovered. These patterns appear 3, 2, 3, and 10 times in the dataset, respectively. Figure 2a shows an attack down the right flank resulting in a cross from the right offensive midfield area to the center offense area followed by a close-range shot going wide. Figure 2b shows an attack through the middle resulting in a long-range shot on target from the right center offensive midfield area. Figure 2c shows an attack down the left flank resulting in a short-range shot off target from the center offense area. Figure 2d shows a corner kick from the right side of the pitch leading to a shot off target from the center offense area.

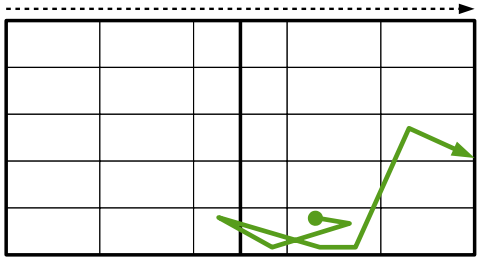
The patterns shown in Figures 2a and Figure 2c suggest a strategy where the team prefers early crosses (i.e., crosses originating further away from the goal line) to late crosses (i.e., crosses originating close to the goal line). This observation could imply a more direct style of play. The pattern shown in Figure 2b suggests a strategy where the team first tries to move the ball to their playmaker in the center of the pitch who then attempts a long-distance shot.

6. CONCLUSIONS

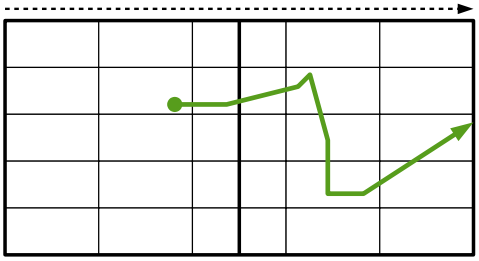
Automatically analyzing playing strategy from rich, complex spatio-temporal data is an interesting and challenging problem. This paper tackled one aspect of this task by trying to automatically discover interesting attacking strategies from event data collected from professional soccer matches.

Acknowledgments

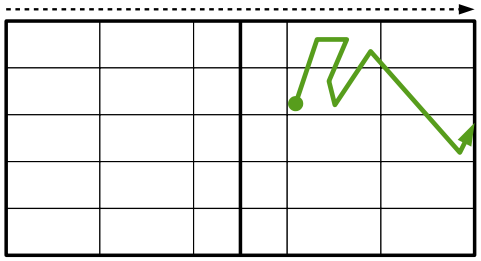
Jan Van Haaren is supported by the Agency for Innovation by Science and Technology (IWT). Jesse Davis is partially supported by the KU Leuven Research Fund (C22/15/015), and FWO-Vlaanderen (G.0356.12, SBO-150033).



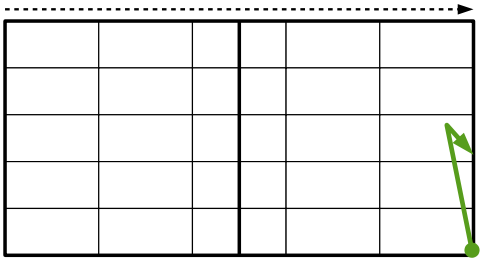
(a) This phase represents an attack down the right flank.



(b) This phase represents an attack through the middle.



(c) This phase represents an attack down the left flank.



(d) This phase represents a corner kick from the right side.

Figure 2: Visualizations of four concrete instances of discovered patterns. During each phase, the ball follows the trajectory shown by the green arrow. The direction of play is from left to right in each visualization.

7. REFERENCES

[1] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews. Identifying Team Style in Soccer Using Formations Learned from Spatiotemporal Tracking Data. In *Proceedings of the Workshop on Spatial and Spatio-Temporal Data Mining*, pages 9–14, 2014.

[2] ChyronHego. http://www.chyronhego.com. Accessed: 2016-02-10.

[3] T. Fernando, X. Wei, C. Fookes, S. Sridharan, and P. Lucey. Discovering Methods of Scoring in Soccer Using Tracking Data. In *Proceedings of the Workshop on Large-Scale Sports Analytics*, 2015.

[4] P. Fournier-Viger, A. Gomariz, T. Gueniche, A. Soltani, C. Wu., and V. S. Tseng. SPMF: A Java Open-Source Pattern Mining Library. *Journal of Machine Learning Research*, 15:3389–3393, 2014.

[5] L. Gyarmati and X. Anguera. Automatic Extraction of the Passing Strategies of Soccer Teams. *arXiv:1508.02171*, 2015.

[6] L. Gyarmati, H. Kwak, and P. Rodriguez. Searching for a Unique Style in Soccer. *arXiv:1409.0308*, 2014.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[8] K. Knauf and U. Brefeld. Spatio-Temporal Convolution Kernels for Clustering Trajectories. In *Proceedings of the Workshop on Large-Scale Sports Analytics*, 2014.

[9] K. Knauf, D. Memmert, and U. Brefeld. Spatio-Temporal Convolution Kernels. *Machine Learning*, 102(2):247–273, 2016.

[10] P. Lucey, D. Oliver, P. Carr, J. Roth, and I. Matthews. Assessing Team Strategy Using Spatiotemporal Data. In *Proceedings of the 19th International Conference on Knowledge Discovery and Data Mining*, pages 1366–1374, 2013.

[11] Prozone. http://www.prozonesports.com. Accessed: 2016-02-10.

[12] D. J. Stracuzzi, A. Fern, K. Ali, R. Hess, J. Pinto, N. Li, T. Konik, and D. G. Shapiro. An Application of Transfer to American Football: From Observation of Raw Video to Control in a Simulated Environment. *AI Magazine*, 32(2):107–125, 2011.

[13] J. Van Haaren, V. Dzyuba, S. Hannosset, and J. Davis. Automatically Discovering Offensive Patterns in Soccer Match Data. In *International Symposium on Intelligent Data Analysis*, volume 9385 of *Lecture Notes in Computer Science*, pages 286–297. Springer, Oct. 2015.