Constructing Spaces and Times for Tactical Analysis in Football

Gennady Andrienko, Natalia Andrienko, Gabriel Anzer, Pascal Bauer, Guido Budziak, Georg Fuchs, Dirk Hecker, Hendrik Weber, and Stefan Wrobel

Abstract—A possible objective in analyzing trajectories of multiple simultaneously moving objects, such as football players during a game, is to extract and understand the general patterns of coordinated movement in different classes of situations as they develop. For achieving this objective, we propose an approach that includes a combination of query techniques for flexible selection of episodes of situation development, a method for dynamic aggregation of data from selected groups of episodes, and a data structure for representing the aggregates that enables their exploration and use in further analysis. The aggregation, which is meant to abstract general movement patterns, involves construction of new time-homomorphic reference systems owing to iterative application of aggregation operators to a sequence of data selections. As similar patterns may occur at different spatial locations, we also propose constructing new spatial reference systems for aligning and matching movements irrespective of their absolute locations. The approach was tested in application to tracking data from two Bundesliga games of the 2018/2019 season. It enabled detection of interesting and meaningful general patterns of team behaviors in three classes of situations defined by football experts. The experts found the approach and the underlying concepts worth implementing in tools for football analysts.

Index Terms—Visual analytics, movement data, coordinated movement, sport analytics, football, soccer.

1 Introduction

Football (soccer) is an exciting sport that attracts millions of players and billions of spectators worldwide. Wikipedia explains the basics as: *"Football is a team sport played with a spherical ball between two teams of eleven players. ... The game is played on a rectangular field called a pitch with a goal at each end. The object of the game is to score by moving the ball beyond the goal line into the opposing goal"* [1]. Although it was sufficient for G.Lineker to use just a single sentence to fully define football as *"Twenty-two men chase a ball for 90 minutes and at the end, the Germans always win"*, in reality football is very complex. 22+ players, the ball and 3 referees move and act in coordination within the teams and in competition between the teams. The game is defined by voluminous rules, characterized by complex interactions, and requires specific skills and sophisticated tactics.

Team managers (coaches) define team tactics and select a plan for each game that needs to be carefully implemented by the players. For winning a game, a team needs (1) skilled players in excellent physical conditions and (2) sophisticated tactics intelligently defined by coaches, effectively taught to players, trained, and carefully implemented in the game. While training is covered well by sport science, tactical analysis is still challenging. Data-driven tactical analysis requires understanding of information hidden in large volumes of game tracking data that include frequently sampled positions of players and the ball and numerous game events such as goal shots and goals, passes, tackles, possession changes, substitutions, fouls etc.

Professional football attracts tremendous interest and therefore is supported by industry and huge investments into infrastructure, players and coaches. Recent progress in football data collection, processing and analysis [2], [3] created new opportunities for providing data-driven insights into the game and, eventually, supporting a variety of stakeholders including coaches, medical staff of clubs, players, scouts, leagues, journalists and general public. Professional clubs nowadays intensively hire data scientists and some major clubs already have their own data analysis departments [4], [5]. Several companies develop software for supporting data collection, processing and statistical analysis and provide services to clubs delivering data and analysis results, including visualizations, which are mainly of illustrative nature. Analytical visualizations and visual analytics at large are still seeking their way to this domain.

This paper results from joint research and co-authorship of a group involving visual analytics researchers, data scientists, and football experts. The research goal of the group was to find approaches to extracting and understanding the general patterns of the **team behaviors** and **dynamics of changes** in relation to **events** and **context** in different classes of game **situations** as these situations develop. In the context of the paper, the term 'situation' refers to a combination of circumstances in which players behave, and the term 'episode' refers to the situation development in which the circumstances dynamically change. The overall goal involves the following sub-goals:

1) enable selection of groups of situations with particular characteristics and extraction of data pieces reflecting the development of these situations;

- *Gennady Andrienko and Natalia Andrienko are with Fraunhofer IAIS and City, University of London. E-mail: gennady.andrienko@iais.fraunhofer.de*
- *Gabriel Anzer is with Sportec Solutions GmbH, Germany*
- *Pascal Bauer is with DFB Akademie, Germany*
- *Georg Fuchs and Dirk Hecker are with Fraunhofer IAIS, Germany*
- *Guido Budziak is with TU Eindhoven, The Netherlands*
- *Hendrik Weber is with DFL Deutsche Fussball Liga GmbH, Germany*
- *Stefan Wrobel is with Fraunhofer IAIS and University of Bonn, Germany*

Manuscript received April 19, 2005; revised August 26, 2015.

2) derive general patterns of team behaviors from the extracted data pieces;
3) enable comparison of general patterns corresponding to different groups of situations.

To achieve these sub-goals, our group has developed a framework including techniques for (1) query and data extraction (filtering), (2) integration of extracted data pieces into aggregate structures, which may involve (2*) space transformation, and (3) visualization of the aggregates for interpretation, exploration, and comparison. These components of the framework are briefly described below.

1. Query and filtering. This component enables selection of time intervals containing game episodes with target characteristics, which may refer to occurrences of specific game events (e.g., shots, passes in a given direction at a certain distance, etc.) and attributes, such as speed or acceleration, of the ball, teams, and selected players. Once a set of target intervals is selected according to event- or attribute-based query conditions, it is possible to further select intervals positioned in time in a specific way in relation to the target intervals, e.g., starting at a given time distance before or after the beginning or the end of each target interval and having a specified duration. This enables exploration of what had happened before and after the target episodes and at different stages of their development.

2. Aggregation. Trajectory fragments extracted from the selected intervals are integrated into aggregate structures, where each structure represents the behavior of one moving object (a player, the ball, the mass center of a team, etc.) and consists of a sequence of generalized positions corresponding to a sequence of interactive selections of time intervals. A generalized position is an aggregate of the positions of the object extracted from all selected intervals. It is represented by a central point, which may be the mean, median, or medoid of the extracted subset of positions, and one or more convex hulls covering chosen fractions (e.g., 50 and 75%) of this subset. Each sequence of generalized positions is represented by a pseudo-trajectory in an abstract temporal domain where the sequence of time stamps corresponds to the sequence of selections. The resulting sets of pseudo-trajectories of all players and the ball provide a generalized representation of collective behaviors in all situations with particular properties.

2*. Space transformation. As an additional means of abstraction and generalization, this component allows putting together similar movements that might take place in different parts of the pitch. The core idea is to replace the positions of the moving objects in the physical space (i.e., on the pitch) by corresponding positions in artificially constructed spaces, such as a team space, which represents the relative placements of the players within a team, or an abstract space with dimensions corresponding to some attributes. Generalized positions and pseudo-trajectories can be constructed from positions in artificial spaces in the same way as from the original positions in the pitch space. The pitch space is good for analyzing the tactics of the team movement while the team space is good for seeing the relative arrangement of the players and how it changes depending on circumstances.

3. Pattern visualization. To enable perception and interpretation of collective behavior patterns by an analyst, a set of aggregates (i.e., pseudo-trajectories) generated from selected episodes is represented visually. For comparative analysis of sets of aggregates generated for different teams, kinds of situations, and games, the pseudo-trajectories are put in a common spatial domain (i.e., the pitch space, team space, or attribute space) and aligned with respect to their abstract times.

While the framework makes use of previously existing techniques and approaches, it also incorporates novel ideas, specifically:

- new primitives for temporal queries allowing specification of relative time intervals (Section 4.1);
- a novel way of aggregating movement data that is suitable for bringing together temporally disjoint data pieces (Section 4.3);
- a data structure for representing aggregated movement data that allows the aggregates to be visualized and explored similarly to trajectories (Section 4.3.2);
- glyphs showing usual relative positions of players in their teams and providing hints at their roles (Section 4.2 and Fig. 4).

We demonstrate the effectiveness of the proposed framework in several case studies using real data from two Bundesliga games [6], [7] of the season 2018-19.

The remainder of the paper has the following structure. Section 2 introduces the main concepts concerning football tactics, describes the collection, contents, and structure of data from a football game, and presents the research problem we have addressed. After an overview of the related work (Section 3), we present our approach and components of the analytical framework in Section 4 and describe how we have applied it to three complex scenarios of team tactics analysis (Section 5). Section 6 discusses the overall approach and outlines directions for further work.

2 BACKGROUND

2.1 Football tactics in a nutshell

Football tactics depends on multiple factors: which team possesses the ball, in what part of the pitch the ball and the teams are located, and how the players are arranged within their teams and in relation to the opponents. When a team possesses the ball, it aims at scoring a goal by offensive actions, although in some rarely occurring cases (such as a lead close to the end of the game) it can be a teams' solely objective to stay in ball possession. When the ball is possessed by the opponents, a team aims at preventing a goal and performs defense. There is an intermediate **turnover** stage between offensive and defensive actions. After winning the ball, a team can either **counter-attack** or **safeguard and build up**. After losing the ball, a team can either **fall back and defend** or perform **counter-pressing**. In some situations, e.g. after fouls or if the ball goes out of the pitch, the game is interrupted and resumes through **set pieces**, such as corners, free kicks, throw-ins, goalie kicks, and penalties.

Players in football teams have different roles. An established term **formation** means a way how 10 outfield players in a team generally position themselves relative to their teammates. Formations typically consist of three or four rows of players and are described, respectively, by

three or four numbers specifying how many players are in each row from the most defensive to the most forward [8]. For example, formation 4-3-3 means that the team has 4 defenders, 3 midfielders and 3 forwards, or strikers. In some formations, intermediate lines appear denoting attacking or defensive midfielders or so-called second forwards playing slightly behind their partner.

Formations usually differ significantly depending on the ball possession, so that each team has an offensive formation and a defensive formation. Schematic figures in media usually show only the offensive formations. First investigations on comparing offensive and defensive formations of the same teams were made in [9] where the average line-ups of two teams were shown both in and out of ball possession. When the ball possession changes, teams strive to arrange themselves as fast as possible into the respective opposite (offensive or defensive) formation. Generally, formations as a major component of football tactics are carefully studied in literature [10]. There exist manuals for coaches (e.g. [11]) and catalogues of offensive formations (e.g. [12]) enumerating possible attacking styles and suggesting efficient defense.

However, not only the chosen formations are important. Football is a highly dynamic game where the players not just take fixed relative positions but constantly move in a coordinated manner, which does not simply mean moving in parallel and thereby keeping the same arrangement. Both the arrangement of the players and their relative movements depend on multiple factors, including which team and for how long possesses the ball, where on the pitch the teams are located, what are the distances to the opponents, what events happened recently, what is the current score of the game, etc. For understanding teams' tactics and their efficiency, it is necessary to see the spatial arrangements of the players and the character and dynamics of their changes in response to game events and other circumstances. In today's practice, this is a very time-consuming process done largely by analysts watching game videos and synthesizing information by reasoning. Several recent research prototypes [13], [14], [15], [16] support this activity by extracting formations and their changes from the data. However, changes of formations and, more generally, changes of movement behaviors do not happen instantly. What is still missing and challenging in supporting game analysis is a possibility to analyze the process of change in the context of game events and situation characteristics. Our paper intends to fill this very important gap.

2.2 Data acquisition, content, and structure

Today, detailed data are collected for almost every football game in major professional leagues. Usually positional data are extracted from video recordings. For this purpose, stadiums are equipped with stationary installations of multiple cameras that record games from different viewpoints. Video analysis software is used for extracting time-stamped positions of the players, referees, and the ball from video footage, usually with a sampling rate of 10-25Hz. Additionally, game events are extracted from video and positional data. Event data include the positions and times of the events and annotations, i.e., attributes describing the event types, involved players, outcomes, etc. The event extraction and annotation is done partly manually, though there exist implementations that facilitate manual annotation using machine learning approaches. Major companies doing data acquisition and processing are ChyronHego [17], OPTA [18], STATS [19], SecondSpectrum [20] and Track160 [21]. Smaller companies (e.g., FootoVision [22]) develop lightweight solutions for extracting data from a single video.

A typical data set for one game consist of general information (date and location, playing teams, names of referees), information about the teams (list of players with their intended positions on the pitch, list of reserve players for substitutions), positional data (coordinates of the ball, players and, sometimes, referees in 2D x,y or 3D x,y,z space with time references) and events (what happened, when and where, with event-specific characteristics). In average, about 140,000 positions for the ball and each player are recorded during one game, roughly 3,500,000 positions in total. In addition to automatically recorded positions, about 1,500 events are annotated manually and then validated using computational methods. As any real-world data, football data require assessment of data quality, plausibility checking, and evaluation of the coverage in space, time, and the set of moving objects [23]. Particularly, it is necessary to make queries with allowing certain tolerance to potential mismatch of times in positional and event data.

In addition to data collection, commercial companies provide basic analytical and visualization services. A typical menu of provided visuals includes depiction of individual events (e.g., positions of fouls and tackles, geometries of passes and shots) and aggregated representations of players' positions on the pitch such as density heat maps. Both types of visuals can be filtered by players and times. However, possibilities for exploration by connecting different aspects are not available.

2.3 Problem statement

The formulation of the research goals comes from the football experts. Basically, their question was: How team tactics can be understood from data reflecting the movements of the individual players and the ball (i.e., their trajectories) and the events that occurred during a game? All partners communicated to clarify the concept of team tactics and, on this basis, define and refine the research goals.

A team tactic can be defined as a general pattern of collective behavior in a group of situations with particular properties. This definition requires further clarification of the concepts of situation properties, collective behavior, and general pattern. Situation properties can be specified in terms of various attributes: which team possesses the ball, how much time has elapsed since the possession change, which team is winning, where on the pitch is the ball and the majority of team players, etc. Collective behavior means relative positions, movements, and actions of the players with respect to their teammates and the opponents. A general pattern means a synoptic representation integrating multiple specific instances of collective behavior that were practiced in similar situations. The patterns differ depending on the situation properties. Hence, understanding of team tactics requires consideration of groups of situations with different properties.

Based on this refinement, the research goals presented in the introduction section were formulated.

3 Related work

3.1 Major approaches to football analytics

Several groups of researchers managed to get access to game tracking data and developed interesting research prototypes. Often the starting point was an adaptation of methods and tools developed for other purposes (e.g. animal tracking or transportation) for football data. A prominent example is the famous Soccermatics book by D.Sumpter [4] that builds on his research on collective animal behavior [24].

A review [25] observes the state of the art, considering the following high-level tasks: playing area subdivision, network techniques for team performance analysis, specific performance metrics, and application of data mining methods for labelling events, predicting future event types and locations, identifying team formations, plays and tactical group movement, and temporally segmenting the game. Some of the considered methods actively use visualization components and thus fall into visual analytics (VA) approaches. Another review [9] takes a different perspective, emphasizing the works with substantial involvement of visualization and identifying the following major approaches:

Analysis of **game events**. A representative example is SoccerStories [26], which summarizes game episodes using visual primitives for game events such as long ball, turning the ball, cross, corner, shot etc.

Analysis of **trajectories** and **trajectory attributes**. A series of works from the University of Konstanz proposed methods for clustering trajectories of players during game episodes [27] and segmenting the game, finding interesting game situations [28], [29] and plays of particular configurations [30], analyzing multiple attributes along trajectories [31] and computing features of team coordination [32].

Analysis of **team formations** and derived features of them. Several papers from Disney Research target at reconstructing team formations and player roles from positional data. The proposed methods identify the role of each player at each time moment allowing the analyst to trace short- and longer-term roles and detect role swaps between players. This approach allowed characterization of team styles in several games [13], [14]. After enumerating offensive and defensive configurations of players, paper [15] evaluates pairwise success statistics. ForVizor [16] uses the dynamics of detected formations for segmenting the game. Another approach for game segmentation is clustering of time moments based on features reflecting relative positions of players or other team configuration indicators [33].

Computation of football-specific **constructs**, such as interaction spaces [31], and **indicators** such as scoring chances, pass options [34], [35], and pressure degrees [9], followed by visual representation of these in spatio-temporal displays. An interesting development is including visualization of computed features directly in video frames [36].

Apart from the research prototypes, there are also commercial systems and services that support coaches and match analysts in their work with positional data. We are aware of four such tools: STATS Edge Viewer [37], parts of the SAP Sports One [38], Second Spectrum [20], and the online match analysis portal offered to the Bundesliga clubs by Sportec Solutions [39]. The functionality provided by these tools can be divided into three categories: calculation of various statistics, which are visualized in business graphics, search for specific game episodes in video records, and replaying of selected episodes augmented with visual representation of calculated features, such as control zones and pass opportunities. Hence, it is possible either to analyze the overall statistics at the level of a whole game or to explore details of individual episodes. Our research fills the gap between these two extremes by developing approaches to extracting general movement patterns from multiple episodes with some common properties. Importantly, it is not limited to computing numeric statistics from selected parts of a game, but it produces more complex spatio-temporal constructs representing movements.

3.2 Relevant visual analytics approaches beyond football

Different visual analytics approaches proposed for analyzing spatio-temporal and movement data [40] are relevant to football analysis, although some of them have been developed for specific application domains such as transportation [41].

Querying and filtering. The structure of movement data suggests possibilities for selection of subsets based on the identities of the moving objects and their attributes, as well as dynamic data items including locations, times, and movement attributes, such as speed and direction [40]. A query may involve a combination of multiple heterogeneous aspects; thus, Weaver [42] discusses interactive cross-filtering across multiple coordinated displays by direct manipulation in the displays. There exist special query devices for temporal sequences of attribute values (e.g. TimeSearcher [43]) and for sequences of events.

The kind of analysis our group aimed to support requires selection of *groups of time intervals* containing game episodes with particular characteristics. Database researchers long ago proposed time query primitives [44], [45] suitable for such purposes. Recently, similar ideas were implemented within an interactive visual analytics environment in a tool called TimeMask [46]. Our approach extends this work by increasing query flexibility, see Section 4.1.

Transformation of space and time provide additional perspectives for looking at movement data. It may be useful to treat selected pairs of numeric attributes as coordinates in an abstract space [47]. A polar coordinate system may be used in such a space if the movement directions or cyclic time attributes are involved [48]. Research on group movement [49] introduces the idea of a group space consisting of relative positions in respect to a central trajectory of the group. This idea was successfully applied for analyzing the distribution of pressure over team formations in football [9].

Analysis of multiple asynchronous trajectories can benefit from transforming the time stamps of the positions to relative time references within relevant time cycles or within the individual lifetimes of the trajectories, which brings them to a common temporal reference system and thus supports comparisons and finding general patterns [50]. In this work, the ideas of space and time transformation have

been further extended, taking into account the new ways of time filtering, see Section 4.2.

Aggregation is one of the most important tools for spatial abstraction and simplification of massive movement data [51]. Review [25] suggests spatial aggregation over Cartesian or polar grids or hand-designed polygons that reflect specific functions of pitch regions. Aggregation results can be automatically re-calculated in response to changes of query conditions.

The existing approaches to aggregation of movement data produce the following major types of aggregates: density fields [9], [52], place-related attributes reflecting various statistics of the appearances of moving objects in the places [53], flows between places [51], and a central trajectory of a set of similar trajectories [54]. The former two types represent the presence of moving objects rather than their movement while the latter two represent the movement of a whole group but not the movements of its members. In previous works, relative positions of group members were represented by density distributions [49] or by their average positions [9], but their movements within the group were not reflected. Hence, none of the previously existing aggregation methods is well-suited for representing collective movement patterns. Therefore, our group has devised a novel way of aggregation producing a novel kind of aggregate – a set of pseudo-trajectories, see Section 4.3. Aggregation operations are combined with time filtering and time transformation and enable assessment of variability within aggregates.

3.3 What is missing

Among the earlier works dealing with collective movement, some focus on detection of occurrences of specific relationships between moving objects, such as close approach, others search for overall patterns of collective behavior. However, the former result in multiple disjoint data pieces that do not make a general picture while the latter, on the opposite, tend to overgeneralize by neglecting essential differences between situations. Football is a dynamic phenomenon with high variability of situations, therefore it is necessary to understand the dynamics of patterns and differentiate individual and collective behaviors depending on the situational context. Hence, there is a need to develop methods for identifying classes of situations, detecting patterns of coordinated movement in subsets of similar situations, and comparing patterns across different subsets.

Another important aspect is a possibility to see relationships between individual behaviors in the overall context of coordinated movement. There are two important aspects of these relationships: (1) the spatial arrangement of individuals within a group and (2) how the arrangement changes in response to different circumstances. Our work aimed at developing appropriate methods for satisfying these needs.

4 Approach

4.1 Temporal queries for episode selection

We use the term *situation* to denote a particular combination of circumstances that may take place in the course of a game. The circumstances may include the ball status (in play or out of play) and possession (one of the two teams), the absolute or relative spatial positions of the players and/or the ball, their movement characteristics, such as direction and speed, the events that are happening currently or have happened before, the relative time with respect to the game start and end, etc. These circumstances dynamically change during the game. We refer to a sequence of changes happening during a continuous time interval as *situation development* and to the corresponding time interval as an *episode* of situation development.

To explore and generalize the behaviors of the players and teams in certain situations, one needs to be able to select all episodes when such situations happened and developed. The selection requires appropriate query facilities for (A) specification of the situations in terms of the circumstances involved and (B) specification of the relative time intervals in which the situation development will be considered. For example, the circumstances may be "team A gains the ball possession when the ball is in the opponents' half of the pitch", and the relative time interval may be from one second before to five seconds after the situation has arisen. An earlier proposed interactive query tool called TimeMask [46] supports (A) but not (B). To support both, we propose the following extended set of query primitives:

(A) Specification of situations
Result: set of target time intervals $T_1, T_2, ..., T_N$, where $T_i = [t_i^{start}, t_i^{end}]$
- Query conditions
  - Attribute-based: selection of value intervals for numeric attributes and particular values for categorical attributes
  - Event-based: selection of particular event categories
- Condition modifier: logical NOT
- Minimal duration of a situation

(B) Specification of relative intervals
Result: set of relative time intervals $R_1, R_2, ..., R_N$, where $R_i = [r_i^{start}, r_i^{end}]$
- Relative interval start r_i^{start}: reference (t_i^{start} or t_i^{end}) and time shift $\pm\delta$, i.e., $r_i^{start} = t_i^{start} \pm \delta$ or $r_i^{start} = t_i^{end} \pm \delta$.
- Relative interval end r_i^{end}: reference (t_i^{start} or t_i^{end}) and time shift $\pm\Delta$, i.e., $r_i^{end} = t_i^{start} \pm \Delta$ or $r_i^{end} = t_i^{end} \pm \Delta$.
- Relative interval duration D and reference r_i^{start} or r_i^{end}, i.e., $r_i^{end} = r_i^{start} + D$ or $r_i^{start} = r_i^{end} - D$.

The primitives for relative interval specification allow this to be done in one of two ways: to set both the start and the end (one of the time shifts δ or Δ may be zero), or to set either the start or the end and the interval duration. Here are examples of possible specifications of relative intervals with respect to a target T:

- select X sec after T: $[R_{start}, R_{end}] \leftarrow [T_{end}, T_{end} + X]$
- select initial X sec of T: $[R_{start}, R_{end}] \leftarrow [T_{start}, T_{start} + X]$
- add X sec before and Y sec after T: $[R_{start}, R_{end}] \leftarrow [T_{start} - X, T_{end} + Y]$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2952129, IEEE Transactions on Visualization and Computer Graphics

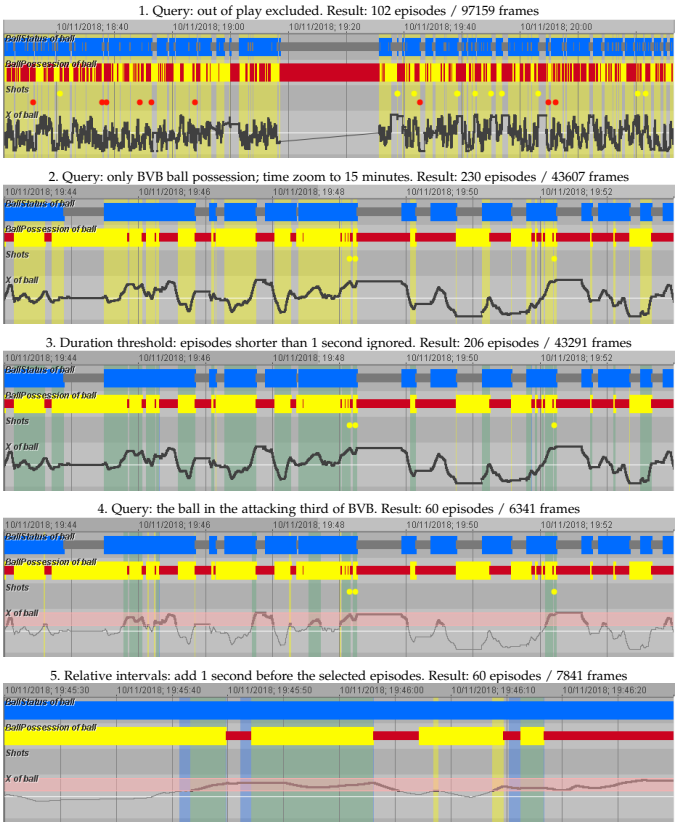

Fig. 1. A sequence of temporal query operations. Yellow vertical stripes mark time intervals selected by queries. Green is used for selected time intervals after ignoring short intervals. Blue shows interval extensions due to modifiers.

Figure 1 provides a visual illustration of a sequence of query operations for episode selection. Here and further on, we use data collected by a commercial service for the game of Borussia Dortmund and FC Bayern München [6], further called by the abbreviations BVB and FCB, respectively. BVB is usually shown in yellow and FCB in red. The data for this game span roughly over two hours (2 half times of 45+ minutes plus a break in between) with 25Hz resolution and include about 170,000 frames in total.

In the images shown in Fig. 1, the horizontal dimension represents time. In the vertical dimension, the images are divided into sections. Each section shows the variation of values of an attribute or a sequence of events. Categorical attributes are represented by segmented bars, the values being encoded in segment colors. Numeric attributes are represented by line charts. Events are represented by dots colored according to event categories. The yellow vertical stripes mark the target time intervals $T_1, T_2, ..., T_N$ selected according to the current situation specification. The blue vertical stripes mark the relative time intervals $R_1, R_2, ..., R_N$. The stripes are semi-transparent; so, the greenish color (a mixture of yellow and blue) appears where relative intervals overlap with target intervals.

The following sequence of query operations is shown: exclude the periods when the ball was out of play (Fig.1.1); select the episodes of BVB ball possession (Fig.1.2); exclude the episodes when BVB possessed the ball for less than 1 second (Fig.1.3); select the episodes in which the ball was in the attacking third of BVB (Fig.1.4); add 1-second intervals preceding the target situations (Fig.1.5).

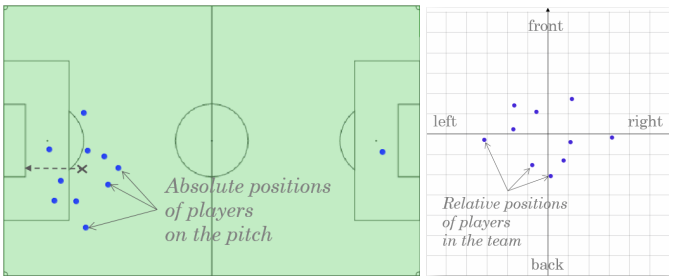

Fig. 2. A schematic illustration of a transformation from the pitch space (left) to the attacking team space (right). The coordinate grid in this and all further team spaces has 5m resolution.

Episode selection works as a temporal filter: only data from currently selected intervals $R_1, R_2, ..., R_N$ (or $T_1, T_2, ..., T_N$ if there is no relative interval specification) are treated as "active", being shown in visual displays and used in computational operations. This filter may be combined with various other filters applicable to movement data [40].

4.2 Space transformation

The space transformation is based on determining the relative positions of points of all trajectories in respect to the corresponding point of a chosen or constructed reference trajectory and its movement vector [49]. Taking into account the nature of the football game, we usually assume the movement vector to be perpendicular to the opponent's goal line, see Fig. 2. This choice can be modified by, for example, treating differently situations when the players are very close to one of the goals, when teams/players often give up their preferred formations.

A reference trajectory may be chosen or constructed in different ways depending on the character of the collective movement and analysis goals. For our goals, none of the existing individual trajectories can be used as an adequate representative of the movement of a team as a whole. Instead, we generate a central trajectory of a team by applying an aggregation operator to the positions of all players of a team, excluding the goalkeeper, at each time moment. The operator may be the mean, median, or medoid (the medoid is the point having the smallest sum of distances to all others). We have extensively tested all three options using data from several games and found that the best is to take the team's mean after excluding the positions of two most outlying field players, i.e., the most distant from the mean of the whole team, excluding the goalkeeper. The central position computed in this way changes smoothly over time, while the medoid and median positions sometimes change abruptly, which leads to sharp kinks in the resulting central trajectory. Depending on the analysis task, it may be useful to compute the central trajectories of subgroups of players, such as the defenders or midfielders, and investigate the behaviors of these subgroups.

Figure 3 demonstrates how movements on the pitch translate to movements in a team space. It presents a short episode of a single attack of FCB on the pitch and in the BVB team space. The movements of the players and the ball are represented by lines, and the tiny square symbols at the line ends show the positions at the end of the selected time interval. The goalkeeper's trajectory is marked in black. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2952129, IEEE Transactions on Visualization and Computer Graphics

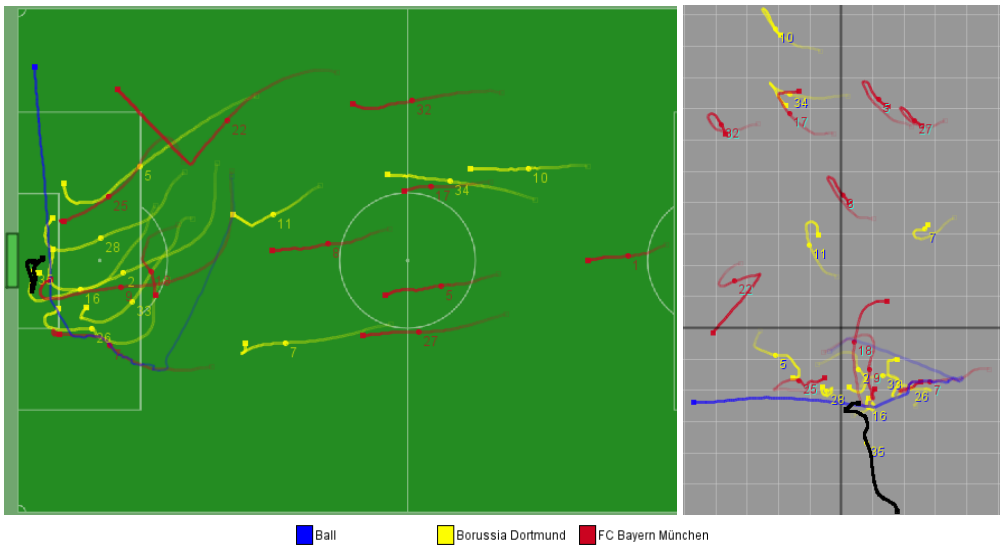

Fig. 3. One attack on the pitch (left) and in the BVB team space (right). In all illustrations, the rendering opacity varies along trajectories so that earlier segments are more transparent.

pitch map shows mostly parallel movements of all players towards the goal of BVB, but the player 22 of FCB, who initially ran in the direction of the goal, made a sharp turn to the right, and the players of BVB who were close to the goal or to the player 22 made similar movements. In the team space, the trajectories of 6 defence players of BVB located in the lower part of the team space are very short, which means high synchronization between them. In the upper part, the shapes of the trajectories belonging to the remaining 4 players of BVB and 5 players of FCB, show that the distance of these players from the team center originally increased (which means that they moved slower than the defenders) and then decreased (as the backward movement of the defenders slowed down). The diagonal orientation of these trajectories corresponds to the movement of the team center first to the left and then to the right.

While transformations to the team spaces are primarily meant for studying relative arrangements and movements of players, they create a useful by-product. By dividing a team space into meaningful zones and aggregating players' duration of presence in these zones, we obtain "fingerprints" of players' typical positions in the teams, which correspond to their roles in the game. We apply a division into a central zone (10m around the team center) and 8 areas around it. The fingerprints can be represented by *position glyphs*, as in Fig. 4. Thus, N.Süle in Fig.4, left, was present mostly on the back-left and back-center and sometimes in the central zone. Such glyphs facilitate identification of players and spotting their appearances in unusual positions and position swapping. Lines below some glyphs indicate the times the players were on the pitch (when it was not the full game) to give an idea of what changed after substitutions. Thus, N.Süle was a one-to-one substitution, which means that he exactly took M.Hummels' position after getting substituted for him. S.Wagner and R.Sanches were 1:1 substitutions to T.Müller and S.Gnabry, respectively, but interpreted their roles differently. This is visible from their aggregated positions in the team space and different distributions of the presence in their fingerprint glyphs. The display of position glyphs is optional. They were intensively used in our cases studies but are rarely present in the illustrations due to the length restrictions.

There is a possible additional use of the presence statistics by zones. The distance of a player to his average position in the team space can be treated as an indicator of how usual his current position is. These measures can be aggregated over the whole team or a selected group of players (e.g. the defence line). Based on the aggregated time series, it is possible to set query conditions for selecting episodes of unusual or usual team arrangements (Section 4.1). On the other hand, the presence statistics can be computed for different groups of selected episodes, and it is possible to choose which set of statistics to use for currently shown position glyphs and distance-to-usual-position calculations.

4.3 Aggregation

We propose a novel method for aggregating movements of an entity under different conditions. The output is a sequence of generalized positions organized in a pseudo-trajectory along an abstract timeline. Here we describe how we construct the positions and times of pseudo-trajectories.

4.3.1 *Obtaining generalized positions*

Aggregation is applied to positions selected by the current combination of data filters, including the episode selection (Section 4.1). It can be done in the pitch space and/or in the team spaces. The currently selected subset of positions of an entity is represented by a generalized position, which can be the mean, median, or medoid of the selected subset. The possibility to switch between these three options can serve as a means for checking the position variability. When the variability is small (i.e., the points are compactly clustered), the mean, median, and medoid positions are very close to each other, and switching between them does not change the general movement pattern obtained through the aggregation. Noticeable changes indicate presence of outliers. In this case, the analyst may look at the whole set of the original points and decide whether the outliers can be ignored. This is possible when the outliers are few, their positions

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2952129, IEEE Transactions on Visualization and Computer Graphics

Fig. 4. Left: examples of position glyphs for N.Süle (below) who substituted M.Hummels (above). Right: glyphs shown at the average positions of the FCB players in their team space. The blue dot represents the average position of the ball.

are randomly scattered, and the remaining points make a compact cluster. If the outliers are not negligible, they need to be examined in detail. To see the pattern formed by the remaining data, the analyst may switch to using the medoid, which is insensitive to outliers (but very costly to compute).

There is no one-size-fits-all rule for including or excluding outliers, but the decision may depend on specific contexts, such as temporary positional changes, different roles of the players in set plays, a change of the tactical system over time, etc. It is crucial that the soccer expert (match analyst, coach, assistant coach) is able to decide those things situation-specific and per interrogation of his/her own. Therefore, the user should be given a high degree of flexibility for investigations and the opportunities to exclude or include outliers and to change between more and less outlier-sensitive aggregates.

To represent the variation among the positions explicitly, we propose to build convex hulls outlining chosen percentages (e.g., 50 and 75%) of the set of original points ordered by their distances to the representative point. Examples can be seen in Fig.7 and Fig.9. Hence, a generalized position of an entity is a combination of a representative point and one or more *variation hulls*. Such a position is constructed for each entity being currently under analysis. The visual representation of the variation hulls can be controlled independently of that of the points; in particular, the hulls can be temporarily hidden for reducing the display clutter.

When the filter conditions change, aggregation can be applied to the new subset of selected positions, which produces a new set of generalized positions of all entities. The new generalized positions can be visualized together with the previous ones for comparing, as shown in Figs. 5 and 6. Figure 5 shows the aggregates in the pitch space, and Figure 6 shows the corresponding aggregates in the team space of BVB. In the left parts of the two figures, the first aggregation operation was performed using the episode filter with conditions $ballInPlay = true \ \& \ ballPossession = BVB$, the second one differs by $ballPossession = FCB$. To support the comparison, the corresponding points are connected by lines. The second points (FCB possession) are marked by dots. The differences and commonalities of players' arrangement depending on the ball possession can be easily seen. On the pitch, both teams move a bit towards the goal line. The most prominent behavioral differences happen with the wing defenders: they move wide under their own team's ball possession and narrow under the opponent's possession. In the team space, we see that the team without the ball gets more compact in both dimensions (all players tend to move towards the team center), while the team with the ball gets wider. The defenders of the attacking team move slower than the other players and thus increase team's covered area.

Another example is shown in the right parts of the Figs. 5 and 6. In the first half of the game, FCB scored a single goal at minute 26. We compare the mean positions of the players and the ball under the BVB possession before and after the goal (the latter are marked by dots), excluding the times when the ball was out of play. The pitch map shows us that both teams have shifted towards the FCB goal and a bit to the right. In the BVB team space, we see that all BVB players shifted synchronously except two central defenders (#2 and #16) who moved about 3m back from the team center. This increased the distance between the lines in the team and thus created opportunities for FCB counter-attacks. This fact may explain why the FCB player #9 moved about 5m aside in the BVB team space, searching for attacking opportunities.

According to the football experts, such representations can be very valuable to coaches by giving an overview of behavior changes after any significant game event. For seeing more than a single change and, in general, comparing generalized positions in more than two sets of situations, we construct abstract timelines for organizing multiple generalized positions in pseudo-trajectories.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2952129, IEEE Transactions on Visualization and Computer Graphics

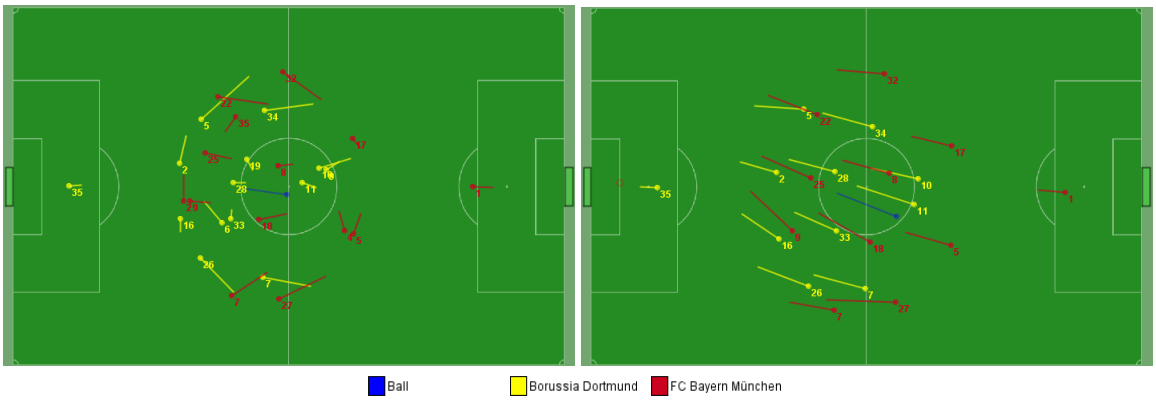

Fig. 5. Comparison of the generalized positions on the pitch in different groups of situations. Left: under the ball possession by the different teams; right: under the BVB possession before and after the goal in the first half of the game.

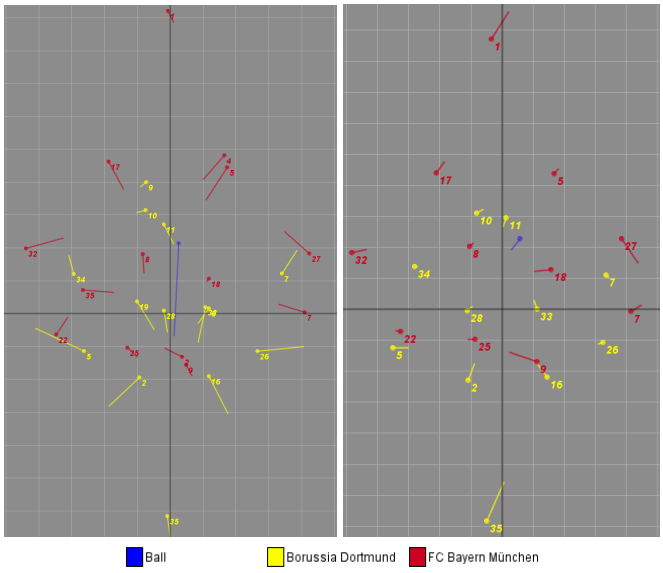

Fig. 6. Comparison of the generalized positions in the BVB team space for the same groups of situations as in Fig. 5.

4.3.2 *Creating virtual times*

A *pseudo-trajectory* consists of two or more linearly ordered generalized positions. A pseudo-trajectory is represented on a map by a line obtained by connecting consecutive positions (more precisely, their representative points). Each position of a pseudo-trajectory has an abstract numeric timestamp (1, 2, 3, ...) that equals the ordinal number of the position. Hence, a pseudo-trajectory has its internal abstract timeline made by the sequence of the position timestamps.

Pseudo-trajectories are generated by successively applying several query + aggregation operations. Each operation generates one position, which is appended after the previously generated position, if any. Hence, the order of the positions in pseudo-trajectories reflects the order of the query + aggregation operations by which they have been obtained. This very simple idea provides high flexibility for creating position sequences with different semantics, as demonstrated by an example below and further on in the case studies.

The queries used for generating pseudo-trajectories may include conditions of any kind, not necessarily time-based. Thus, the example in Fig. 7 demonstrates pseudo-trajectories of the players and the ball obtained by queries concerning the position of the BVB team center on the pitch. We made a sequence of 10 queries with a common condition $ballPossession = BVB$ and the differing conditions referring to the position of the BVB team center along the X-axis with respect to the pitch center: $x < -40m, -40m \leq x < -30m, ..., x \geq 40m$. The queries did not include explicit time constraints, but each query selected a set of time intervals when the x-coordinate of the team center was in a specific range.

Two upper images in Fig. 7 show the footprints of the pseudo-trajectories as lines on the pitch (left) and in the BVB team space (right). On top of the lines corresponding to the players, the position glyphs of the players are shown. The glyphs are drawn at the middle positions of the players' pseudo-trajectories, which correspond to the BVB team center position being in the interval $[-10..0)$ meters. The relative arrangement of the glyphs of each team reflect the formation used by BVB for preparing an attack and the defensive formation 4-4-2 of FCB. We can also see consistent monotonous changes of the player's positions from left to right and changes of the teams' widths along the pitch. Complementary to this, the team space demonstrates changes in the team compactness in both dimensions.

For illustrative purposes, the remaining images in Fig. 7 include the 50% variation hulls for selected 5 players of BVB. The hulls are shown in the pitch space and the BVB team space in 2D maps and 3D space-time cubes. The images of the space-time cubes are provided for merely illustrative purposes, to demonstrate that the pseudo-trajectories and the hulls are spatio-temporal objects, albeit the time domain in which they exist is abstract rather than real. These objects can be treated in the same ways as "normal" spatio-temporal objects existing in real time domains.

The colors of the variation hulls from violet through yellow to orange depict the virtual times of the positions. We can observe stable shapes and sizes of the hulls and their very stable locations in the team space, except for the two wing defenders who first moved outwards from the center and then back towards the center. Generally, such stability checks are necessary for all aggregates, and they were consistently performed during the analysis. However, the page limit does not permit having many such illustrations in the paper.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2952129, IEEE Transactions on Visualization and Computer Graphics

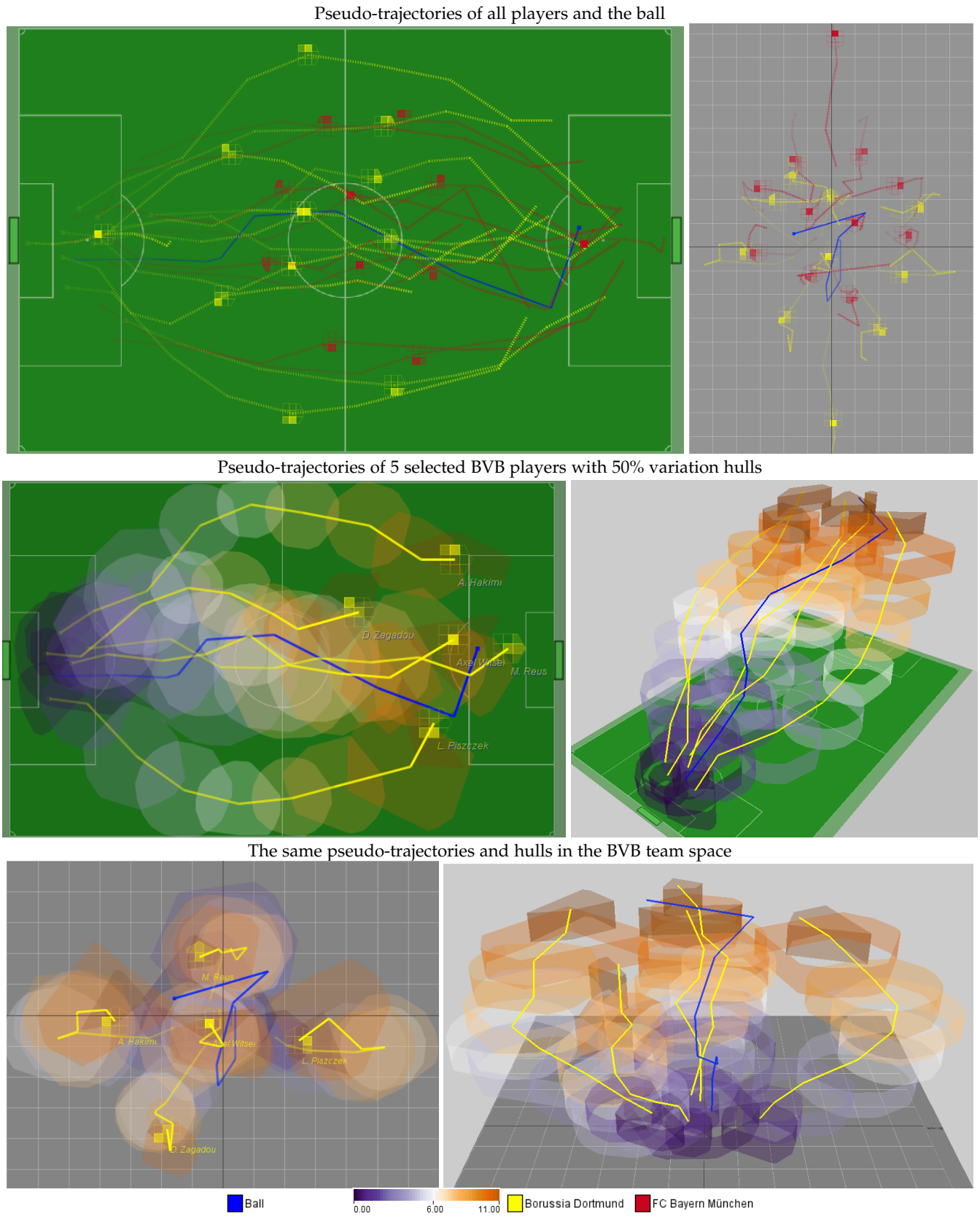

Fig. 7. Sequences of 10 generalized positions of the players and the ball under the BVB ball possession corresponding to different positions of the BVB's team center on the pitch along the X-axis.

4.4 Interaction between the framework components

Since our paper aims at presenting the general framework, which can be implemented in different ways, rather than a specific implementation, we refrain from describing techniques for user-computer interaction, which can differ between possible implementations. What we describe here is how the components of the general framework are supposed to work together within the analysis process.

The process begins with creation of the team spaces (Section 4.2). Then, the following sequence of steps is repeatedly executed:

1. Temporal query (Section 4.1):
   1.1. Specify and find situations of interest (Section 4.1(A)).
   1.2. Specify a sequence of relative intervals for extracting the situation development episodes (Section 4.1(B)).

   Result: set of episodes.
2. Aggregation (Section 4.3):
   2.1. Automatically aggregate the query result and generate pseudo-trajectories in the pitch space and in the team spaces.
   2.2. Put the pseudo-trajectories as new information layers in the respective spaces.

   Result: set of pseudo-trajectories.
3. Visualization and comparative analysis:
   3.1. Represent the pseudo-trajectories within each space in an interactive visual display using techniques suitable for ordinary trajectories. The displays need to be linked through common visual encoding and by interactive techniques, such as brushing [55].
   3.2. Compare the new set of pseudo-trajectories with one or more of the previously obtained sets resulting from other temporal queries.

Our group has performed this analytical process in the case studies described in Section 5.

4.5 Evaluation of the framework

In our research project, it was not intended to design and develop software tools according to specific users' tasks and requirements. For developing and testing the components of the framework, the partners specializing in visual analytics and data science, further referred to as *the analysts*, utilized existing software tools. Among others, they used a research-oriented software system V-Analytics (http://geoanalytics.net/V-Analytics/). The analysts extended its base functionality by implementing new query, aggregation, and data transformation techniques. These software developments were necessary for achieving the research goals, but the key result of the project is the analysis methods and not the tools.

In the course of the research, the methods under development were constantly evaluated by the football domain experts according to the following criteria: (A) the possibility to select multiple situations with common properties and the flexibility in specifying the properties of interest; (B) the possibility to extract, visualize, and interpret general patterns of situation development. To assess the query facilities (A), the experts described the situations that were interesting for them, and the analysts translated the descriptions into queries, extracted the corresponding portions of the data, and provided the experts with tools for sampling some of the selected situations and reviewing them with the use of animated maps and corresponding fragments of the game video. To assess the situation generalization facilities (B), the experts were provided with visual displays of the pseudo-trajectories, which represented the extracted patterns.

The evaluation was carried out in a series of case studies, in which the experts set the analysis tasks, the analysts performed operations according to the framework, and the football experts interpreted and evaluated the results and posed further questions.

5 Case studies

The objective of professional football is to win matches and entertain the public. To win matches, you have to score more goals than the opponent. This requires a good balance between the offensive and defensive strategies of a team. That is why it is important that a team has a tactical plan defining the desired team formations and behavior in different states of the game. Examples of states are own and opponent's ball possession, transitions between them after loosing the ball or recovering it, counter attacks, set pieces, such as corners etc. In the following, we consider several categories of situations that were interesting to analyze for our football experts. We use data from two Bundesliga games [6], [7] of the season 2018-19. The data from each game consist of two parts: (1) trajectories of the players and the ball, i.e., sequences of their positions in the pitch recorded every 40 milliseconds, and (2) records of the game events with their attributes, including the event type, time of occurrence, and players involved; see Section 2.2.

5.1 Ball possession change

The two switching moments between own and opponent's possession are getting more and more attention in football. The reason is that the desired field occupancy of the players and team tactics in situations where the team possesses the ball is completely different from the ideal field occupation and tactics in situations in which the opponent has the ball. As soon as a team loses the ball, the field occupation is often disorganized from a defensive perspective. It takes time for a team to adapt to the new situation ('switching cost') in which it has to apply its defensive tactics. This temporary 'chaos' is something the team with ball possession can take advantage of.

5.1.1 Checking the five-seconds rule

The famous P.Guardiola's 5 seconds rule for successful counterpressing says [56]: *"After losing the ball, the team has five seconds to retrieve the ball, or, if unsuccessful, tactically foul their opponent and fall back"*, that has been key to Manchester City conceding considerably fewer goals. The football experts were interested to see if BVB and FCB applied this tactics in the game [6].

To extract the episodes of interest, the analysts applied temporal query operations described in Section 4.1. They first selected the moments of change in the ball possession, excluding those when the ball got out of play. Next, they consecutively applied a series of operations for specification

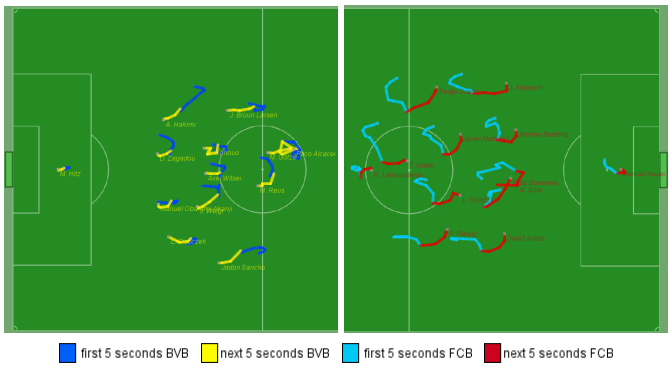

Fig. 8. Pseudo-trajectories of the players during 5+5 seconds after loosing the ball. Left: BVB, blue for the starting 5 sec and yellow for the following. Right: FCB, in cyan and red, respectively.

of relative intervals:
$[R_{start}, R_{end}] \leftarrow [T_{end} + X sec, T_{end} + (X+1)\ sec]$,
$X = 0, 1, 2, ..., 9$
producing 10-steps long pseudo-trajectories from the mean positions and their 50% variation hulls. During the aggregation, irrelevant parts of the original trajectories that were shorter than 10 seconds (e.g. due to the ball going out of play or another change of possession) were discarded by an attribute-based query.

Figure 8 shows the results of the aggregation separately for BVB (left) and FCB (right). The pseudo-trajectories of the players are painted in two contrasting colors corresponding to the first 5 seconds (blue for BVB and cyan for FCB) and to the following 5 seconds (yellow for BVB and red for FCB). The images show that almost all BVB players do not move back during the first 5 seconds and gradually fall back in the next 5 seconds. This agrees with the fact that BVB is known for their pressing style of playing. The patterns of the FCB players are different. The players continue moving forward for the initial 2 seconds on the average and then start moving to the left and back.

Figure 9 shows the players' pseudo-trajectories and 50% variation hulls in the team spaces. The colors of the hulls encode their relative times. For BVB, the hull colors vary from dark blue to dark yellow, so that the shades of blue correspond to the first 5 seconds and the shades of yellow to the following 5 seconds. For FCB, the hull colors vary from dark cyan to dark red, respectively. The images show that BVB tended to reduce the team width and depth whereas FCB kept the width constant while slightly reducing the depth. The stacks of the hulls with the colors representing the relative times show that the hulls of the majority of the players were getting notably smaller over time. This means that the players were successfully reconstructing their planned defensive formations and then were keeping their intended relative positions. This is in accordance to the general football philosophy that suggests creativity while attacking and organization and structure while defending.

5.1.2 Comparison of behaviors in two games

For experts it is important to understand how a team adapts to different opponents and circumstances. To contrast the game 3:2 FC Bayern [6], they want to compare BVB's behavior with their 7:0 game against FC Nürnberg (FCN) [7]. Following the procedure described in Section 5.1.1, in each

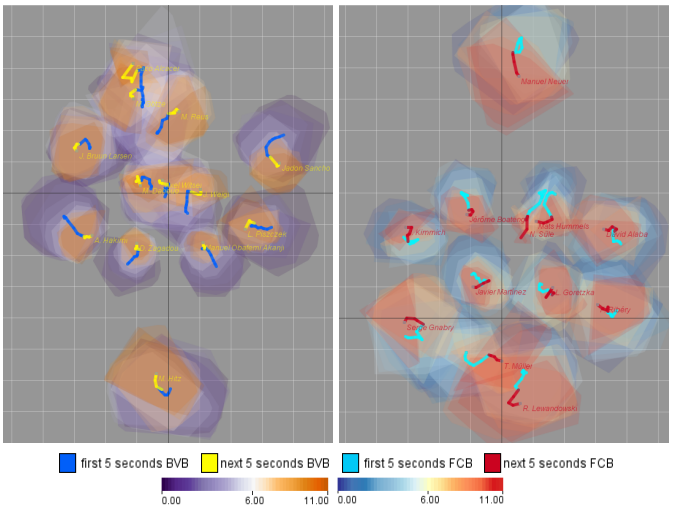

Fig. 9. The same aggregated data as in Fig.8 are presented in the team spaces (left: BVB, right: FCB) using the same colors together with the 50% variation hulls.

game for each player the analysts constructed two 10-steps long pseudo-trajectories summarizing the transitions to the attacking and defensive formations (Figs. 10 and 11).

The images on the top of Fig. 10 and on the left of Fig. 11 correspond to the game against FC Nürnberg, the other two images correspond to the game against FCB considered earlier. The BVB team tactics when loosing and regaining the ball are represented in black and yellow, respectively. The pitch images show that, while in the game against FCB (right) BVB players pressured for 5 seconds after loosing the ball, in the game against FCN they were falling back during the initial 5 seconds and only then put pressure on the opponents. The team was narrowing down in the FCB game but kept constant width in the FCN game. After ball regaining, BVB players attempted to perform fast counter attacks against FCB but preferred careful, slowly moving forward attack preparations against FCN.

The team space images are useful for seeing the synchronization among the players and changes of their arrangement. Game against FCN (left): synchronous movement in the transition to defense, except the central forwards, and widening of the team in the transition to offense. Game against FCB: increasing compactness of the team for defense and fast expansion in both direction for offense. The team depth in the game against FCN was about 30m in contrast to about 40m against FCB.

5.2 Long passes

One of the most important instruments for advancing the attack, finding open spaces on the pitch, and forcing opponents to make mistakes, is long passes. After examining the distribution of the pass lengths with respect to the X-axis (i.e., how much the ball moved in the direction to the opponents' goal) in the game [6], the experts and analysts jointly chose a query threshold of minimum 20m for selecting long passes. After inspecting the selected passes, the experts understood that the passes made from the own goal box need to be excluded. Figure 12 shows the footprints of the remaining selected long balls. To produce them, the analysts used two queries based on the starting and ending

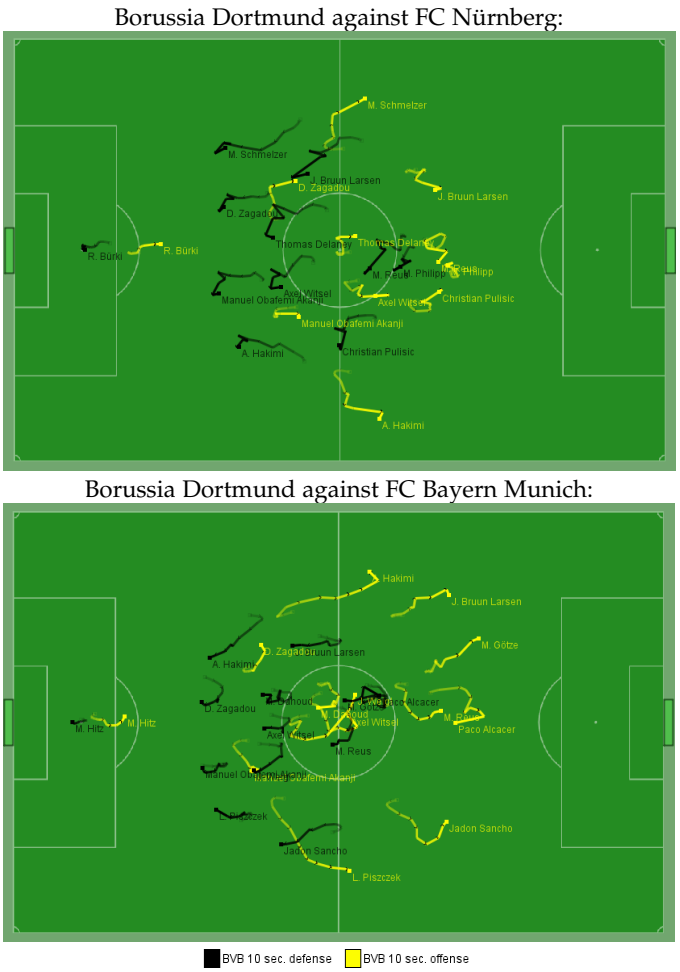

Fig. 10. Comparison of movements of Borussia Dortmund players in ball possession transition periods in games against FC Nürnberg [7] (top) and FC Bayern Munich [6] (bottom). Only players who were present on the pitch for at least 30 minutes are shown. Black lines show transitions to defense: from -1 to +10 seconds after loosing the ball control; yellow lines represent transitions to offense: from -1 to +10 seconds after gaining the ball. Different tactical patterns appear prominently (see explanations in section 5.1.2 for details).

moments of the selected passes:
$[R_{start}, R_{end}] \leftarrow [T_{start} - \delta, T_{start} + \delta]$ and
$[R_{start}, R_{end}] \leftarrow [T_{end} - \delta, T_{end} + \delta]$,
where $\delta = 0.2sec$ was applied for tolerating possible mismatches in the times between the position data and manually annotated passes. To include more information about the conditions in which these long balls were made, the analysts visualized the pass lines together with aggregated positions of all players (colored dots) on top of the density field summarizing the distributions of all players on the pitch during the execution of the selected passes.

An interesting diagonal configuration of the players during the long balls of BVB can be seen in Fig. 12, top. The passes were sent either along or across this dense concentration. Most of the long passes of both teams were directed to the left flank of BVB and right flank of FCB, so the same side of the pitch was used actively by both teams. These passes were very important in this game as they were involved in 3 out of 5 attacks that resulted in goal scoring. To consider them separately, the analysts added spatial filters by the pass destinations and obtained aggregates for the subset of the passes (Fig. 13).

It is interesting to observe that, although the pass targets were quite widely distributed on the pitch, they were compact in the spaces of the defending teams. All BVB passes were targeted in the area behind the FCB's right central defender J.Boateng, on the average 5 meters behind and 10 meters aside of him. He had to move back during these passes, breaking the last defensive line. The FCB passes targeted at a point about 25-30 meters aside of the BVB team center. These passes forced the defending team to shift left.

It can be concluded that the long forward passes of BVB were intended to make immediate danger to the goal. The attacking group of the BVB players moved far forward during these passes. The shape of the team became long but rather narrow. The long passes of FCB resulted in changes of the attacking direction with the players moving to the right. It should be noted that FCB striker R.Lewandowski was balancing around the offside line at the moment of the reception of the selected long passes, so it would be dangerous to pass to him immediately.

5.3 Building up for shots

Goals are scored after successful shots, which require not only high individual skills of a striker but also work of the whole team for reaching situations in which good shots become possible. To help the experts to investigate this teamwork in the BVB-FCB game, the analysts used queries to select, first, the moments of the shots and, second, the episodes preceding them. They made a series of queries $[R_{start}, R_{end}] \leftarrow [T_{start} - X\ sec, T_{start} - (X-1)\ sec]$, $X = 10, 9, 8, ..., 1$, where T_{start} is the moment of a shot, and obtained the corresponding sets of pseudo-trajectories of the ball, team centers, and all players separately for the shorts of BVB and FCB. As a measure of position variation, the analysts also computed the median distances of the representative points of the generalized positions to the original positions from which they had been derived. Since changes of the ball possession could happen during the 10-seconds intervals before the shots, the analysts applied attribute filters discarding irrelevant parts of the episodes in which the build up was shorter than 10 seconds.

The pseudo-trajectories of the players and the ball are shown in Fig. 14 and the pseudo-trajectories of the team centers in Fig. 15. These two figures demonstrate different levels of aggregation and abstraction applied to the same data. The position variation indicators associated with the points of the pseudo-trajectories are represented by proportional sizes of the dots thus marking 1-second segments. Please note that this is an abstract, symbolic representation using the visual variable 'size' to encode numeric values. The sizes of the symbols are not related to the map scale. This representation is essentially different from the representation of the variation by hulls, which are spatial objects. Unlike the dot sizes, the hulls occupy particular areas in space and have particular shapes.

The individual aggregates in Fig. 14 and team aggregates in Fig. 15 consistently show that the two teams tended to use different ways to reach their opponents' goal. FCB mostly used the right flank and then turned towards the center of the penalty box. The overall shape of the BVB attacks looks like an arrow targeted straight at the FCB goal.

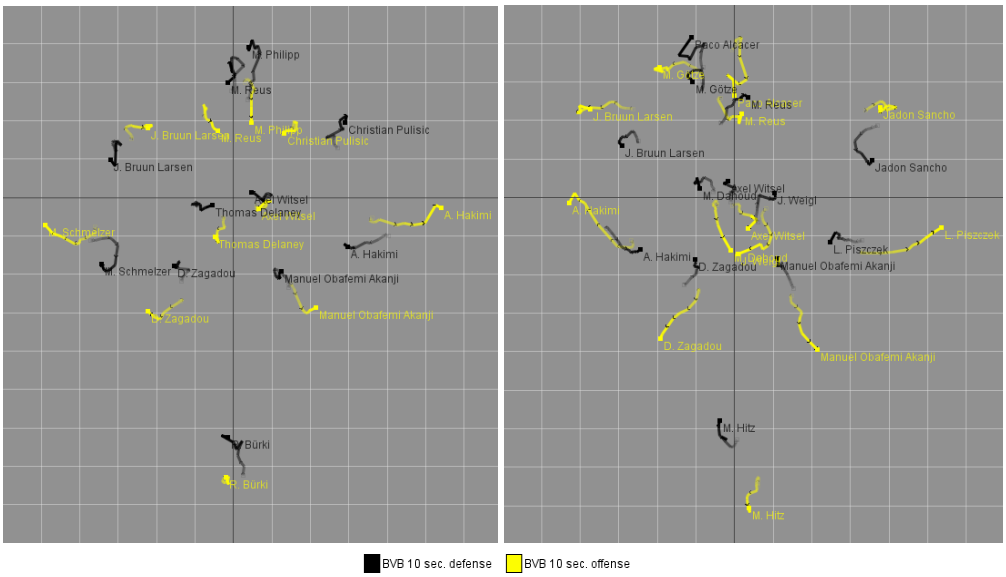

Fig. 11. Aggregated movements of BVB players in their team space after ball possession changes in games [7] (left) and [6] (right).

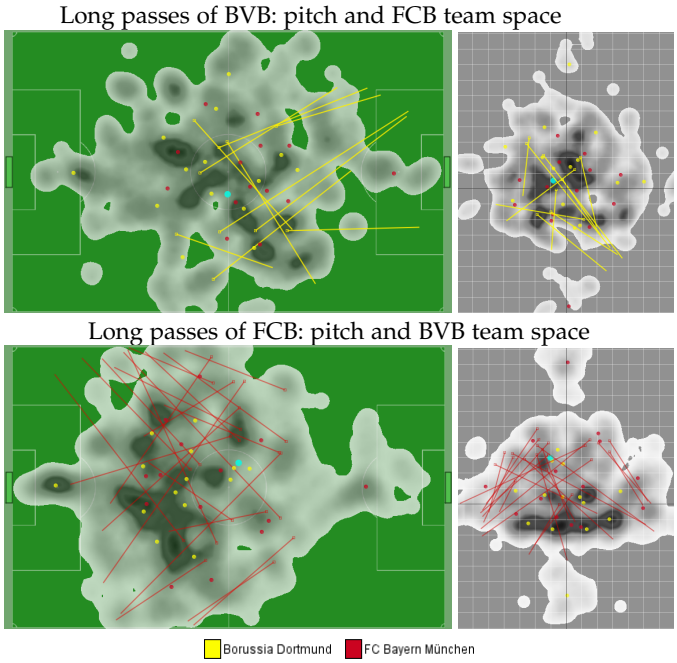

Fig. 12. Long forward passes of BVB (top) and FCB (bottom) and the average positions of all players (shown by colored dots) on top of the density fields of all players during the passes.

The position variation indicators (i.e., the median distances to the representative positions) can be compared along and across the pseudo-trajectories. For example, the variation of the positions of the FCB's right midfielder is much higher than that of the left midfielder. This can be related to the fact that the ball was transferred to the penalty box mostly from the right flank, and the opponents were putting more pressure on that side forcing the right midfielder to vary his positions.

Our group considered two options for studying further details for smaller subsets of similar shots. One option was grouping by the shot location. However, by inspecting the episodes preceding the shots, it was found that the variation of the shot positions does not match the variation of the trajectories of the ball, teams, and players. Similar build-ups do not necessarily lead to making shots from similar positions. Another option was to cluster the shots by similarity of the last preceding passes or by similarity of particular trajectories, e.g., of the ball and/or the team centers. We evaluated several variants of grouping using clustering of trajectories by relevant parts [54]. They produced either heterogeneous clusters with too small differences between them or homogeneous clusters that were too small for valid generalization. This procedure, however, appears to have a good potential when applied to a larger number of situations extracted from multiple games of the same team.

5.4 Conclusions from the use cases

Even if the top leagues and clubs are aware of the necessity of acquiring event and tracking data, the potential of using them in the team's daily business is not yet tapped. Since the game is very complex and the interpretation of situations is very subjective even for experts, a lot of data-driven projects fail in terms of communication between data-science and soccer experts. The football experts concluded that the proposed VA approach is a great step towards making the complex spatio-temporal tracking and event data understandable and so usable for professionals.

Application of visual analytics approaches allowed our group to find many interesting patterns that would be very difficult or even impossible to detect by means of watching game footage on TV or an animated visual representation of the data. We were able to identify patterns, compare them, and synthesize further higher-level patterns. Moreover, obtaining interesting and meaningful results motivated many further analysis scenarios such as identifying formations during long periods of ball possession, assessing efficiency of counter-pressing efforts, comparing evolution of playing style of the same team over a season and across multiple seasons, searching for impact of changes of team coaches and/or key players. Considering such scenarios would be out of question if only the state-of-the-art techniques were available.

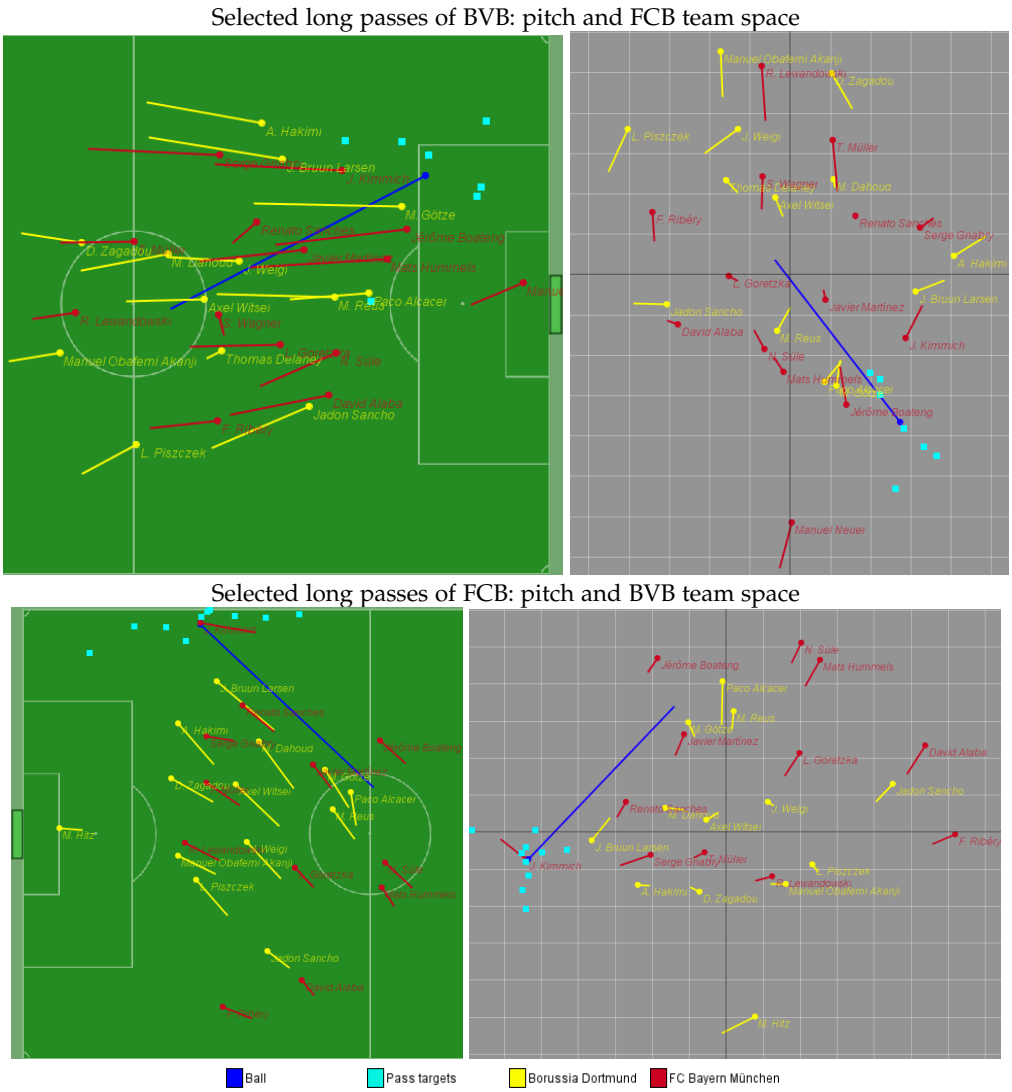

Fig. 13. Changes in aggregated positions of players and the ball during the selected long passes of BVB (top) and FCB (bottom). Pass targets are marked by cyan dots.

6 Discussion

This paper presents results of a research project involving visual analytics researchers, data scientists, and football domain experts with the aim to find effective approaches to gaining practicable knowledge from real complex data. The football experts were impressed by the capabilities of the visual analytics techniques that were developed. They said that the selection of similar game situations based on underlying data and extraction of general patterns of the teams' and players' behaviors in such situations has been so far an unsolved challenge. Hence, appropriate query and generalization techniques would bring a big benefit for experts, especially for scouts and match analysts. The proposed framework has a very high potential to bring all the data-insights finally on the pitch and thus produce a substantial impact on professional football.

In the experts' opinion, the power of the framework can be further increased by involving Key-Performance-Indicators (KPI) for the soccer game metrics such as expected goals, dangerosity, pass options based either on measuring the difficulty of performing passes or representing possible gain if a pass is successful (e.g. packing rate), indicators of team compactness and structure such as space occupation, team shape damage, stretch index etc., and pressure indicators. These KPIs are on the rise and could be utilized in making episode queries to both validate and improve the metrics and to select game situations of interest, enabling further application scenarios.

To put the results of this research into practice, it is necessary to develop software tools that can be easily utilized by end users. As there exist different categories of potential users, as discussed below, user-centered design and development may need to be done specifically for each category, taking into account the specific tasks, requirements, and capabilities of the target users. Different classes of users need different interfaces with different levels of interactivity and visual complexity, but, irrespective of these, achieving high level of automation in making queries, constructing pseudo-trajectories, and putting them in visual displays is of great importance. There exists a technical possibility to implement the presented framework in the form of automated procedures oriented to specific analysis tasks. Such automated procedures can be used for extracting patterns from large databases containing data from many games.

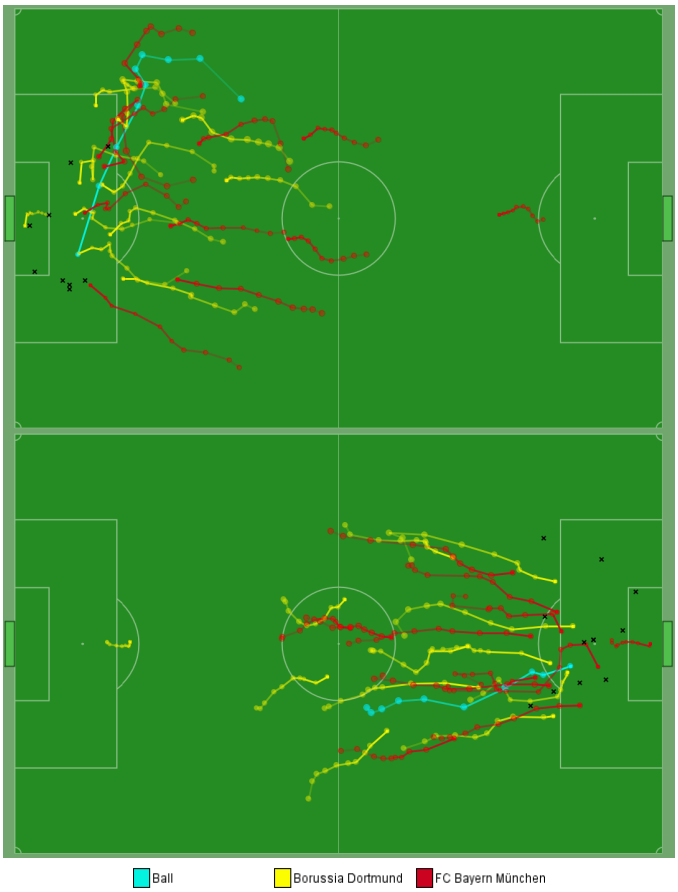

Fig. 14. Build-up for the shots by FCB (top) and BVB (bottom). The shot positions are marked by black crosses. The cyan line corresponds to the ball. The dots with the sizes representing the variation mark the generalized positions, which are separated by 1-second intervals.

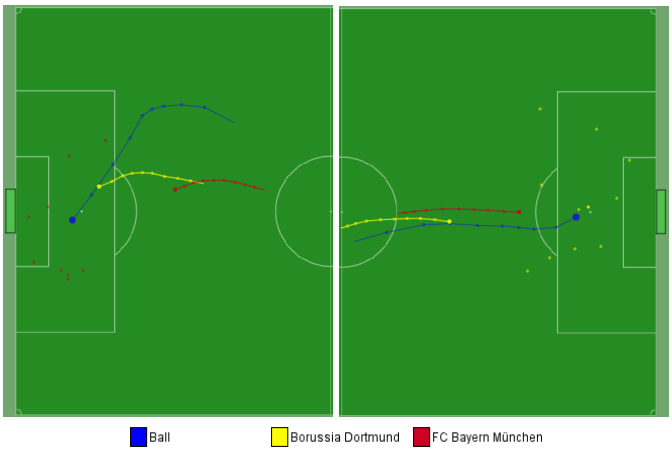

Fig. 15. Build-ups for shots: pseudo-trajectories of the team centers before the FCB shots (left) and BVB shots (right).

We anticipate that the following user categories could benefit from specific applications based on the framework. **Match analysts** could evaluate the efficiency of their own team in previous games and create automatically a catalogue of tactical schemes of opponents over a big set of their previous games. Such automatically acquired tactical schemes conditioned over different classes of situations could be a great hint-giving and decision-supporting means. **Medical staff** of clubs could examine movements of players during episodes characterized by fast running at different times in the games. **Scouts** could evaluate players' movements and actions in different classes of situations and spot their strong and weak abilities. **Journalists** could present tactical schemes and compare them in pre-game and post-game articles or TV shows or even during game breaks. **Leagues** could provide services to clubs and also enhance their media products. Some user categories (particularly, the latter two) require tools not only for analysis but also for communication of the insights gained to certain audiences, including the general public. This requires specific approaches for synthesizing audience-targeted stories from results of tactical analyses [57].

While the presented framework have been developed with an orientation to football data and analysis tasks, it is potentially generalizable to various kinds of coordinated movements of multiple objects in applications where the task of extracting general behaviour patterns under different circumstances is relevant. Examples are movements of players in other team sports, such as ice-hockey or basketball, behaviors of animal groups, or movements of people in crowded environments. We also envision potential applications in domains of air and sea traffic management. The main components of the approach, i.e., the query facilities for episode selection, the method for generalization and aggregation, the data structure for representing generalization results, and the visualization techniques, can be adjusted to the specifics of various application domains.

7 Conclusion

Our contribution can be summarized as proposing an analytical framework involving interactive **queries**, **generalization and aggregation** of query outcomes, and **comparative visual exploration** of resulting general patterns. The framework makes use of an interesting and fruitful interplay of physical and constructed spaces (pitch and team spaces) and times (absolute and relative times). The query primitives enable selection of sets of time intervals containing situations with specified characteristics and, moreover, further selection of sets of intervals having particular temporal relationships to the previously selected intervals. This can be used, in particular, for considering the episodes of situation development step-wise or for studying what happened before or after them. The aggregation method produces a novel type of movement aggregate, pseudo-trajectory, consisting of generalized positions arranged along an abstract timeline. The aggregation results are visualized in ways enabling exploration, comparison, and assessment of the variation of the original movements summarized in the aggregates. The techniques proved useful for discovery of general patterns of collective movement behavior in diverse classes of situations.

Acknowledgments

This research was supported by Fraunhofer Cluster of Excellence on "Cognitive Internet Technologies" and by EU in project SoBigData.

REFERENCES

[1] Wikipedia, *Association football*, 2019 (accessed February 14, 2019), https://en.wikipedia.org/wiki/Association_football.

[2] D. Memmert and D. Raabe, *Revolution im Profifußball. Mit Big Data zur Spielanalyse 4.0*. Springer, 2017.

[3] D. Link, *Data Analytics in Professional Soccer*. Springer, 2018.

[4] D. Sumpter, *Soccermatics: mathematical adventures in the beautiful game*. Bloomsbury Publishing, 2016.

[5] J. Ladefoged, *How Data (and Some Breathtaking Soccer) Brought Liverpool to the Cusp of Glory*, 2019, https://www.nytimes.com/2019/05/22/magazine/soccer-data-liverpool.html (accessed June 18, 2019).

[6] Bundesliga, *Borussia Dortmund - FC Bayern München 3:2*, 10.11.2018 (accessed February 14, 2019). [Online]. Available: https://www.bundesliga.com/de/bundesliga/spieltag/2018-2019/11/borussia-dortmund-vs-fc-bayern-muenchen/stats

[7] ——, *Borussia Dortmund - 1. FC Nürnberg 7:0*, 26.09.2018 (accessed February 14, 2019). [Online]. Available: https://www.bundesliga.com/de/bundesliga/spieltag/2018-2019/5/borussia-dortmund-vs-1-fc-nuernberg/stats

[8] Wikipedia, *Formation (association football)*, 2019 (accessed February 14, 2019), https://en.wikipedia.org/wiki/Formation_(association_football).

[9] G. Andrienko, N. Andrienko, G. Budziak, J. Dykes, G. Fuchs, T. von Landesberger, and H. Weber, "Visual analysis of pressure in football," *Data Mining and Knowledge Discovery*, vol. 31, no. 6, pp. 1793–1839, 2017.

[10] J. Wilson, *Inverting the pyramid: the history of soccer tactics*. Nation Books, 2013.

[11] A. Zauli, *Soccer: Modern Tactics*. Reedswain Inc., 2002.

[12] M. Lucchesi, *Attacking soccer: A tactical analysis*. Reedswain Inc., 2001.

[13] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. A. Matthews, "Large-scale analysis of soccer matches using spatiotemporal tracking data," in *2014 IEEE International Conference on Data Mining, ICDM 2014, Shenzhen, China, December 14-17, 2014*, R. Kumar, H. Toivonen, J. Pei, J. Z. Huang, and X. Wu, Eds. IEEE, 2014, pp. 725–730.

[14] A. Bialkowski, P. Lucey, G. P. K. Carr, Y. Yue, S. Sridharan, and I. A. Matthews, "Identifying team style in soccer using formations learned from spatiotemporal tracking data," in *2014 IEEE International Conference on Data Mining Workshops, ICDM Workshops 2014, Shenzhen, China, December 14, 2014*, Z. Zhou, W. Wang, R. Kumar, H. Toivonen, J. Pei, J. Z. Huang, and X. Wu, Eds. IEEE, 2014, pp. 9–14.

[15] J. Perl, A. Grunz, and D. Memmert, "Tactics analysis in soccer–an advanced approach," *International Journal of Computer Science in Sport*, vol. 12, no. 1, pp. 33–44, 2013.

[16] Y. Wu, X. Xie, J. Wang, D. Deng, H. Liang, H. Zhang, S. Cheng, and W. Chen, "Forvizor: Visualizing spatio-temporal team formations in soccer," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 65–75, Jan 2019.

[17] ChyronHego, *https://chyronhego.com/*, 2019 (accessed June 18, 2019).

[18] OPTA, *https://www.optasports.com/*, 2019 (accessed February 14, 2019).

[19] STATS, *https://www.stats.com/*, 2019 (accessed February 14, 2019).

[20] S. Spectrum, *https://www.secondspectrum.com/*, 2019 (accessed June 18, 2019).

[21] Track160, *https://track160.com/*, 2019 (accessed June 18, 2019).

[22] FootoVision, *https://www.footovision.com/*, 2019 (accessed February 14, 2019).

[23] G. Andrienko, N. Andrienko, and G. Fuchs, "Understanding movement data quality," *Journal of Location Based Services*, vol. 10, no. 1, pp. 31–46, 2016.

[24] D. J. Sumpter, *Collective animal behavior*. Princeton University Press, 2010.

[25] J. Gudmundsson and M. Horton, "Spatio-temporal analysis of team sports," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, p. 22, 2017.

[26] C. Perin, R. Vuillemot, and J.-D. Fekete, "Soccerstories: A kick-off for visual soccer analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2506–2515, 2013.

[27] D. Sacha, F. Al-Masoudi, M. Stein, T. Schreck, D. Keim, G. Andrienko, and H. Janetzko, "Dynamic visual abstraction of soccer movement," *Computer Graphics Forum*, vol. 36, no. 3, pp. 305–315, 2017.

[28] D. Sacha, M. Stein, T. Schreck, D. A. Keim, O. Deussen *et al.*, "Feature-driven visual analytics of soccer data," in *IEEE Conference on Visual Analytics Science and Technology*. IEEE, 2014, pp. 13–22.

[29] L. Shao, D. Sacha, B. Neldner, M. Stein, and T. Schreck, "Visual-interactive search for soccer trajectories to identify interesting game situations," *Electronic Imaging*, vol. 2016, no. 1, pp. 1–10, 2016.

[30] M. Stein, H. Janetzko, T. Schreck, and D. A. Keim, "Tackling Similarity Search for Soccer Match Analysis: Multimodal Distance Measure and Interactive Query Definition," in *Symposium on Visualization in Data Science (VDS) at IEEE VIS 2018*, 2018.

[31] M. Stein, H. Janetzko, T. Breitkreutz, D. Seebacher, T. Schreck, M. Grossniklaus, I. D. Couzin, and D. A. Keim, "Director's cut: Analysis and annotation of soccer matches," *IEEE Computer Graphics and Applications*, vol. 36, no. 5, pp. 50–60, 2016.

[32] M. Stein, J. Häußler, D. Jäckle, H. Janetzko, T. Schreck, and D. A. Keim, "Visual soccer analytics: Understanding the characteristics of collective team movement based on feature-driven analysis and abstraction," *ISPRS International Journal of Geo-Information*, vol. 4, no. 4, pp. 2159–2184, 2015.

[33] G. Andrienko, N. Andrienko, G. Budziak, T. von Landesberger, and H. Weber, "Coordinate transformations for characterization and cluster analysis of spatial configurations in football," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2016, pp. 27–31.

[34] J. Gudmundsson and T. Wolle, "Football analysis using spatio-temporal tools," *Computers, Environment and Urban Systems*, vol. 47, pp. 16–27, 2014.

[35] M. Horton, J. Gudmundsson, S. Chawla, and J. Estephan, "Automated classification of passing in football," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2015, pp. 319–330.

[36] M. Stein, H. Janetzko, A. Lamprecht, T. Breitkreutz, P. Zimmermann, B. Goldlcke, T. Schreck, G. Andrienko, M. Grossniklaus, and D. A. Keim, "Bring it to the pitch: Combining video and movement data to enhance team sport analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 13–22, Jan 2018.

[37] S. E. Viewer, *https://www.stats.com/edge/*, 2019 (accessed June 18, 2019).

[38] S. S. One, *https://www.sap.com/germany/products/sports-one.html*, 2019 (accessed June 18, 2019).

[39] S. Solutions, *http://www.bundesliga-datenbank.de/*, 2019 (accessed June 18, 2019).

[40] G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel, *Visual Analytics of Movement*. Springer, 2013.

[41] G. Andrienko, N. Andrienko, W. Chen, R. Maciejewski, and Y. Zhao, "Visual analytics of mobility and transportation: State of the art and further research directions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 8, pp. 2232–2249, Aug 2017.

[42] C. Weaver, "Cross-filtered views for multidimensional visual analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 2, pp. 192–204, March 2010.

[43] H. Hochheiser and B. Shneiderman, "Dynamic query tools for time series data sets: Timebox widgets for interactive exploration," *Information Visualization*, vol. 3, no. 1, pp. 1–18, Mar. 2004.

[44] S. K. Gadia, "A homogeneous relational model and query languages for temporal databases," *ACM Trans. Database Syst.*, vol. 13, no. 4, pp. 418–448, Oct. 1988.

[45] C. S. Jensen, J. Clifford, S. K. Gadia, A. Segev, and R. T. Snodgrass, "A glossary of temporal database concepts," *SIGMOD Rec.*, vol. 21, no. 3, pp. 35–43, Sep. 1992.

[46] N. Andrienko, G. Andrienko, E. Camossi, C. Claramunt, J. M. C. Garcia, G. Fuchs, M. Hadzagic, A.-L. Jousselme, C. Ray, D. Scarlatti, and G. Vouros, "Visual exploration of movement and event data with interactive time masks," *Visual Informatics*, vol. 1, no. 1, pp. 25 – 39, 2017.

[47] T. von Landesberger, S. Bremm, T. Schreck, and D. W. Fellner, "Feature-based automatic identification of interesting data segments in group movement data," *Information Visualization*, vol. 13, no. 3, pp. 190–212, 2014.

[48] N. Andrienko, G. Andrienko, J. M. C. Garcia, and D. Scarlatti, "Analysis of flight variability: a systematic approach," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 54–64, Jan 2019.

[49] N. Andrienko, G. Andrienko, L. Barrett, M. Dostie, and P. Henzi, "Space transformation for understanding group movement," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2169–2178, Dec 2013.
[50] G. Andrienko, N. Andrienko, H. Schumann, and C. Tominski, *Visualization of Trajectory Attributes in Space–Time Cube and Trajectory Wall*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 157–163.
[51] N. Andrienko and G. Andrienko, "Spatial generalization and aggregation of massive movement data," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 2, pp. 205–219, 2011.
[52] N. Willems, H. Van De Wetering, and J. J. Van Wijk, "Visualization of vessel movements," *Computer Graphics Forum*, vol. 28, no. 3, pp. 959–966, 2009.
[53] G. Andrienko and N. Andrienko, "Spatio-temporal aggregation for visual analysis of movements," in *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (IEEE VAST) 2008*, 2008, pp. 51–58. [Online]. Available: https://doi.org/10.1109/VAST.2008.4677356
[54] G. Andrienko, N. Andrienko, G. Fuchs, and J. M. Cordero-Garcia, "Clustering trajectories by relevant parts for air traffic analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 34–44, Jan 2018.
[55] A. Buja, J. A. McDonald, J. Michalak, and W. Stuetzle, "Interactive data visualization using focusing and linking," in *Proceedings of the 2Nd Conference on Visualization '91*, ser. VIS '91. Los Alamitos, CA, USA: IEEE Computer Society Press, 1991, pp. 156–163. [Online]. Available: http://dl.acm.org/citation.cfm?id=949607.949633
[56] J. Candil, *Pep's five-second rule, the key to City's success*, 2018, https://en.as.com/en/2018/07/26/football/1532614241_079674.html (accessed February 14, 2019).
[57] S. Chen, J. Li, G. Andrienko, N. Andrienko, Y. Wang, P. H. Nguyen, and C. Turkay, "Supporting story synthesis: Bridging the gap between visual analytics and storytelling," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2019.

Gennady Andrienko is a lead scientist responsible for visual analytics research at Fraunhofer Institute for Intelligent Analysis and Information Systems and part-time professor at City University London. Gennady Andrienko was a paper chair of *IEEE VAST* conference (2015–2016) and associate editor of *IEEE Transactions on Visualization and Computer Graphics* (2012–2016), *Information Visualization* and *International Journal of Cartography*.

Natalia Andrienko is a lead scientist at Fraunhofer Institute for Intelligent Analysis and Information Systems and part-time professor at City University London. Results of her research have been published in two monographs, "*Exploratory Analysis of Spatial and Temporal Data: a Systematic Approach*" (2006) and "*Visual Analytics of Movement*" (2013). Natalia Andrienko is an associate editor of *IEEE Transactions on Visualization and Computer Graphics*.

Gabriel Anzer is the lead data scientist at Sportec Solutions GmbH, a subsidiary of the Deutsche Fußball Liga (DFL). He holds a M.Sc. in Financial Mathematics and Actuarial Sciences. His research focuses on using spatio-temporal positional data to analyze individual and team based performances of soccer players.

Pascal Bauer has a background in mathematics and data-science in applied research at Fraunhofer and worked as a coach/speaker at Fraunhofer Big Data & Artificial Intelligence Alliance. He holds a UEFA A-level coaching license with almost nine years of experience as a head coach and has a passion for soccer. At the DFB Academy, he is responsible for a wide field of data analysis & machine learning applications in soccer, including talent identification, injury prediction models, tactical match analysis based on positional data, and much more.

Georg Fuchs is a senior research scientist heading the Big Data Analytics and Intelligence division at Fraunhofer IAIS. His research work is focussed on visual analytics and Big Data analytics, with a strong emphasis on spatio-temporal and movement data analysis. His further research interests include information visualization in general, Smart Visual Interfaces, and computer graphics. He has co-authored 55+ peer-reviewed research papers and journal articles, including a best short paper award at Smart Graphics 2008 and a VAST challenge award in 2014.

Guido Budziak holds a master's degree in Computer Science. He played professional football for 10 years in the Netherlands, including Dutch national youth teams. He founded Connected.Football, a football technology company aimed at making expert football knowledge accessible to youth academies and amateur football clubs. His research work focuses on communicating football tactics and tactical performance analysis.

Dirk Hecker is vice-director of the Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS and a member of the board of directors of Fraunhofer Academy. His current topics of work include spatial analytics, deep learning and trustworthy AI. He has authored multiple publications on these topics and works as expert and auditor in several boards concerned with artificial intelligence.

Hendrik Weber leads the area of innovation and sports technology for the German Football League (DFL) and is managing director of the subsidiary Sportec Solutions responsible for Sport Tech Operations. He holds a PhD in business administration and is frequent lecturer and author for performance analysis in sports.

Stefan Wrobel is Professor of Computer Science at University of Bonn and Director of the Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS. His work is focused on questions of the digital revolution, in particular intelligent algorithms and systems for the large-scale analysis of data and the influence of Big Data/Smart Data on the use of information in companies and society. He is the author of a large number of publications on data mining and machine learning, is on the Editorial Board of several leading academic journals in his field, and is an elected founding member of the "International Machine Learning Society".