Feature-Driven Visual Analytics of Soccer Data

Halldór Janetzko, *Student Member, IEEE*, Dominik Sacha, Manuel Stein,
Tobias Schreck, *Member, IEEE*, Daniel A. Keim, *Member, IEEE*, and Oliver Deussen

Abstract—Soccer is one the most popular sports today and also very interesting from an scientific point of view. We present a system for analyzing high-frequency position-based soccer data at various levels of detail, allowing to interactively explore and analyze for movement features and game events. Our Visual Analytics method covers single-player, multi-player and event-based analytical views. Depending on the task the most promising features are semi-automatically selected, processed, and visualized. Our aim is to help soccer analysts in finding the most important and interesting events in a match. We present a flexible, modular, and expandable layer-based system allowing in-depth analysis. The integration of Visual Analytics techniques into the analysis process enables the analyst to find interesting events based on classification and allows, by a set of custom views, to communicate the found results. The feedback loop in the Visual Analytics pipeline helps to further improve the classification results. We evaluate our approach by investigating real-world soccer matches and collecting additional expert feedback. Several use cases and findings illustrate the capabilities of our approach.

Index Terms—Visual Analytics, Sport Analytics, Soccer Analysis

1 Introduction

The visual analysis of soccer data is interesting for two reasons: first of all it is scientifically interesting since it is an instance of a geo-spatial analysis problem with complex, interdependent trajectories and events. On the other hand, soccer is a very popular sport and actively played by approximately 270 million people [14]. Soccer plays a huge role in public media coverage and also, poses analytical needs by sport decision makers. Recently, GPS- and video-based tracking technology became available which allows to record spatio temporal data of players at high frequency and accuracy. The arising data is interesting to analyze for two main purposes:

- Scouts are looking for high-performance players, where performance needs to be assessed by many measurable parameters or combinations thereof, in relation to other players and play situations, and over time. For example, these attributes may be the accuracy of shots, the quality of passes, or the willingness to run in the last minutes of a long and exhausting match. Depending on these high-level attributes of a player, the underlying analysis must focus on different sets of basic features. The willingness to run, for instance, depends on the time of the match, the speed of the player, and the current game situation. If the player's team is already three points ahead of the other team, it is not that important to run very fast in the last minutes of the match.
- Coaches are analyzing matches to improve the overall performance. The analysis can be performed either in real-time, in the halftime break, or after the match. Depending on when the analysis is performed the focus is different, which has to be reflected by the analysis process. In defending situations, coaches are interested in dangerous situations, how they occurred and how the team resolved those. For instance, an analysis of the back-four formation can help in assessing the quality of the defense.

Soccer data is a representative of spatio-temporal datasets and therefore already inherently challenging. Compared to standard movement data, the spatial restriction of the movement stands out. Movement data of soccer matches is located on an approximately 105 by 68 meters pitch. As 22 players are moving in a relatively small area, the resulting data is very dense and difficult to visualize by a single visualization. Furthermore, the observed movement patterns are very complex as the movements of each player depend on the movement of all the other players. Nearly every movement action causes a reaction, because of the high interdependencies between all players. Compared e.g., to the flock movement of birds, there are two opposing teams aiming for different targets and trying to hinder the other team. Simple leaderships rules, as e.g., in flock movement theory, are therefore not applicable. To make matters worse, soccer is a very dynamic game as tactics and strategics change over time. Depending on the current game situation a team might, for example, switch their overall game-play from a defensive to an offensive one being directly reflected in the observed movement patterns. Even historically the formations changed from a sweeper up to the late nineties over the 4-4-2 formation to today's most often used 4-2-3-1 formation (see also Section 4.2). Though the outcome of the game, basically who wins and who loses, is not necessary reflecting superiority. Some goals or non-goals are lucky or due to incorrect referee decisions. But what is important when assessing games is why a team won or lost. In order to assess the quality of a team, it is important to take the respective context, e.g., strategy or movement patterns, into account. Switching from offensive to defensive gameplay when losing the ball for instance can be an important clue for coaches.

In this design study, we analyze soccer data with Visual Analytics methods using single-player, multi-player and event-based features. We apply feature analysis techniques to present the most important features to the analyst depending on the respective analysis task. We combine data mining techniques detecting interesting game events with interactive visualizations allowing immediate user feedback to the data mining process. Our focus is to help coaches in investigating interesting and dangerous game situations. There are two points to tackle when trying to support coaches. First of all, interesting game situations have to be identified and presented to the coach and, second, the coach should be able to analyze features of players with respect to these situations. Analyzing a certain game situation can be performed on different levels, such as taking only one single player into account or consider even several players. Our implemented prototype is depicted in Figure 1.

The remainder of this paper is structured as follows. We outline existing and related approaches and system in Section 2. The analysis of a single-player is described in Section 3, followed by a discussion of our multi-player analyses in Section 4. Furthermore, we perform event-based analyses in Section 5. We implemented a modular, layer-based system allowing an easy integration of existing data mining and visualization techniques, described briefly in Section 6 with further

• *All Authors are with the University of Konstanz. Email: forename.lastname@uni-konstanz.de*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication xx xxx 2014; date of current version xx xxx 2014.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

outlook to the Visual Analytics capabilities. Our design is evaluated by use cases and interesting findings together with expert feedback in Section 7. We finally conclude our paper and give an outlook to future work in Section 8.

2 Related Work

We first discuss related work in general visual analysis of sports data in Section 2.1, followed by specific works organized according to the considered analysis perspective in Sections 2.2 and 2.3. Section 2.4 positions our approach within the aforementioned works.

2.1 Visual Analysis of Sport Data in Research Interest

The visual analysis of data related to sports has recently come into focus of research and application [7]. The interest is seen driven by advances in acquisition of high-resolution sports data, and in advances in visualization and analysis of sensor and movement data. Sports analysis is expected to foster many new applications for end users, sports coaches, and sports managers alike [7]. Analytical goals in these applications include overviewing and comparison of player and team performance, prediction and correlation of behavior, and understanding changes over time on the short, medium and long term perspective. Commercial systems are very hard to compare to as the there are high financial interests behind the scenes. We had some discussions with a professional soccer analyst telling us that existing automatic approaches cover more or less only single-player statistics. In-depth team analyses are typically performed by manual inspection.

We just mention two of the most recent sport analysis systems here as examples, before surveying more in the following paragraphs. A recent work on visual analysis of sport data includes [29], where a visual search system for scenes in a Rugby match was introduced. The approach is based on the configuration of team players and their movement during a match, where this data is extracted by means of video analysis. The approach offers a sketch-based query processing for movement patterns extended by Visual Analytics methods. Instead of using movement sketches, we directly look at manually annotated important and dangerous situations and extract similar dangerous ones. We compute semantically meaningful features with respect to soccer and use them for our data mining process. Regarding soccer, by means of a design study, in [33] a tool was developed which combines different perspectives on soccer match data with the aim of creating play reports. The data set used included raw player positions and movement, as well as manually annotated match events like goals, fouls or ball contacts. Thereby, the match data was segmented into meaningful units, which can be visualized in different views. The matches were for instance partitioned by looking at shots and going back in time until the team gained the ball. We extend this work by detecting interesting event and phases semi-automatically by integrating statistical features.

2.2 Movement and Constellation-based Analysis

In general, many approaches for sports analytics consider trajectories extracted for players and teams as a basic abstraction of the data to be analyzed. Consequently, methods of spatio-temporal data analysis are applicable [3, 4]. Important data analysis methods in this area include the segmentation, abstraction, correlation, clustering or classification of trajectories. Today, many applications for trajectory-based data analysis have been identified, including studying of traffic data [43], movements of pedestrians in office spaces [23], or analyzing eye tracking data in context of user studies [31]. Further applications of trajectory-based analysis include understanding of animal movements [38], or analysis of time-dependent measurements in a 2D diagram space [37, 42]. In general, key to successful trajectory-based analysis is finding a meaningful trajectory representation [5].

The trajectory of even a *single* player can already be useful for sports analysis of a game, and it certainly is useful for measuring the performance of a given player. However, often also properties of *groups* of players are relevant. To this end, certain approaches first detect specific constellations among groups of players which may then again, be described by trajectories or other time-dependent group features. Examples for soccer analysis include [25], where player formations are analyzed. Specifically, the spatial constellation between all defenders of one team are analyzed over time, which can reveal tactical maneuvers. In [15], the area on the field where a given player showed a particularly strong influence during the game, was identified. In [40], speed and direction were considered as features in such areas of interest. In [16], distances between player, puck and goal within hockey games were used as features of analysis. Further extensions of the approach of associated areas can be found in [24, 26, 30].

Other works detect specific scenes of interest during a match. [39] detect attacks of a team based on trajectory segmentation in conjunction with specific heuristics on the game progression including change of ball posession and distance to goal. In a further work [18], pass alternatives and their specific contextual difficulty are visualized. Furthermore, paths frequently taken by individual players are considered in that work.

2.3 Analysis Based on Temporal and Statistical Aspects

Besides trajectory-based analysis, also methods from time series and multivariate analysis are applicable to sports data analysis. In general, any relevant measure which is recorded over time (including properties of trajectories) can give rise to time series analysis approaches [20]. Examples include comparison and correlation of measurements among players, or analyzing for cyclic behaviors of measurements [4]. In addition, time-dependent measurements can also be aggregated by descriptive statistics such as mean, variance or other statistical moments of interest.

In [27], it was evaluated which statistic measures correlate with the outcome of a game. The temporal development of geometric statistics, like the convex hull, circumference, or center of a team were analyzed in [12]. Also, statistics were used in [11] to differentiate between players of different positions. A number of commercial and academic software solutions for the analysis of statistical sports data exists. In [35, 36] an interactive statistical tool for coaches is introduced, enabling to analyze and compare players. Furthermore, domain-dependent tools exist e.g., Matchpad [28], CourtVision [17] and SnapShot [34].

Statistical measures can, among other transformations, be defined based on a relational perspective on data: Passing networks can be seen as a rich source for investigating soccer matches even further. Then, statistics can be extracted from a network (or graph-based) representation of the data. E.g., in ball sports, the passing network indicates which player passes the ball to which other players over time. In [32], the performance of players is measured by aggregates of the ball passing network. In [13], additional nodes for "shots to goal" and "shots wide" are added to the passing network description.

2.4 Summary and Positioning of our Work

We distinguish two classes of analysis of sports data. Approaches based on low-level features extract measurements from movement or other sensor data and perform statistical and correlation analyses on the (possibly, pre-processed) data. On the other hand, approaches being oriented toward higher-level representations, such as semantic annotations of data, exist. These can stem, e.g., from manual annotation by human experts or crowds; or by recognition of specific constellations of interest, based on heuristics or Machine Learning approaches.

The work most closely related to ours is [33]. Similarly, we present an interactive system for explorative analysis of soccer data. Our system is flexible in that it incorporates both low-level features (based on trajectory features, see Section 6) and semantic annotations (based on recognition of play configurations, see Section 5) for the analysis. Our system flexibly allows to draw on either of these analysis perspectives, based on the user task. Our semi-automatic selection of features helps to cope with the otherwise difficult problem of feature selection by users. We achieve this by incorporating a user-configurable classifier which allows to detect further events in the movement data, based on a number of example events and input features. Thereby, our system

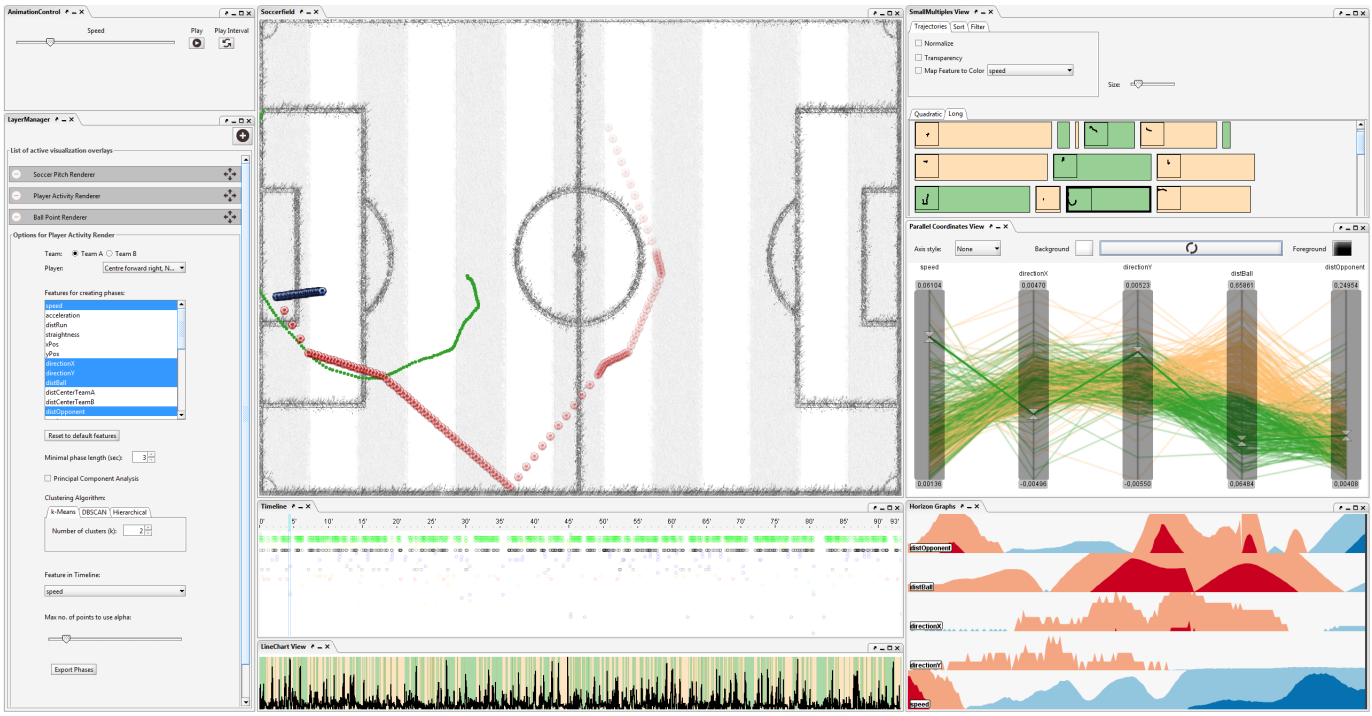

Fig. 1. Interesting phases of a single player can be automatically found by applying the clustering approach presented in Section 3. In this case, we analyze a forward and are interested in the attacks in that the player was involved. Resulting phases can be inspected using the small-multiples view (top-right panel) in combination with the other rendering layers and Horizon Graphs (left and bottom panels).

is not limited to detect a certain number of pre-configured situations, but helps in configuring detectors for many events of interest.

3 Single Player Analysis

The single player analysis investigates the performance and features of one player at a time. We want for example to detect when a player is actively participating during a match. Certain player features, as speed or distance to the ball, will be of use for this kind of analysis. More abstract, we divide the different behavior and motion patterns of a player into different phases. The features being relevant for a single player analysis can be divided into three categories: *Individual Characteristics* (e.g., coordinates and speed), *Game Context* (e.g., distance to ball), and *Events* (e.g., shots, receptions and fouls) features. These features can be seen as numerical time series with changing values over time, with events transformed into a binary time series with singletons.

In order to segment the match into different phases, we apply clustering and hereby detect similar phases. Phases derived from the clustering results should be as homogeneous as possible with respect to the underlying numerical features. The overall analysis process in depicted in Figure 2. We first partition all time series into small, fixed-size intervals and aggregate the values into a numerical feature vector describing the respective time interval. The values are linearly normalized to avoid any biases during the distance calculation. Additionally, we can apply dimension reduction techniques such as PCA to remove noisy dimensions if necessary. The PCA is performed by WEKA [19] automatically reducing the number of dimensions with a threshold of 95 percent of the variance being still explained. The intervals are afterwards clustered resulting in a certain number of clusters (depicted by small letters in Figure 2). In our analyses, we apply k-Means (allowing us to control the number of resulting clusters) and DBSCAN (being a robust clustering technique with respect to noise and outliers). Finally, we merge similarly clustered and adjacent intervals to phases.

We visualize the analysis results using colored trajectories, line charts, parallel coordinates [22] and small multiples [41] all linked via Brushing & Linking. In Figure 1, we present the visual interface

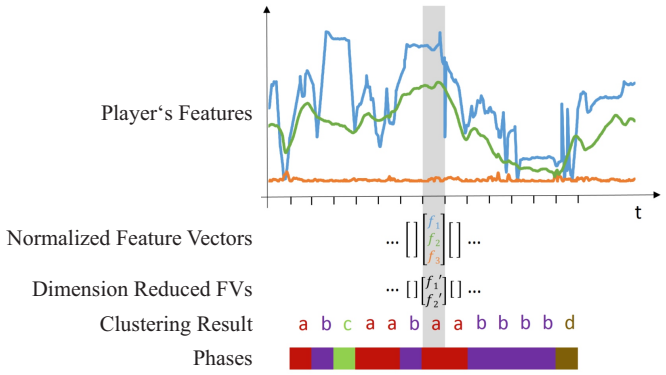

Fig. 2. Feature-based approach to detect similar activity phases of a single player.

showing the results analyzing a forward described in more detail in Section 7.1. It is very crucial in the analysis process to understanding the semantical meaning of found clusters or phases. We therefore integrated several views onto the segmentation results and the human analyst can bring in his expertise.

A first overview is provided by a linechart with a freely selectable feature and background coloring depicting the phases (bottom of Figure 1). Parallel coordinates help to understand the distributions of feature values in the respective clusters. Finally, small multiples offer several interaction possibilities like filtering, sorting (according to a feature, phase similarity or time), or visualization options (e.g., mapping feature values to the trajectory's color).

A typical workflow for this kind of analysis is shown in Figure 3. We start specifying parameters including the clustering parameters and features to regard during the segmentation process. Afterwards, the analyst uses the line chart and parallel coordinates in combination with the small multiples view allowing filtering, highlighting and inspecting

phases. Furthermore, all other implemented visualization layers can be applied to analyze the selected situations of the soccer match in more detail. As our system is interactively reflecting changes to the clustering settings, all steps of the workflow may be revisited several times.

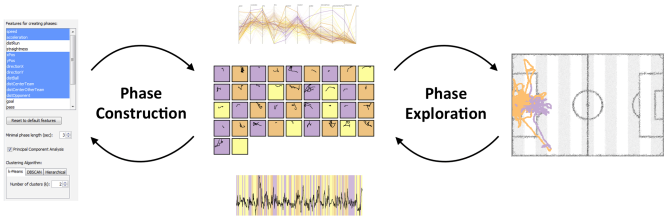

Fig. 3. Schematic workflow for the analysis of a single player.

4 Multi Player Analysis

Regarding more than only one player in the analysis process is very important as soccer is a team sport. This section introduces our methods for the analysis of soccer matches with respect to the movement patterns of multiple players.

4.1 Player Comparison

We enable the analyst to compare several players visually by providing Horizon Graphs [21] for selected players and features. In Figure 4, we show the speed of all field players of one team in the first three minutes of a soccer match. The correlation and the similarity of the speed feature is clearly visible. There are phases with high speed (blue) and also phases with almost no speed (red) showing that the players act as a team. The bottom player can be seen as an outlier to the coherent movement behavior. The bottom Horizon Graph represents a forward who does usually not participate in all defense actions explaining the observed pattern.

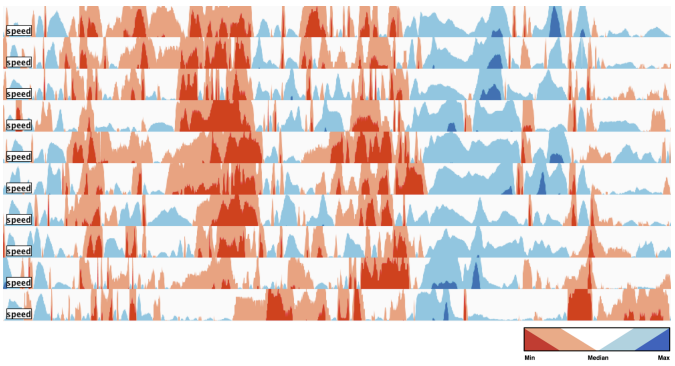

Fig. 4. Speed feature of all field players of one team in the first three minutes of a soccer match.

We furthermore extend the single-player segmentation process described in the previous section towards a multi-player analysis. The combination of phases together with the possibility to inspect selected features visually can reveal interesting patterns. E.g., in Figure 5 we analyze two central defense players. The trajectories are colored by detected phase and the speed features are visualized by Horizon Graphs for the selected time interval (blue rectangle in the timeline). Interestingly, both defense players act very similar, which is reflected in both, the movement features and the phase coloring.

We help the analyst in selecting interesting features to visually inspect by predefined sets of features for the analysis of certain event types. Further details can be found in the use case Section 7.

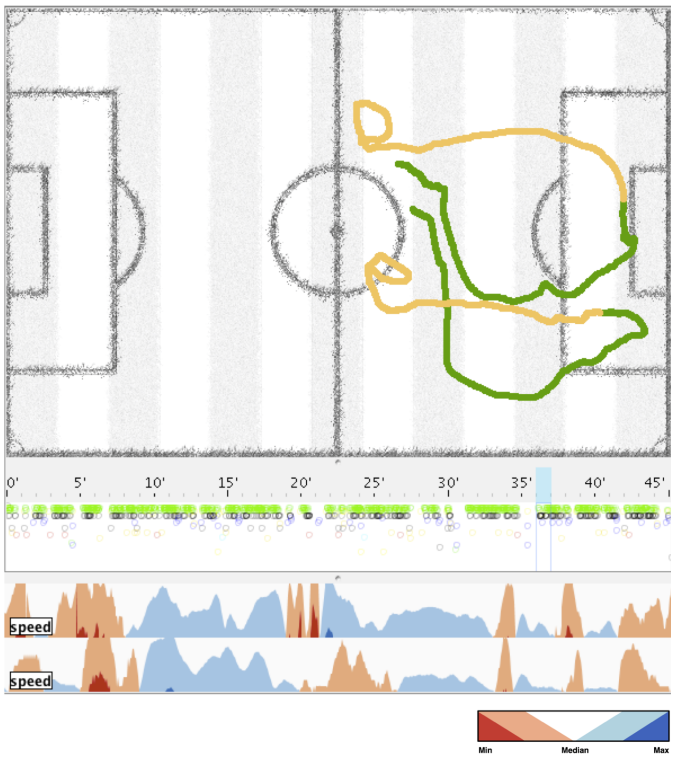

Fig. 5. Activity phases and trajectories for two defense players.

4.2 Constellations and Formations

In addition to low-level and statistical trajectory features, the analysis of spatio temporal team formations in soccer games is very important. This is because formations reveal semantically meaningful patterns and may relate to tactics or strategies of the teams. Formations tell us more about tactics than single player analysis. There exists a variety of formations in modern soccer, like the nowadays very widely known and used 4-2-3-1, the 4-4-2 (a.k.a. “Diamond”), or the 4-3-2-1 (a.k.a. “Christmas Tree”) formation. In this section, we focus on the analysis of the defensive lines and more specifically on the back-four formation. Other defensive structures, such as the defensive triangle, could be also easily automatically assessed. Further descriptions of different formations can be found in [2].

The crucial point when analyzing the back-four formation is to assess the quality with means to the defensive effectiveness. We therefor need a definition for a good and a bad back-four formation. The main task of the back-four formation is to defend their own goal. Nowadays, zonal marking is the widely used defense strategy. Consulting soccer literature and training handbooks, we found some criteria how the back-four formation should react to attacks [10]. There exists an *ideal line* parallel to the ground lines of the pitch, where all back-four players should be. Basically all players should be on the same height, which is also very relevant for the offside trap. Scoring the defensive formation is then simply computing the average distance from the ideal line. However, there exist different kinds of attacks that have to be dealt with differently, resulting in a more complicated assessment. Incoming attacks can be differentiated by the following criteria: As long as the distance between ball and goal is larger than 24 yards, the back-four formation shall use the ideal line described above. If the ball is closer to the goal, we will have to distinguish between an attack from the middle and one from the side. Attacks from the side should be answered by a sickle-like formation. Further details can be seen in this Youtube video [1]. From the computational perspective, we need to check the curvature between outside and central defender of the respective side. Furthermore, the distance between central and outside defender must not be too big, because the outside defender might need

help. The defenders of the side which is not attacked should then build an ideal line reflecting the positions of the other defenders. Defense triangles are the correct reaction to attacks from the middle. The computational assessment is performed by angle computations between the affected defenders. We score the defensive triangle by computing the angles and also include the distances of the involved players. Figure 6 shows two examples for a bad and a good back-four formation.

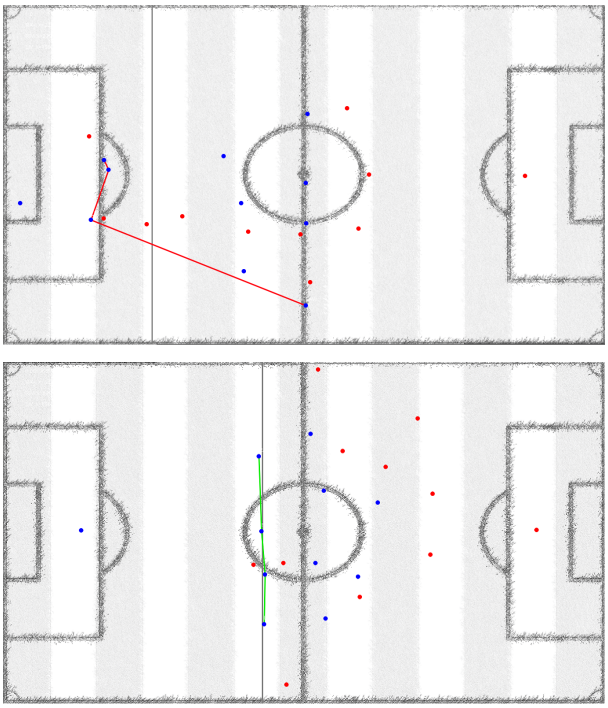

Fig. 6. An example for a bad (upper) and good (lower) evaluated back-four formation based on the ideal line.

5 Event-Based Analysis

Soccer matches are not only continuous movements of players, but there are also incisive events. Besides goals and fouls there are also events like passes or crosses. These events are manually annotated and added to our datasets. We use these events as a basis for event-specific feature pattern exploration. We support two modes of analysis within our system. The first visualizes the development of selected features around user-chosen event types. The second analysis applies a classification technique to support discovery of previously unnoticed candidate events of interest.

5.1 Interactive Feature Analysis

If the analyst wants to analyze a certain kind of events, we visualize features in a time frame around the events with Horizon Graphs. Player specific features are derived from the involved players and additionally game context features as ball specific features are available. We render for each feature and event a single Horizon Graph, and lay them our in a tabular way. A line within each visualization indicates the time point when the event occurs. We included also Brushing & Linking to enable the selection of single events being reflected in all other shown visualizations.

Figure 7 illustrates as an example Horizon Graphs for all crosses occurring in one half of a soccer match. Feature patterns for standard crosses like corners or free kicks are visible as opposite players are typically not near of the executing player. Furthermore, the speed of the executing player is very low at the beginning of the interval, as the player is waiting until he is allowed to perform the free kick. This visualization serves also as a verification for the similar phase analysis presented in the next section.

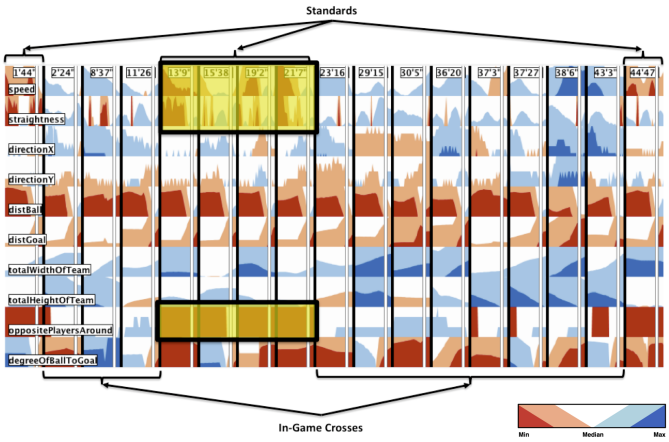

Fig. 7. Features for all crosses occurring in one half of a match. Standard crosses like corners or free kicks can be clearly distinguished from crosses that happen within the match. Before a standard cross speed and straightness are similar lower than other crosses and there are almost no opposite players around the ball.

5.2 Similar Phase Analysis

Using manually annotated data comes with the advantage that human knowledge is added to the data. Though at the same time, there is no guarantee that all events have been detected. We want to make use of the labeled events and learn, how these events can be described from a feature perspective. In this work, we focus on important events as shots on goal, fouls, crosses, and assists. We analyze how specific features, some related to only the involved player and some related to the team, develop right before these events over certain time intervals (2, 5 and 10 seconds). Our goal was to define and train a classifier enabling us to distinguish between intervals where something, like a "shot on goal" event, happened and intervals without this kind of event. Furthermore, we were interested in finding which features are important for this differentiation. We use the resulting classifier to detect similar phases in our data and validate the new found events in our tool as described further in Section 6.4.2. We use KNIME [8] as a state-of-the-art data mining framework and applied all widely used classifiers as Neural Networks, Decision Trees, Probabilistic Models, and Support Vector Machines. Evaluating the classifiers by n-cross-fold validation, we used finally Decision Trees as the classification results were reasonably good and at the same time the model can be easily inspected. Specifically, as the decision tree implies an order of the features which are used for distinguishing the events, we can take these as a reading of the most important features for a given analysis. This information can in turn be used for more in-depth exploration of game situations for these features.

6 System

In this section, we describe the developed components of our system more technically. Our developed *Java* prototype for the analysis of soccer data is depicted in Figure 1. We implemented a layer-based soccer-pitch visualization, with several visualization techniques available (e.g., player position renderer and heat map). The visualization layers can be added interactively and the order and further parameter settings can be controlled by the left control panel. Furthermore, we integrated a timeline visualization and additional panels related to the analyses described in the previous sections. We designed the system in a modular and expandable way in order to enable an easy development of new layers or visualizations being connected to all the other components.

6.1 Features

Most of our visualizations and analyses rely on different kinds of features (see previous sections). These features are extracted, derived,

and finally delivered to all other components. Player-specific features are computed and available for each player. Furthermore, team- and ball-related features are calculated as well. In Table 1 we list all features that are already implemented and available in our system. The extension of this list is an ongoing process triggered by new use cases and analysis needs emerging by tool usage and expert interviews.

Single Player Analysis	
Speed	Acceleration
Position	Direction of movement
Distance covered	Straightness
Distance to next opposite	Distance to ball
Distance to own team center	Distance to opposite team center
Multi Player Analysis	
Width of team shape	Height of team shape
Opposite players around player	Back-four formation
Event Based Analysis	
Shots on goal	Passes
Fouls	Off-site
Cards	Reception
Goal	Clearance
Running with ball	Assists
Game Specific Analysis	
Ball-goal distance	Ball position
Angle of ball to goal	

Table 1. Features implemented in our system.

6.2 Visualization Components

Our prototype offers several panels where visualization can be plugged into and also provides synchronization functionality between the components. The analyst can control the currently visualized time windows by using the timeline component showing the selected time interval and event occurrences. We furthermore developed a layer manager where several layers can be registered and rendered on a soccer pitch area simultaneously. For each layer it is possible to integrate an option panel handling the layer's configuration (e.g., clustering parameters). Finally, we offer a feature export component allowing to export features based on selected players, events, or time intervals. We make use of the export capabilities integrating external software components, described in more detail in Section 6.4.2.

6.3 Visualizations

Depending on the analysis task, we provide different visualizations. Most of the visualizations are realized as layers that can be drawn on a soccer pitch. In order to get details of a soccer scene, we offer a player and ball renderer visualizing a selected scene. For larger time windows, we provide a heat map that can be computed for every spatio-temporal object (e.g., player, ball, event position). Selected features may be analyzed through line charts or horizon graphs. We provide specific views being useful in combination with each other. For example, the single player analysis view consists of the colored trajectory on the soccer pitch, the small multiples view, a colored line chart and the parallel coordinates plot. Another example is the back-four formation layer that renders formation dependent lines and colors on top other layers and also adds information to the timeline component.

6.4 Analysis Facilities

This section briefly describes how our system integrates analysis functionality.

6.4.1 WEKA Clustering Integration

We integrated the WEKA-library [19] in order to support state-of-the-art analysis techniques. WEKA takes care of the cluster analysis described in section 3. We integrated the clustering components K-Means, DBSCAN, and hierarchical clustering for the single player analysis. The classification capabilities of WEKA are used in KNIME for the data mining part of our Visual Analytics pipeline.

6.4.2 Visual Analytics Integration and Machine Learning

As stated above, we are interested in gaining knowledge from investigating features of annotated events. We want to study which features and values are significant for different kinds of events. Furthermore, we want to use this knowledge for finding new events that were not annotated but can fulfill the found criteria. We set up a KNIME workflow and integrated it into the analysis process depicted in Figure 8. We export all extracted and computed features into the KNIME workflow and partition the time series data into fixed-length intervals. Intervals including an event are marked as class A, while all others are marked as class B. After preprocessing, we train all available KNIME and WEKA classifiers with a 33% data sample and evaluate with the remaining data. We take the best five classifiers (LMT, LibSVM, Logistic Base, FT, and Decision Stump) according to their accuracy measured by their confusion matrix. The accuracy of the best classifiers ranges from 72 to 90 percent. We consider for our decision also the amount of false positives, which should be reasonable. False positives indicate new potential interesting intervals not yet annotated in our data. The classification results are then imported back into our prototype allowing the analyst to investigate time points labeled as class A. Furthermore, we integrate a feedback loop enabling the analyst to confirm found, previously untagged events and use them as additional training data for the classifier. This feedback loop may be repeated as often as the analyst wishes to.

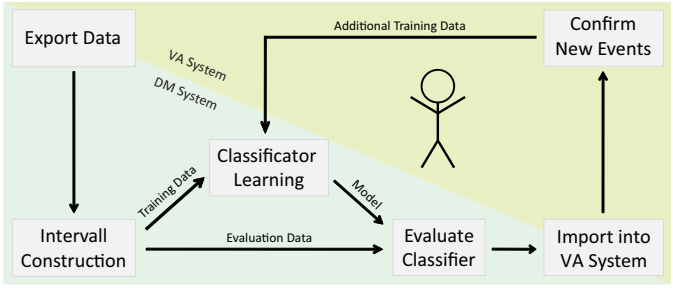

Fig. 8. Analysis process for the detection of similar events and feedback loop to the classifier.

6.5 Interaction and Animation

Every developed component offers several interaction possibilities allowing the analyst to steer his analysis. Linking & Brushing is supported among all visualizations enabling multi-view data exploration. Besides mouse interactions and parameter setting controls, we provide common keyboard shortcuts in order to facilitate power user operations (e.g., animation control). Additionally, animation of selected soccer scenes turned out to be useful in order to verify results or to understand and investigate longer phases avoiding overplotting issues.

7 Use Cases and Evaluation

In this section, we demonstrate in applicability of our prototype in different analytical use cases. We present several analyses and findings in which a domain expert could gain further knowledge about his team. We analyze a single player, detect similar situations in the soccer game, and investigate team formations as the back-four formation.

The data analyzed in our use cases is provided by prozone/mastercoach. The data set is not publicly available and was anonymized as it was a professional game. For each of the 22 players timestamped, two-dimensional position data are available with a

temporal resolution of 100 milliseconds. Furthermore, the data includes manually annotated events containing information about position, time, and event-specific information as the involved player. These events are less frequent and lack in accuracy as they are manually tagged.

The use-cases were designed to show how our prototype can help coaches in analyzing the offensive and defensive qualities of their team. We will first analyze a single player and focus on his active phases. Afterwards, we investigate the offensive gameplay and the defensive back-four formation. These use-cases reflect some of the most important training aspects for a successful training, basically the attacking and defending skills. The last paragraph will cover some expert feedback we received when showing the tool to a subject matter expert.

7.1 Analysis of a Forward

Grouping and clustering interesting phases of a single player can be performed automatically by applying the clustering approach presented in section 3. In this use case, we analyze a forward and are interested in the attacks where he was involved. Therefore, we select the features *Speed, Direction of Movement (x and y-dimension), Distance to nearest opposite Player, Distance to Ball* and apply a k-Means clustering with two desired clusters in order to divide interesting from non-interesting phases. The resulting phases can be inspected using the small-multiples view in combination with the other rendering layers and the Horizon Graphs. In Figure 1, we show the analysis results of the forward's attacks.

If we want to investigate the two clusters, we will use the parallel coordinates plot showing the feature values for all phases. Though the labels of the parallel coordinates plot show the original data space, we used normalization before applying clustering. Obviously, green phases are defined by large distances to the ball. These green phases are uninteresting phases which we can ignore in our analysis. The uninteresting green phases can be hidden from the small multiples view to focus only on the interesting phases. As a next step, we take a closer look at the interesting (orange) phases where the player was very active and near to the ball. We sort the small-multiples according to his x-position in order to see the phases where the player was closest to the opposite goal first. Selecting one small multiple will make all other components showing the selected phase. Figure 1 shows the third phase the system found, in which the forward receives the ball after he started to sprint and scores his first goal. The player is rendered by an orange trajectory and the phase can be animated as well. As a next step, the coach could inspect the other phases or arrange the small-multiples by similarity in order to find similar patterns. Another option would be to explore the player's features using horizon graphs as described previously.

7.2 Shot-Event Feature Pattern Analysis

As described above, we try to gain knowledge from the manually annotated events. We focus in this section on the most important event of a soccer event, namely the shot on goal. We applied and investigated the Decision Trees mentioned above in Section 6.4.2 to classify the events. We found that the most relevant features are *x-Position* (near to left or right goal), *Total Width of Team* (in dangerous situations the team width in x-dimension is greater that usual), and *Opposite Players around* (more opposite players are around trying to prevent shots). Furthermore, the *speed* feature turned out to be useful for all kind of events. Crosses and shots are events easily detectable by classifiers, whereas fouls and assists are difficult to detect.

7.2.1 Annotated Shot Events

We visually inspect all pre-annotated shots on goal by plotting them next to each other using Horizon Graphs for the most relevant features. In Figure 9 we investigate all relevant features of the first half of the game by Horizon Graphs in combination with the soccer pitch, players, and ball rendered for the second shot event. Similar to crosses (analyzed in Figure 7), we can detect one direct free kick (6th event) as there are no opposite players around and there is no speed before

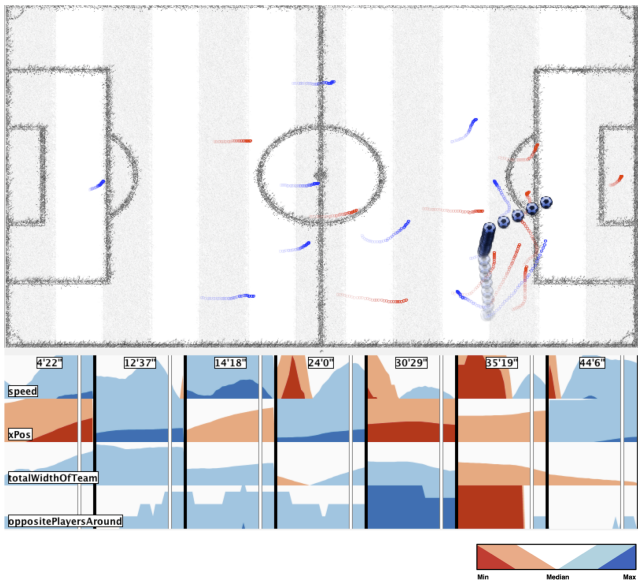

Fig. 9. Horizon Graphs for the relevant features of all shots on goal events in first half of the game. The second shot event is shown on the soccer pitch above. The time point of the event is represented by a vertical white line.

the shot (the player is waiting until he is allowed to perform the free kick). During all other shot events there are many of opposite players around and the x-position is near to the relevant goal. In most of the events the team width is higher than usual indicating that there is a fast movement of the offensive players towards the goal.

7.2.2 Shot Events Found by Classification

The analyst may also be interested in similar, dangerous, and interesting situations not yet being marked in the data. We therefore exported the transformed soccer data into the KNIME workflow as described in Section 6.4.2. We trained and evaluated our classifiers and imported the results back into our prototype. Several new shot on goal events could be detected by our classifier but were not yet marked in the original game data. Figure 10 illustrates the classification results. Where green bars depict correctly found events, red represent not found events, and yellow bars stand for potentially interesting events.

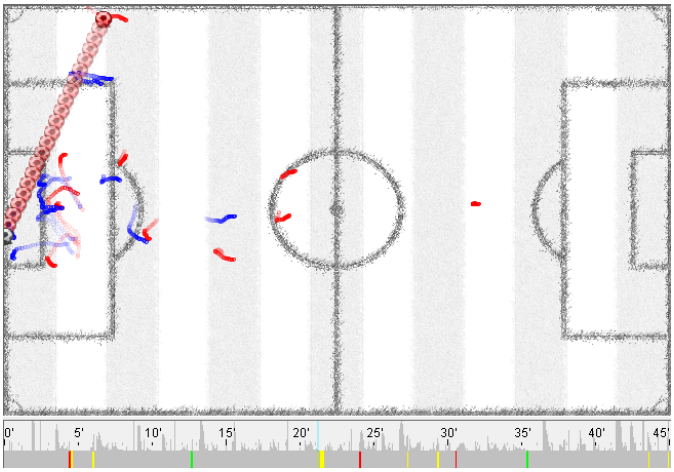

Fig. 10. Analysis of detected new shot on goal events. Green colored bars indicate correct classified events, red represent not found events, and yellow bars show events found by the classifier but not tagged in the original input data.

The analyst is able to validate new found shot on goal events and mark correct found as new shot on goal events. Following the Visual Analytics pipeline it is possible to add the new events to our KNIME workflow and to update the classifiers. It is therefore feasible to extend, update and improve the classifiers to gain more insights.

For our example, we inspected all found shots on goal events not annotated before and marked the correct ones. We retrained our classifiers with the additional training data and imported the classification results into our tool. By this single iteration we discovered eight new events of which five were relevant. An excerpt of the newly found events can be seen in Figure 11.

It seems that the extension of our classifiers with additional interesting events helped to move away from pure shot on goal events to overall dangerous events. The upper image in Figure 11 shows a new not yet marked shot on goal event, whereas the middle and lower image show dangerous situations. In the bottom row for example a striker tried to enter the penalty area, but was stopped in the very last moment. We see the discovery of overall dangerous situations as a prove that the Visual Analytics pipeline helps in improving the classification results.

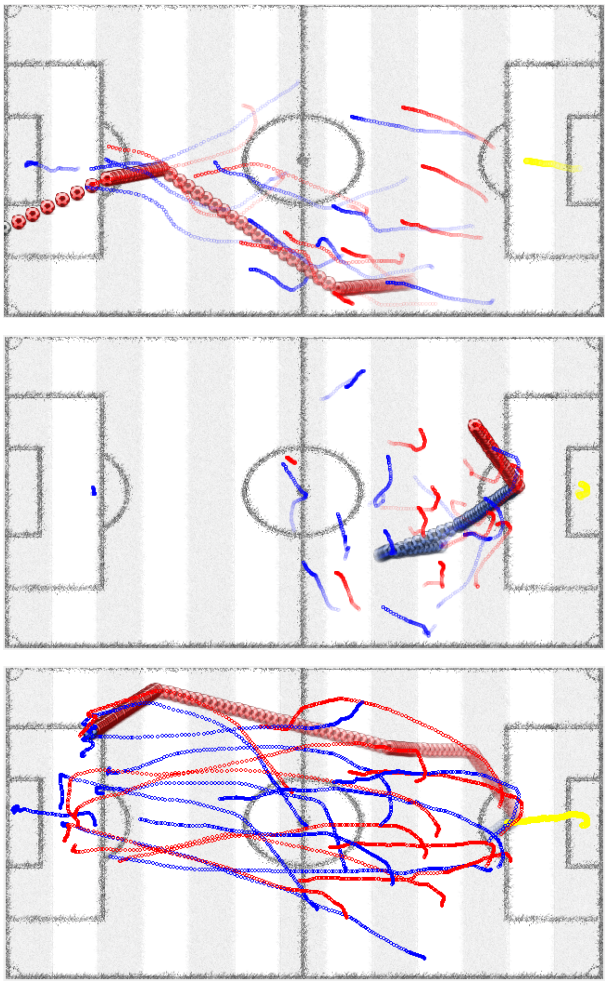

Fig. 11. New events found after adding confirmed events to the classifier's training data. The classifier returns not only shots on goal (top) anymore but also semantically dangerous situations (middle and bottom).

7.3 Back-Four Formation

In this use case, we want to evaluate how the back-four formation performed right before a goal was scored. We investigate a short period before the goal of use case 1 happens (Section 7.1).

The key scene of the failure is shown in upper Figure 12. Our previously described assessment of the back-four formation detects that there seems to be something wrong with the back-four formation resulting in a red coloring. Investigating this time frame we can see why: the back-four formation seems to have problems with their coordination. The nearest midfield player to the right-back is not fulfilling any correct defensive tasks. Unfortunately, the central right defender decides wrong and moves out to the sideline in order to cover another opposite player. Instead, he should have stayed near his usual position to cover the central areas in front of the goal. Although, a free opposite player at the sideline is not good, it is much more dangerous to have large distances between defenders and uncovered opposite players near the middle. A simple pass through the resulting free space leads to a situation with again too much free space for the opposite striker. Three own defending players are consequently outplayed and not involved in the defense anymore.

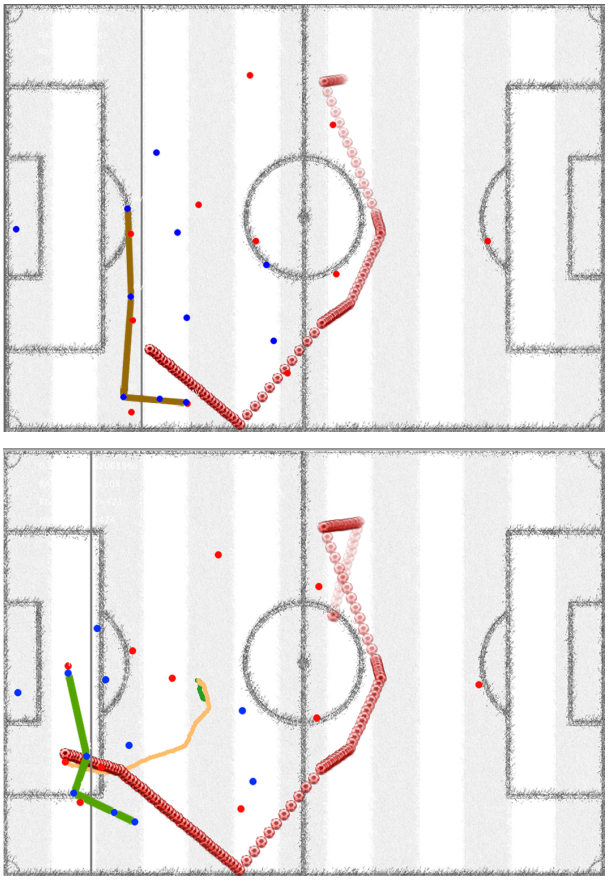

Fig. 12. Back-four formations immediately before the goal occurs. The connecting line is colored from red to green representing the computed quality of the back-four formation.

In lower Figure 12, the back-four formation has improved their positions and tries hard to recover from their previous mistake. As the central right defender moved back, the overall formation is better than before resulting in greenish coloring. Though the mistake was too severe to recover from and the opposite players is already on his way to score a goal.

The coach of this team can learn from the analysis and teaches his central-back players to stay near the center area and avoid any free spaces in the center. Furthermore, the coach should improve the collaboration and coordination of defensive midfield players and the back-four formation as well. If the midfield player at the sideline had covered his opposite number, the central right defender would not have needed to assist at all.

7.4 Expert Feedback

We had contact to a soccer expert, who is involved into playing soccer since 23 years and into coaching since nine years. Currently, he is working for FC Bayern München being an international successful German soccer club. He is certain about the benefits a semi-automatic tool has and that such tools can be implemented in professional soccer sports. The semi-automatic analysis will help coaches in cases where there is not enough time for a manual analysis and it allows analyzing more games in the same amount of time compared to a pure manual analysis. Current developments in soccer show that coaches want to decide less by intuition but more by hard facts and figures. We showed the capabilities of our current prototype and the use-cases to the soccer expert and asked for his feedback and opinions. The overall feedback was quite good, but he came immediately with suggestions for improvements that will be included in future versions of this tool.

We were especially interested in the effectiveness of the implemented Horizon Graphs. Horizon Graphs were not intuitive to the soccer expert and were explained to him by showing the visual process of transforming a line chart into a Horizon Graph. After the explanations, he was not only able to read the visualizations but was also convinced that this visualization technique supports him better than traditional line charts. He was amazed by the possibility to see the team's coherence for certain features as speed or acceleration supported by the color changes around the quartiles. In his opinion, Horizon Graphs are most beneficial when comparing the same attribute across several players. Comparing several players reflects the spirit of soccer being a team sport.

Detecting dangerous situations and potential shots semi-automatically was regarded positively, especially with respect to fast analysis tasks. During half-time breaks, the detection of potentially dangerous situations can be very helpful. He mentioned that it would be also interesting to get hints about, why a certain attack did not succeed and lead to a goal.

With respect to future improvements and capabilities of our tool he sees the following potential: Coaches could validate their – maybe intuitive or experience-based – hypotheses in our tool, by looking for a certain kind of situation specified by the coach. Thereafter, the system should automatically derive the corresponding features and detect similar situations and display them.

8 Conclusion

We presented a Visual Analytics approach to investigate soccer data and gain new insights. Based on the analysis of single-player, multi-player, and event-based we were able to easily detect standard situations as crosses for example. The integration of state-of-the-art data mining techniques helps to find and understand interesting events. Additionally, even not previously annotated interesting events could be found by Visual Analytics methods. Currently, our prototype is set up as an expert tool. We followed a data-driven tool design, namely, we aimed to combine visual analytics techniques deemed useful to answer analytical questions in context of high-resolution soccer sensor data. These techniques include interactive and automatic data filtering, visual representation of trajectories on a soccer field, and compact time series visualization using horizon graphs. As we expect the types of required movement features to vary between different analytical questions, we decided to compute a large number of features from which the expert can chose. In addition, inspired by similar recent work [9], we incorporated an interactive classifier which can help to discover events of potential interest, based on example event annotation, relying on a broad basis of features. Given our system is set up as an expert system, we recommend it being used in Pair Analytic scenarios [6].

Our future work includes to provide our prototype to coaches and to support the most often used analyses by predefined configuration settings and the definition of task-driven views. The analysis of soccer features is at an early stage, but this pre-study showed already some information available in the data. We want to extend our approach to a semi-automatic detection of mistakes of a team to help the coach in finding critical situations. Furthermore, we want to integrate video material into our system whenever available and implement more assessment criteria for formations. Furthermore, we want to integrate a better visualization for the movement of players and soccer-specific artifacts as free spaces or running paths. Especially when visualizing longer time windows, a more abstract visualization technique adapted to soccer is necessary. Additionally, we will integrate the expert's feedback in order to support the coach in validating hypotheses.

Acknowledgments

The authors wish to thank David Perlich for his implementations and Moritz Stefaner for the fruitful discussions. Further thanks for fruitful discussions on the subject are due to Michael Behrisch. We furthermore wish to thank Andreas Riederer for his valuable feedback and discussions. The soccer data used in this publication was generously provided by prozone/mastercoach.

References

[1] Behavior of a back-four formation (German). http://www.youtube.com/watch?v=P_pFKGDEgUc, 2014.

[2] Formation (association football). http://en.wikipedia.org/wiki/Formation_%28association_football%29, 2014.

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

[4] N. V. Andrienko and G. L. Andrienko. *Exploratory analysis of spatial and temporal data - a systematic approach*. Springer, 2006.

[5] N. V. Andrienko, G. L. Andrienko, L. Barrett, M. Dostie, and S. P. Henzi. Space transformation for understanding group movement. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2169–2178, 2013.

[6] R. Arias-Hernández, L. T. Kaastra, T. M. Green, and B. D. Fisher. Pair analytics: Capturing reasoning processes in collaborative visual analytics. In *HICSS*, pages 1–10, 2011.

[7] R. Basole, E. Clarkson, A. Cox, C. Healey, J. Stasko, and C. S. (Organizers). First IEEE visworkshop on sports data visualization, Oct. 14, 2013.

[8] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, 2007.

[9] H. Bosch, D. Thom, F. Heimerl, E. Puttmann, S. Koch, R. Krüger, M. Wörner, and T. Ertl. Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2022–2031, 2013.

[10] S. Chapman, E. Derse, and J. Hansen, editors. *LA84 Foundation Soccer Coaching Manual*. LA84 Foundation, 2012.

[11] V. Di Salvo, R. Baron, H. Tschan, F. Calderon Montero, N. Bachl, and F. Pigozzi. Performance characteristics according to playing position in elite soccer. *International journal of sports medicine*, 28(3):222, 2007.

[12] R. Duarte, D. Araújo, H. Folgado, P. Esteves, P. Marques, and K. Davids. Capturing complex, non-linear team behaviours during competitive football performance. *Journal of Systems Science and Complexity*, 26(1):62–72, 2013.

[13] J. Duch, J. S. Waitzman, and L. A. N. Amaral. Quantifying the performance of individual players in a team activity. *PloS one*, 5(6):e10937, 2010.

[14] Fédération Internationale de Football Association. Big Count, Mar. 2014. http://www.fifa.com/worldfootball/bigcount/index.html.

[15] S. Fonseca, J. Milho, B. Travassos, D. Araújo, and A. Lopes. Measuring spatial interaction behavior in team sports using superimposed voronoi diagrams. *International Journal of Performance Analysis in Sport*, 13(1):179–189, 2013.

[16] A. Fujimura and K. Sugihara. Geometric analysis and quantitative evaluation of sport teamwork. *Systems and Computers in Japan*, 36(6):49–58, 2005.

[17] K. Goldsberry. Courtvision: New visual and spatial analytics for the nba. MIT Sloan Sports Analytics Conference, 2012.

[18] J. Gudmundsson and T. Wolle. Football analysis using spatio-temporal tools. In *Proceedings of the 20th International Conference on Advances in Geographic nformation Systems*, pages 566–569. ACM, 2012.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Nov. 2009.

[20] J. Hamilton. *Time series analysis*, volume 2. Cambridge Univ Press, 1994.

[21] J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1303–1312. ACM, 2009.

[22] A. Inselberg and B. Dimsdale. Parallel coordinates. In *Human-Machine Interactive Systems*, pages 199–233. Springer, 1991.

[23] Y. Ivanov, C. Wren, A. Sorokin, and I. Kaur. Visualizing the history of living spaces. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1153–1160, 2007.

[24] C.-H. Kang, J.-R. Hwang, and K.-J. Li. Trajectory analysis for soccer players. In *Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on*, pages 377–381. IEEE, 2006.

[25] H.-C. Kim, O. Kwon, and K.-J. Li. Spatial and spatiotemporal analysis of soccer. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 385–388. ACM, 2011.

[26] S. Kim. Voronoi analysis of a soccer game. *Nonlinear Analysis: Modelling and Control*, 9(3):233–240, 2004.

[27] C. Lago-Peñas, J. Lago-Ballesteros, A. Dellal, and M. Gómez. Game-related statistics that discriminated winning, drawing and losing teams from the spanish soccer league. *Journal of sports science & medicine*, 9(2):288, 2010.

[28] P. A. Legg, D. H. Chung, M. L. Parry, M. W. Jones, R. Long, I. W. Griffiths, and M. Chen. Matchpad: Interactive glyph-based visualization for real-time sports performance analysis. In *Computer Graphics Forum*, volume 31, pages 1255–1264. Wiley Online Library, 2012.

[29] P. A. Legg, D. H. S. Chung, M. L. Parry, R. Bown, M. W. Jones, I. W. Griffiths, and M. Chen. Transformation of an uncertain video search pipeline to a sketch-based visual analytics loop. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2109–2118, 2013.

[30] R. Nakanishi, J. Maeno, K. Murakami, and T. Naruse. An approximate computation of the dominant region diagram for the real-time analysis of group behaviors. In *RoboCup 2009: Robot Soccer World Cup XIII*, pages 228–239. Springer, 2010.

[31] K. Ooms, G. L. Andrienko, N. V. Andrienko, P. D. Maeyer, and V. Fack. Analysing the spatial dimension of eye movement data using a visual analytic approach. *Expert Syst. Appl.*, 39(1):1324–1332, 2012.

[32] J. L. Peña and H. Touchette. A network theory analysis of football strategies. *arXiv preprint arXiv:1206.6904*, 2012.

[33] C. Perin, R. Vuillemot, J.-D. Fekete, et al. Soccerstories: A kick-off for visual soccer analysis. *IEEE transactions on visualization and computer graphics*, 2013.

[34] H. Pileggi, C. D. Stolper, J. M. Boyle, and J. T. Stasko. Snapshot: Visualization to propel ice hockey analytics. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2819–2828, 2012.

[35] A. Rusu, D. Stoica, and E. Burns. Analyzing soccer goalkeeper performance using a metaphor-based visualization. In *Information Visualisation (IV), 2011 15th International Conference on*, pages 194–199. IEEE, 2011.

[36] A. Rusu, D. Stoica, E. Burns, B. Hample, K. McGarry, and R. Russell. Dynamic visualizations for soccer statistical analysis. In *Information Visualisation (IV), 2010 14th International Conference*, pages 207–212. IEEE, 2010.

[37] T. Schreck, J. Bernard, T. Tekuov, and J. Kohlhammer. Visual cluster analysis of trajectory data with interactive Kohonen maps. *Palgrave Macmillan Information Visualization*, 8:14–29, 2009.

[38] D. Spretke, H. Janetzko, F. Mansmann, P. Bak, B. Kranstauber, S. Davidson, and M. Mueller. Exploration through Enrichment: A Visual Analytics Approach for Animal Movement. In D. Agrawal, I. Cruz, C. S. Jensen, E. Ofek, and E. Tanin, editors, *Proceedings of the 19th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 421–424, New York, NY, USA, 2011. ACM.

[39] F. Staals and J. Gudmundsson. Detecting attack patterns in football trajectory data & home region computation.

[40] T. Taki and J.-i. Hasegawa. Visualization of dominant region in team games and its application to teamwork analysis. In *Computer Graphics International, 2000. Proceedings*, pages 227–235. IEEE, 2000.

[41] E. R. Tufte and P. Graves-Morris. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 1983.

[42] T. von Landesberger, S. Bremm, T. Schreck, and D. Fellner. Feature-based Automatic Identification of Interesting Data Segments in Group Movement Data. *Information Visualization Journal*, 2013.

[43] Z. Wang, M. Lu, X. Yuan, J. Zhang, and H. V. D. Wetering. Visual traffic jam analysis based on trajectory data. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2159–2168, 2013.