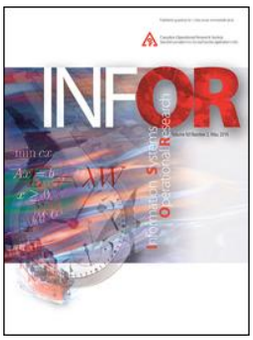

INFOR: Information Systems and Operational Research

ISSN: 0315-5986 (Print) 1916-0615 (Online) Journal homepage: http://www.tandfonline.com/loi/tinf20

Interactive exploratory soccer data analytics

Emrullah Delibas, Ali Uzun, Mehmet Fatih Inan, Onur Guzey & Ali Cakmak

To cite this article: Emrullah Delibas, Ali Uzun, Mehmet Fatih Inan, Onur Guzey & Ali Cakmak (2019): Interactive exploratory soccer data analytics, INFOR: Information Systems and Operational Research

To link to this article: https://doi.org/10.1080/03155986.2018.1533204

Published online: 07 Jan 2019.

Submit your article to this journal

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=tinf20

INFOR: INFORMATION SYSTEMS AND OPERATIONAL RESEARCH
https://doi.org/10.1080/03155986.2018.1533204

Taylor & Francis
Taylor & Francis Group

Check for updates

Interactive exploratory soccer data analytics

Emrullah Delibas, Ali Uzun, Mehmet Fatih Inan, Onur Guzey and Ali Cakmak

Department of Computer Science, Istanbul Sehir University, Istanbul, Turkey

ABSTRACT

Spatiotemporal soccer data enables in-depth analysis of a soccer game. However, the amount and the nature of the data makes it challenging for analysts to easily uncover insights from the data. In this article, we introduce an interactive visualization tool that uses novel data mining and machine learning methods to enable coaches and analysts to work on large amounts of data by moving most of the complicated models to the backend and presenting interactive visualization that can be manipulated in real-time. A unique interactive replay and modification feature enables creation of what-if scenarios on existing game data to explore alternative situations, such as a defensive player taking a different position or an offensive player choosing another pass, while making the experience seamless to the users.

• Information systems → Information system applications • Applied computing → Computers in other domains.

ARTICLE HISTORY
Received 8 November 2016
Accepted 30 September 2018

KEYWORDS
Exploratory data analysis; visual analytics; classification; soccer analytics; Voronoi diagrams

1. Introduction

In recent years, computerized tracking systems (Castellano et al. 2014) that can collect spatiotemporal data from soccer games have become common place in major leagues and international competitions. These systems are capable of sampling the locations of players, referees and the ball multiple times per second to provide location, acceleration and direction data throughout the game. This large amount of data makes performing large-scale data analytics on the game of soccer possible.

Although, analytics had a large impact in other sports, such as baseball (Baumer and Zimbalist 2014) and basketball (Goldsberry and Weiss 2013), their influence on soccer has been relatively limited. A portion of this limited effect can be appropriated to the low scoring nature of soccer. Baseball and basketball provide clear hit or miss opportunities in the form of thrown balls and shots taken. In comparison, soccer games have few goals scored, which require more complicated analysis to be involved. In other words, soccer data has fewer low hanging fruits for data analytics.

CONTACT Ali Cakmak alicakmak@sehir.edu.tr Department of Computer Science, Istanbul Sehir University, Istanbul, Turkey.
A quick demo video is available at: http://idea.sehir.edu.tr/sport_analytics_demo.mp4
Source code is available at: https://bitbucket.org/onurguzey/futbol-data-analysis/src
© 2018 Canadian Operational Research Society (CORS)

An interactive soccer data analytics tool can be utilized by coaches, scouts or TV analysts. Coaches can analyze past game performances of individual players or the entire team, and anticipate competitor's tactics in future games. Scouts can gather information about potential acquisition targets and see how they would fit in their teams. TV analysts can perform real-time analysis with visualizations that convey relevant information, such as interactive heatmaps, that would not be readily available otherwise.

In this article, we propose an interactive visualization tool that can be used to perform exploratory analysis on soccer data. The visualizations are enriched by utilizing various data mining and machine learning algorithms, such as support vector machine (SVM) (Scholkopf and Smola 2001), in the backend. The tools provide advanced exploration options such as pass success, ball ownership and optimal shooting point prediction. Our models work on spatiotemporal data collected in games. In addition, these models can be augmented or personalized with the addition of expert survey data collected from experienced soccer coaches.

A crucial component of the tool is its ability to interactively create what-if scenarios on top of the existing game data. Using this interactive game replay and modification feature, users can jump into any point in the game and change locations of players to create new scenarios. The tool re-builds its models, performs all necessary calculations and displays updated visualizations in real-time. This enables in-depth exploration and analysis of game data. A coach or TV analyst, can go over the game and explore alternative scenarios or analyze the importance of a single player through the user-interface.

The rest of the article is organized as below. Section 2 summarizes related work. Section 3 introduces interactive data mining and machine learning components of the tool. Visual elements, such as the interactive heatmaps, are explained in Section 4. Section 5 presents the high-level analysis features and Section 7 concludes the article.

2. Related work

A number of studies focused on evaluating the performance of players and teams so far. Gudmundsson and Wolle (2014) presents a collection of tools that are developed specifically for analyzing the performance of individual players and teams. The tools mainly focus on passing, pass sequences, and correlation-based analysis. Perin et al. (2013) proposes a visualization interface, named SoccerStories, to support analytics in exploring soccer data and communicating interesting insights. This interface is designed to support the current practice of soccer analysts and to enrich it, both in the analysis and communication stages. It also provides different interfaces for high-level and detailed view of game phases, and their aggregation into a series of connected visualizations. Each visualization focuses on important game moments such as a series of passes or a goal attempt.

To a coach, the summary statistics (e.g. the percentage of ball possession) lack sufficient details to act on. Reading the spreadsheet is time-consuming and making decisions based on the spreadsheet in real-time can be challenging. Legg et al. (2012) presents a visualization solution to this problem in real-time sports performance

analysis. Cox and Stasko (2006) also introduces a tool, called SportVis, which uses visualization to help people discover meaning in the massive amount of statistics generated during sporting events. Kang et al. (2006) proposes a model to quantitatively express the performance of soccer players. The study focuses on analyzing the trajectories of players and the ball, since they all interact with each other and produce a trajectory that has certain properties. The proposed model exploits the relationships between trajectories of 22 players and the ball and allows evaluation of the performance of several players in a quantitative way.

Taki and Hasegawa (2000) proposes the concept of 'dominant region' for each player. Inspired by Voronoi cells, the dominant region of a player is defined as the set of points that s/he can reach before other players on the field/court. The authors define the properties of dominant region and provide an approximation algorithm for the computation of dominant regions. As an application of the dominant region, a motion analysis system of team ball games was developed. In this system, the dominant region is used for quantitative evaluation of basic teamwork. Taki and Hasegawa (1998) proposes the same basic feature for quantitative measurement and evaluation of group behaviour, and introduces a motion analysis system of soccer games as an application of it. The purpose of the system is to evaluate the teamwork quantitatively based on movement of all the players in the game. Nakanishi et al. (2009) introduces a method for real-time calculation of the 'dominant region' by stating that it takes an important role to analyze the current situation in the game, and it is useful for evaluating the suitability of the current strategy. Another advantage of its real-time calculation is that it makes prediction of pass success or failure possible.

To the best of our knowledge, ProZone Sports is one of the leading commercial tools exist around. Below, we present a detailed comparison with it.

- Both tools offer general statistics in terms of passing. Our tool enables the user to examine the 'quality' of given passes as well by providing an embedded pass effectiveness algorithm (Cakmak et al. 2018), whereas ProZone only offers additional statistics such as one-to-one passing between players, etc.
- The functionality of trace exists in both tools, but ProZone, additionally, enables users to get all passes that leads to a goal for example.
- Both tools offer player-based statistics such as total running distance, total coverage, etc. ProZone lets the user make comparisons between different games and provides performance-based visualizations on top of that. On the other hand, our tool enables the user to generate even more sophisticated heat-maps, but on a single game only.
- Both tools provide displaying and statistics on game events but our tool additionally enables the user to examine what-if scenarios as well by letting the user manipulate the positioning of players.
- ProZone mainly focuses on 'what happened' in the game only and provides a mass of analytics on top of that where our tool also provides slightly comparable analytics as well, and, additionally, provides some prediction based analytics by discussing 'what may be happened if' cases.

The above-summarized works provide a collection of useful features to interpret game data. When compared to our proposed analytics platform, even though there are partial overlaps, we see major differences in the following aspects:

- These tools do not provide an *exploratory* data analytics platform. In contrast, our tool allows to modify model settings for different features and instantly presents users the updated results.
- In these tools, *interactivity* aspect is limited in the sense that user may see and/or analyze what has happened in a game. On the other hand, our tool is designed to be fully interactive to the extent that at any time, the user can modify the game data by changing player positions, defining new passes, and see the resulting effect instantly. This enables the user to create and run custom what-if scenarios and interactively explore what could happen in addition to what has happened in a particular moment in the game.
- Our tool blends both visual and conventional analytics in a well-balanced manner, and integrates these capabilities with data mining and analysis methods running in the backend. In contrast, some of the above tools emphasizes mostly visualization capabilities (e.g. Perin et al. 2013), while others focus on mostly conventional data analysis (e.g. Kang et al. 2006)
- Some of the above tools are commercial and proprietary, while some of them are just models or prototypes, which are not ready for prime time use. In contrast, our tool is open source, fully functional, and accepts soccer data in popular XML and JSON formats to allow others to analyze their own data.

3. Interactive data mining and machine learning features

3.1. Restricted pass success modelling

Once calculated, the probability of successfully completing an attempted pass in a particular region of the field can be useful both as a tactical aid and an input to more complicated calculations. As a tactical tool, it can be used to instruct players to take less risk when they are passing in low probability areas. More complicated metrics, such as a metric that calculates the passing efficiency of players can consider the difficulty of the passes a player attempts. For example, a defensive player, who routinely completes high risk, high reward passes can be considered to have high passing efficiency.

The main difficulty of calculating passing success is the number of different variables that effect the success of a pass. The positioning of the opposing players, the directions they are facing, their current speed, their individual attributes, such as their height, are just a few of the variables that effect the success of a pass. Acknowledging this difficulty, a restricted pass success model that only considers starting and ending points of a pass is developed. Although simplified, if sufficient data is provided, this model can still provide valuable information. Consider passes within the opposing team's penalty area. These passes are inherently high risk due to positioning of the other team, and the data will show that these passes failed with high probability in the past. A positive of property of this model is its high calculation speed. Since, the

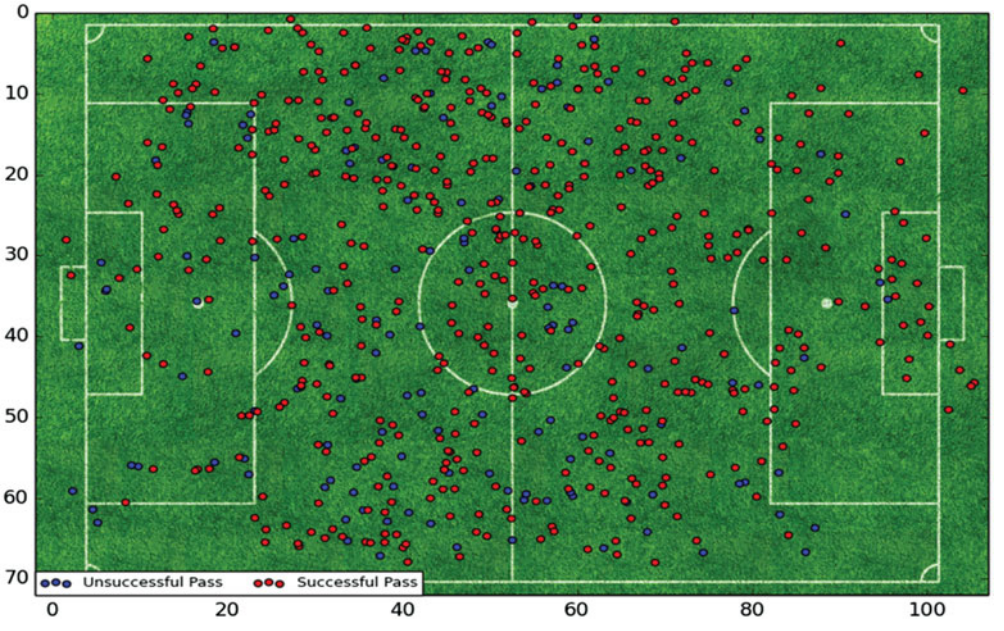

Figure 1. Ending points for successful and failed passes in a game.

model considers fewer parameters, once it is trained it can be used in real-time applications. The results can even be cached in a fast data structure to provide constant time access to the pass success results given the starting and ending points of a pass.

This feature employs a supervised learning approach to predict whether passes originating from a certain region R1 would be successful or not. More specifically, we train a SVM classifier on the successful and failed passes made during a game. SVM classifier is from an open-source package called scikit-learn and uses Radial Basis Function (RBF) as kernel. The features are pass starting and ending positions (i.e. [12.3, 30.0, 45, 60.2, Fail]). A total of 168 passes occurred in the analyzed game and 132 of them ended as successful, 36 of them ended as unsuccessful. Then, using the learned model, we predict the success of a pass originating from a user-defined region. Users may define the region by clicking on a point on the field, and providing a radius that defines a circular region around the clicked point. Then, our tool virtually creates all possible passes from the user-defined region to every other point in the field. The trained model is then employed to predict the success of each pass. The tool has an accuracy score of 83.75% with 10-fold cross validation. The results are shown to the user in the form of a heat map which indicates the regions of the field where a team or a player pass with high accuracy. Figure 1 shows the training data from a game, and Figure 2 shows the analysis result in the form of a heat map. Dark red coloured regions show that passes ending at those regions are highly likely to be successful. The visualization also includes the real passes from the game, which originate from the specified region (centered at ($x = 35$, $y = 20$)). Users may perform exploratory analysis by repeatedly running this feature after adjusting several parameters through the interface. User-adjustable parameters include training data scope

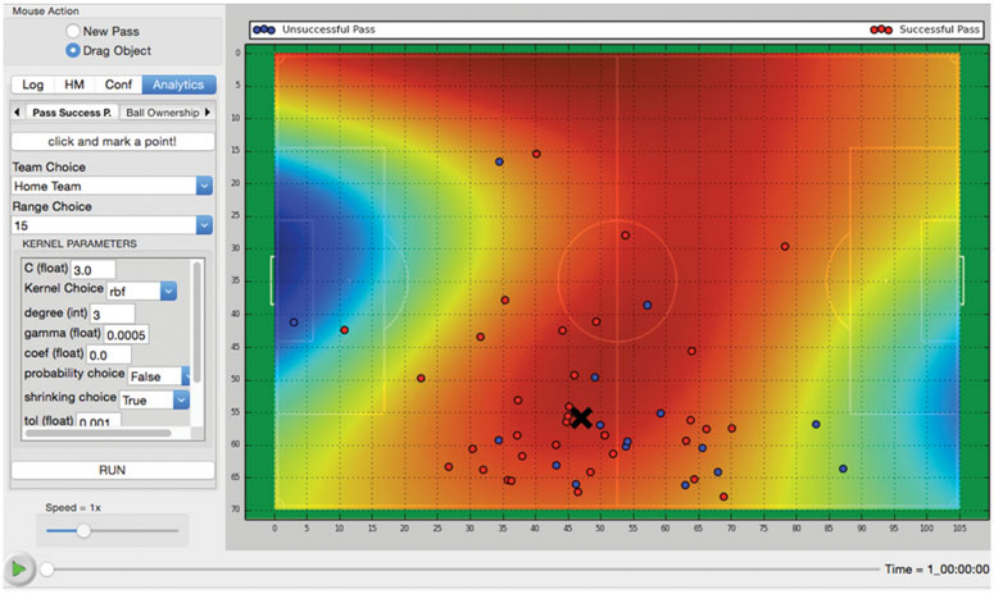

Figure 2. Heat map showing predictions for the success of passes originating from a user-selected point (i.e. marked with X) and its surrounding region as defined by the **user**.

(e.g. passes that belong to a particular team or player), region center and radius, SVM kernel and several other SVM parameters.

3.2. Ball ownership prediction

An important analysis aspect to compare the teams in a soccer game is the amount of time that each team has the ownership of the ball. Due to its small size and the technical limitations of the current sensor and camera technologies, it may not be always possible to detect the location of the ball accurately at all time points during a game. As a remedy, our soccer data analytics tool offers a tool to predict the team who owns the ball at any time point in a game.

We model this problem as a classification problem, and we train an SVM classifier to predict the ball owning team. Similarly, we used SVM from scikit-learn package with RGB as kernel. We define a rich set of features (currently 21 of them) to represent data points at each time unit. For instance, as one feature, we compute the correlation between movements of players. The idea is that the team who does not have the ball will need to chase the opponent team players either to get the ball or to prevent a player from getting a pass from his teammate. Hence, the number of correlated players in terms of their movements may give an idea regarding attacking (i.e. ball-owning) and defending teams. Other features include average speed of a team's members, total number of changes in speed and direction of players, covered distances, number of players running at high speed, etc. The accuracy of our prediction system is computed as 87% on sample game data with cross validation (svm parameters: c = 5, cache size = 300, gamma = 0.05 and kernel = rbf. The rest are default values). Currently, this analysis is done in the backend by an independent module, and integration into the

interface is work in progress. The tool will soon allow users to modify model parameters (e.g. kernel, regularization coefficient, features to be used, cache size, gamma, etc.).

3.3. Optimal shooting point prediction

When a player receives a pass, he can either make a pass to another teammate or shoot the ball to score a goal, if he is close enough the goal area. Furthermore, when the player decides to shoot, he has two options: (i) he may either move with the ball towards goal area to get closer and increase his chance to score, or (ii) shoot immediately without moving. Predicting the optimal shooting position may be important for two goals: (i) to evaluate the goodness of a received pass at a particular position, (ii) to assess and improve the decision-making capabilities of players.

In order to predict the optimal shooting point, we adopted a simulation type approach. More specifically, we virtually run a number of what-if iterations where we move the player with the ball towards the goal area, and compute the likelihood of scoring a goal from the new position. While the player with the ball moves towards the goal, we assume that nearby opponent players also move to block the attacking player as early as possible. To this end, for each nearby defending player, we compute an intervention point P on attacking player's moving path to the goal area such that the defending player is expected to reach P at the same time or earlier than the attacking player is. Then, each defending player is assumed to move to his intervention point while the attacking player moves towards the goal area.

We consider the current speed and direction of each player in our player movement model as described next. We first calculate the angle that the player has to turn to be headed towards his target point. Since the player can turn towards his target location either from his left or right, the angle that he needs to turn is chosen as the smaller of the two alternatives. Then, we get the components of the player's speed vector on the x and y axes. We assumed that x-axis lies on the line between player and target point, and y-axis is then positioned perpendicular to x-axis. In order to turn to towards his target point, the player has to increase or decrease his speed on each axis depending on the angle. For instance, in Figure 3, player 77 has a velocity vector in a direction shown by the purple arrow, and player 17 is running in a direction shown by the yellow arrow. Player 77 has to turn degree of Qt in order to catch player 17 at the target point marked by a star. Then, we consider the components of the velocity and acceleration vectors on x- and y-axes individually. Player 77 would decrease his speed on x-axis with deceleration of ax, which is the component of acceleration vector of player 77 on axis. At the same time, s/he would increase his speed on y-axis with acceleration of ay. The slow down and speed up process keep going until the angle Qt becomes 0. When we stop iteration, the final speed is player's new initial speed toward the target point. From there on, his speed will increase up until max speed ($v_{\max}$) with acceleration a. We consider $v_{\max} = 10\,\text{m/s}$ (FIFA Stats 2014) and $a = 2.8\,\text{m/s}^2$ (Akenhead et al. 2013) for all the players.

At each time unit, based on the updated position of the attacking and defending players, we compute (Cakmak et al. 2018) the chance of a goal if the attacking player decides to take a shot at his current location. We compute the goal chance of a player P using the following simple formula:

Figure 3. Optimal shooting point (marked with X) as predicted for player #8.

$$\text{Goal Chance}(P) = \frac{1}{d_1} * \frac{\min(\alpha,\ 180-\alpha)}{90} * \frac{1}{1+\text{Risk(defenders)}}$$

where d_1 is the distance of P to the goal area, α is the angle that P's path makes with the goal line. The overall risk component is inversely proportional to the distance of the defending player to the intervention point and proportional to the distance of the attacking player to the intervention point. The risk is further weighted by the likelihood of the required speed that the defending player needs to assume.

Finally, the optimal shooting point is predicted to be the one where a goal chance is computed to be the largest among all points on the path that the attacking player follows. Figure 4 shows the result of optimal shooting point prediction for player #8.

The proposed optimal shooting point prediction feature is relevant to the studies by Miller et al. (2014) and Lucey et al. (2014). Miller et al. (2014) proposes an unsupervised feature extraction method to represent the shooting success tendencies of individual NBA players. The authors consider the shooting point choices as a Poisson distributed process, and apply nonnegative matrix factorization to identify the descriptive features that minimally summarizes the shooting behaviour and outcome of a player. The proposed approach merely relies on the shooting point location in the court, and does not take into consideration the position and movements of other players.

Lucey et al. (2014) builds a supervised machine learning model to predict the likelihood that a shoot will result in scoring a goal in soccer. As different from Miller et al.'s work, the authors also consider other game features such the opponent team

Figure 4. Optimal shooting point (marked with X) as predicted for player #8.

players locations and counts, distance between defending players, etc. which we also consider in our model.

As different from these works, our goal is not only to assess the effectiveness of a shot from a particular position, but also to explore if there is a better alternative location to make the shot. That is, when a player receives a pass from another teammate close to the goal area, he/she needs to make a decision: should s/he make a shoot right away, or move with the ball to get closer to the goal area and then make his/her shoot for a better chance of scoring goal? Our tool is designed to act as a decision support system to help players make the most optimal decision during a training session with the help of a wearable device similar to Google glass with instant feedback. None of these related studies has such a setting in that they are directly focused on evaluating the current location, and do not consider alternatives. Hence, our proposed tool is more extensive and proper for player training purposes. This is consistent with the general 'what-if' scenario analysis focus of our study that mainly differentiates it from the other works. Having said this, the machine learning models employed in the above works may also be plugged into our approach as well, once sufficient amount of soccer data with location tracking becomes publicly available to the researchers. Currently, there is no such freely available repository to properly train machine-learning models.

4. Visual interactive data analytics

4.1. Interactive game replay and modification platform

The main screen of the tool offers the following features:

- The user can replay a portion of the game on the field in varying speeds (from 1× to 5×), pause the game, or go to a particular time point in a game using slider controls.

Figure 5. Interactive Game Replay – The main screen of our tool.

- The user may change the location of any player by dragging and dropping at any point.
- A yellow-bordered circle indicates the player who currently has the ball.
- Thin arrows show the paths the ball followed in most recent two passes. Between two passes if the player does not give a pass immediately and instead plays with the ball (i.e. moves with the ball), the movement is shown with dashed lines. The arrows and dashed lines are coloured the same as the team that has the ball. The current owner of the ball and his path are shown in yellow colour, and later, they turn into red.
- Important game events (foul, out, offside, etc.) are displayed at the center of the field for a temporary amount of time when they happen.
- The user can change the speed, direction, acceleration parameters for each player individually, or for all players at once.
- The user can save a game moment on the disk through file menu, and later load it back from the disk through the same file menu.

Figure 5 shows the main screen of the tool with interactive game replay in action.

4.2. Pass evaluation, definition, and modification

Passes are among the essential game elements in soccer that may define the winning team. Hence, performing pass analytics is invaluable in soccer data analysis. The tool at a high level offers two essential capabilities to this end: (i) quantitative evaluation of passes, (ii) new pass definition or the modification of existing passes.

4.2.1. Pass evaluation

The tool includes a built-in customizable pass evaluation model (Cakmak et al. 2018). In this model, a pass is evaluated based on the following dimensions:

- *Gain:* An effective pass usually decreases the number of players between the ball and opponent team's goalkeeper.
- *Pass Advantage:* An effective pass is targeted to a player who is in advantageous position to potentially make a good pass to another teammate or score a goal (i.e. 'assist' case) after receiving the pass.
- *Risk:* An effective pass usually involves low risk that increases the likelihood that pass will succeed, and does not put the pass target into a difficult position.
- *Triggering Another Effective Pass:* An effective pass usually leads to another effective pass (recursive definition)

Based on the above properties, we define *pass effectiveness score* as follows.

Definition *(Pass Effectiveness Score)*: Given a pass(P_1, P_2) from player P_1 to player P_2, assume that pass(P_1, P_2) is part of a pass sequence S in which it is followed by another pass pass(P_2, P_3) from player P_2 to player P_3. Then, *pass effectiveness score* is defined as follows:

$$\begin{aligned}\text{Effectiveness}\Big(\text{pass}(P_1,P_2)\Big)_{\text{NextPass: pass}(P_2,P_3)} = {} & w_1 \times \text{Gain}\Big(\text{pass } (P_1,P_2)\Big)+ \\ & w_2 \times \text{Pass advantage}(P_2)+ \\ & w_3 \times \text{Goal chance}(P_2)+ \\ & w_4 \times \text{Decision time}(P_2)+ \\ & w_5 \times \text{Effectiveness}\Big(pass(P_2,P_3)\Big)\end{aligned}$$

where referred functions (Gain, Goal Chance, etc.) are defined in Cakmak et al. (2018), and not included here for brevity. We refer readers to Cakmak et al. (2018) for details. w_i's are weights associated with each component of the effectiveness computation. The default values for these weights are learnt through training with a genetic optimization algorithm (Segaran 2007). However, a user may change the weights, and choose which components to include and exclude in the effectiveness model through the bottom left hand side panel on the interface (please see Figure 5).

During the game replay, for each pass, the tool visually shows the computed effectiveness real-time on the field next to the pass with subcomponents of the score shown on the left hand-side panel for each pass (please see Figure 5).

There are also several related studies in the literature. Maheswaran et al. (2012) employs a machine learning model to predict what team will get the rebound if a particular shot is missed. (i.e. a kind of binary classification). The model uses player locations and ball height as features. Similarly, Horton et al. employs a machine learning model to classify passes into three labels, that is, 'good', 'ok', and 'bad'. Although their model is complex and takes advantage of a large number of spatial and other features, it requires a manually created pass labels, and even then, the assigned pass labels are too coarse and not differentiable (78% of passes had the same

label, 'ok'). In our pass evaluation feature, our goal is to quantitatively evaluate the 'effectiveness' of a pass, rather than doing a binary classification, that is, 'good pass' or 'bad pass'. We could still employ machine learning models to this end, if there was labeled training data that includes a large number of passes in different settings with their true effectiveness values as determined by experts. Such data is not available, and is costly to generate. Therefore, our approach is more of a heuristic approach due to the lack of labeled pass evaluation data.

The EPV approach of Cervone et al. (2016) is particularly interesting. The authors propose to compute expected point value (EPV) for each movement in basketball games. To some extent, it may be applicable to soccer data as well, but, obviously, more research is needed. The immediate concerns (hence, research questions) on the applicability of EPV approach on soccer data would be as follows:

i. EPV approach employs a kind of Markov Model with states and transitions between these states. The field is divided into discrete regions. The number of states is proportional to [number of players] * [number of field regions] * [whether the player is being defended]. Given the number of players and the size of the soccer field, the number of states would be somewhat large. Training such a model would require an enormous amount of training data to have a sufficiently generic model for each player. It may be challenging to obtain such data given the size of probable path space.
ii. Basketball is a game where a large number of points are scored in relatively smaller sequence of movements. Hence, connecting each movement to a score/point value is more intuitive. On the other hand, in soccer, a lot of games end with no goals scored, and in many others, one or two goals would be scored despite hundreds (and sometimes thousands) of passes made during the entire game. Hence, it may be challenging to compute EPVs for soccer.
iii. Related to item (i), the computational complexity of coming up with the EPVs for passes in soccer may be too high, and real-time pass assessment may not be possible. In our current setting, real-time evaluation is critical, as one of the main motivations of our proposed tool is to give instant feedback to players and coaches during training sessions.

The whole focus of EPV is how each movement will contribute to a team's goal of scoring. Inherent to this approach is the assumption that the movement is successful. On the other hand, our assessment of a pass directly considers the likelihood that whether such as pass can be made successfully. The goal here is to value high gain low risk movements over high risk high gain movements to guide players and coaches on possible pass alternatives. EPV approach ignores this aspect.

4.2.2. Manual pass definition and modification

The tool allows the user to define a new pass interactively to run *what-if scenarios*. In order to define a pass, the user first selects 'New Pass' button at upper left corner. Then, the user draws a pass with the mouse from one player who initiates the pass to another player who receives the pass. The pass is visually represented by an arrow

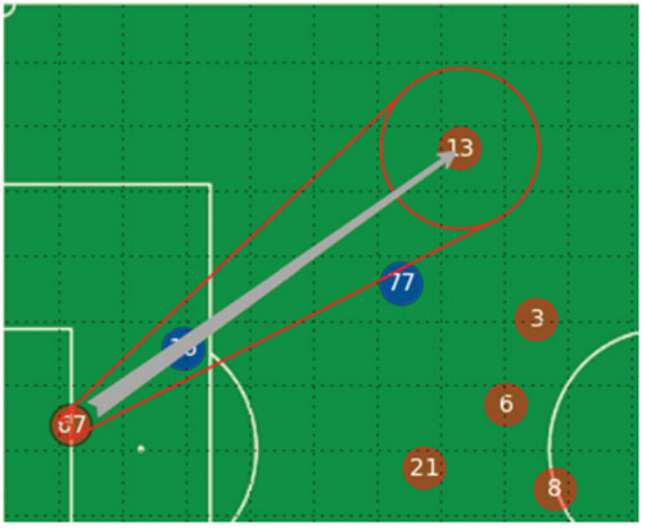

Figure 6. Visualization of the risk area for a pass.

from pass source to the target player. This feature enables the user to see the result of modifying the organization of players or some configuration parameters. For instance, if the user changes the location of any player for a certain pass, there is no need to redefine the pass. Instead, the tool recalculates the output of pass effectiveness model instantly and shows them on the logger area (on the left panel).

For each studied pass, a *risk* area is determined and visualized around the pass (see Figure 6). The risk area is defined such that any opponent team player included in the risk area may intervene the pass. The intervention risk of a pass by an opponent player in the risk area is inversely proportional to his distance to the pass path and distance to the target of the pass.

4.2.3. Case study: interactive pass analytics

In this case study, we illustrate a scenario where a user defines a pass, and then changes the organization of players. Then, she instantly sees the effect of this change. Next, s/he investigates with the help of the proposed tool why the observed pass evaluation changes take place.

Figure 7 is a screenshot of a position during the game and blue team is on attack. Player 88 passes the ball to player 25 (shown with an arrow). The panel on the left of the field provides detailed quantitative evaluation of the pass (*overall_risk:11.30, gain: 48.9, pass_advantage: 1.47, goal_chance:1.50, effectiveness: 768.06*). Now, the user changes the organization of players a little bit by moving player 25 to north east slightly. Figure 8 shows the organization of the players on the field after the change. The user observes that the small movement that she introduced caused dramatic changes in some pass evaluation components. For instance, the *overall risk* goes to 0 from 11.3.

Then, she turns on risk area visualization for the pass through the view menu of the tool to investigate this change. Risk area of a pass (shown in red borders in Figure 9) represents the region where the pass may be intervened by an opponent player. She notices that the reason for the observed change in evaluation is due to the

Figure 7. A sample passing scenario created manually by a user.

Figure 8. The same as Figure 7 except that player 25 has been moved slightly.

fact that opponent players 8 and 20 are now out of the risk area so that they do not pose any risk to intervene the pass.

4.3. Heatmaps

One of the ways that the tool employs visual analytics in a novel way is to provide analysis of what may have happened in a game. Heatmaps allow users to visually

Figure 9. Risk area visualization for the pass in the previous figure.

study and inspect those scenarios over all soccer field positions. This provides a global view of the game dynamics as well. In this section, we briefly describe three sample heat map-based visual analytics features that our tool offers. These visualizations focus on the effect of position taking decisions that players make during making a pass, receiving a pass, and defending against the opponent team.

Leading existing tools (such as ProZone, Sentio, SoccerStories and etc.) generally focus on what happened in the game, and provide player/team based heatmaps which represent the region that each player/team dominates. Besides, heatmap visualizations summarizing player/team performance are also featured in these tools. In addition, some also provide heatmaps demonstrating the frequency of events, such as shooting points, successful passes, etc. In comparison, our tool also features similar heatmap visualizations. Moreover, what makes our tool unique in terms of heatmaps is that we provide heatmaps not only on 'what happened' in the game but also on 'what might have happened' by allowing the user to modify the player positions, passes, etc. In other words, in addition to analytics based heat-maps, we also provide prediction-based ones, for example, pass success prediction. Furthermore, since we have an embedded pass effectiveness model (Cakmak et al. 2018), we also provide unique heatmaps regarding the 'quality' aspect of passes.

The tool computes the heatmaps as follows: Depending on the chosen heatmap type and resolution setting, the selected player (passing player, targeted player, or a defender) is imaginarily moved to all (x, y) coordinates on the field. At each point, all pass evaluation components are computed and stored in a memory cache. Then, with respect to the chosen pass evaluation component, the corresponding v_min and v_max values for the chosen heatmap type are automatically determined by normalizing the computed values based on mean and standard deviation. Finally, each point is

coloured according to its computed value, and the resulting heat map is displayed to the user.

We provide sample heat map visualization screenshots in Section 4.3.1 and do not repeat them here for brevity.

- *Position taking of a passing player:* When a player receives a pass, one of the important decisions the player needs to make is whether to continue keeping the ball and possibly move on the field with the ball, or make a pass to another teammate. This decision may directly affect whether the pass will be a successful one or not. For a given pass (artificially created or happened in a game), our tool allows to perform a visual analytics on the risk/goal chance/gain/pass advantage/effectiveness assessment of the pass based on the position of the pass-initiating player. In other words, it gives a sense of how would the risk involved in a pass will change if the pass-initiating player was positioned at different points on the field.
- *Position taking of a pass target:* In this type of analytics, for the above setting, the analysis focus on the position of the pass target, and how it affects the pass evaluation.
- *Position taking of a defender:* While the opponent team is attacking, one of the main aims of defending players is to prevent the passing between attacking team players. In this visual analytics feature, the focus is on evaluating the effect of a defender's position taking on the assessed value of a possible pass between attacking team members. Before initiating this analysis, the user chooses a defender player that s/he wants to analyze in terms of the player's position taking decision.

To use the above visual analytics features, users first choose the type of analytics that they want to perform (e.g. the effect of position taking of a defender on a pass's effectiveness). Then, interactively create a pass, and run the corresponding analytics, which instantly provides the resulting heat map visualization. The user may also adjust the resolution of the analytics which determines for many distinct points on the field the corresponding what-if cases will be investigated.

4.3.1. Case study: visual analytics of player position taking when receiving a pass

In this section, we present a sample case study to perform visual analytics to investigate the relationship between the position taking of a pass target and the risk of the pass. A user defines a pass from player 5 to 9, and runs the analysis with highest resolution, which involves evaluating the possibility that player 9, could be at any point on the field when she received the pass. Figure 10 shows the resulting heat map.

When the user moves any players on the field, the tool instantly updates the risk heat map to take into account the change. In this case study, the user wants to analyze how could player 6 in the away team be better positioned to pose a risk for a pass that player 5 in the home team would make near goal area. Using the drag-and-drop feature of the tool, she moves the node that represents player 6 of the away team to slightly north of its current position. Figure 11 shows the re-computed heat map after this change. In the recomputed heat map, the user observes that the red region (high risk) around away team's goalkeeper expands. Based this visual analytics

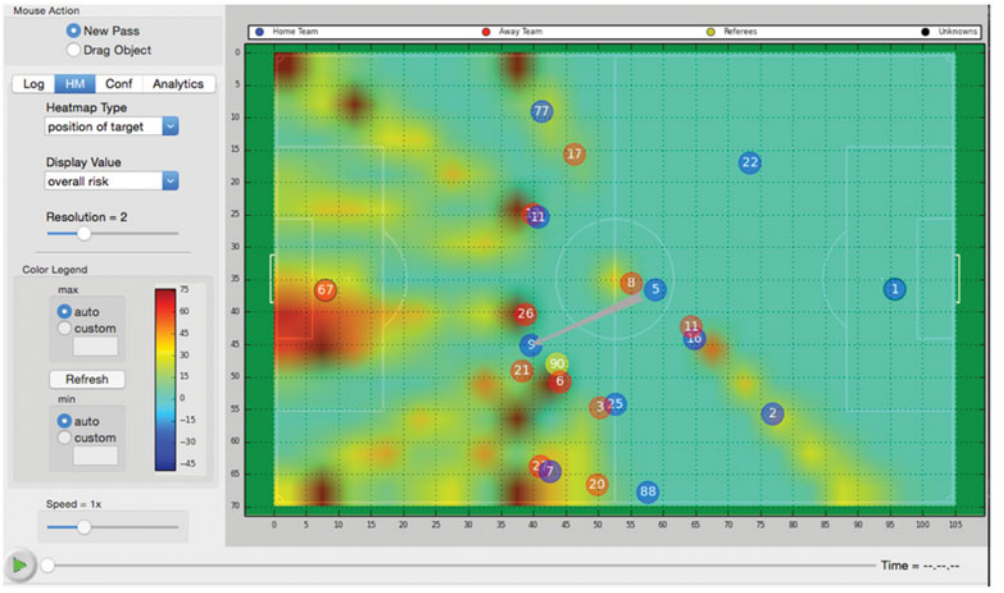

Figure 10. Heat map illustrating the effect of position taking when receiving a pass.

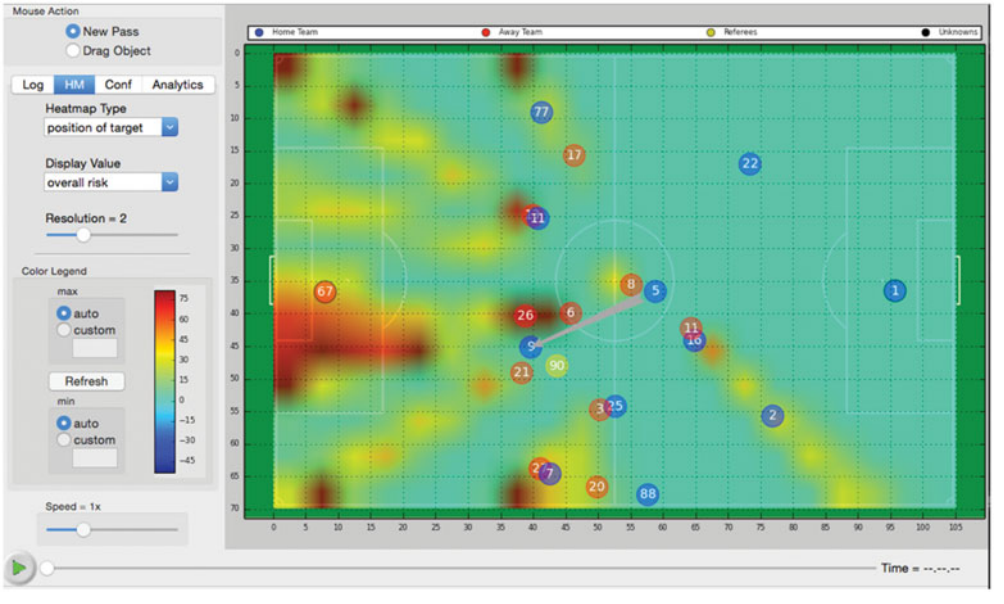

Figure 11. Re-computed heat map after moving player 6 (away team) to the north.

study, she concludes that if player 6 was located in the new position she could better defend her team's goal area.

4.2. Dominant region analysis

In modern soccer, teams usually organize themselves to dominate all or critical regions of the field during the game. Hence, the analyzing regions that each player

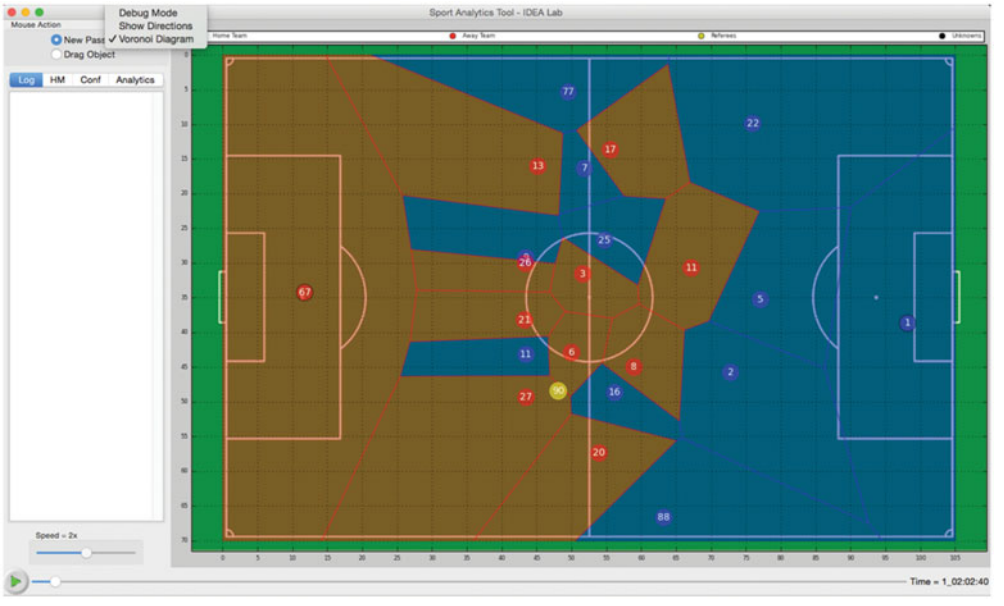

Figure 12. Voronoi cells drawn real-time during game replay.

dominates during any time point in the game may provide invaluable insights regarding the organization capabilities of a team. To this end, our tool offers a visual analytics feature which displays the region controlled by each player through both Voronoi diagram (Aurenhammer 1991) in real-time and Dominant Region (Taki and Hasegawa 2000) in almost real-time with respect to resolution setting while the game is being replayed.

In the Voronoi Diagram, for each player, the tool computes a Voronoi cell which represent the region of the field that the player dominates. Voronoi cell of a player is coloured with his team's colour in order to facilitate interpreting teams' weaknesses and strengths in dominating the entire field. Figure 12 illustrates Voronoi diagram visualization during game replay.

The standard Voronoi considers only the positioning of players. Hence, it may not give accurate results since players also have direction and speed. Our tool also supports an enhanced version of the diagram, also known as the 'Dominant Region' (Taki and Hasegawa 2000) that considers players' speed and direction. More specifically, for each coordinate on the field, the arrival time of each player to that coordinate is calculated by considering both direction and speed in addition to his positioning. Then, this coordinate is assigned to that player and added among the set of coordinates that s/he dominates. After the assignment process is done for all coordinates in the field, for each player, concave-hull is extracted by computing the alpha shape out of the set of coordinates. Finally, this concave-hull is converted to a polygon and coloured with player's team in order to facilitate interpreting teams' weaknesses and strengths in dominating the entire field. Figure 13 illustrates a sample dominant region visualization. When considering speed and direction of players to compute arrival time of a player at a certain point on the field, we employ the same approach as summarized in Section 3.3.

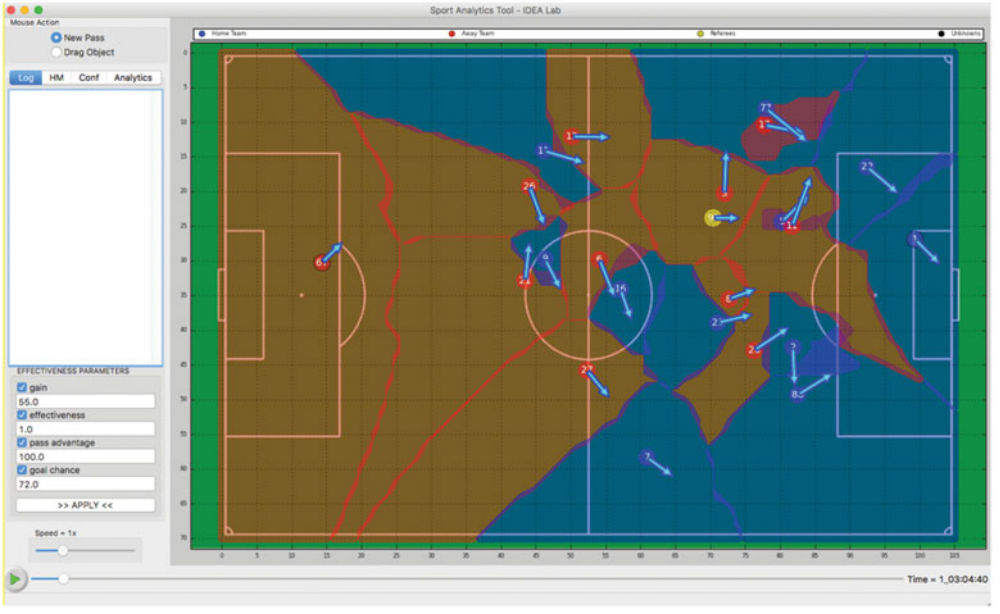

Figure 13. Dominant regions of players.

Our tool also offers an exploratory visual analytics feature that allows users to perform dominant region analysis at team and player levels with various customizations. As an example, a user may choose to run dominant region analytics for a team only when the team is defending against the opponent (i.e. the opponent has the ball) just to evaluate the position taking of players during defense. Alternatively, independent of who has the ball, a coach may focus on a particular interval of the game (by setting start and end times of analysis on the interface) and run analysis on how successful his team was in pressing in the opponent team's field (by restricting the analysis to opponent team's field). Figure 14 illustrates the result of a sample dominant region analysis where the analysis is restricted to the first five minutes of a game.

5. Quick, high-level data analysis features

A number of analysis features are expected of all sports data analytics tools such as running distance. In this section, we are listing features that are included as part of the visual analytics tool for completeness although their research contributions are limited. (Castellano et al. 2014) and (Gudmundsson and Horton 2017) provide a thorough overview of some of the tools that provide similar features in the literature.

5.1. Running distance computation

A player's contribution to the game at a high level may be analyzed by computing player's total running distance. Our tool offers such an analysis feature with several customizable configurations such as choosing a team, defining filters, such as excluding player movements that are under a certain speed, or happens during game

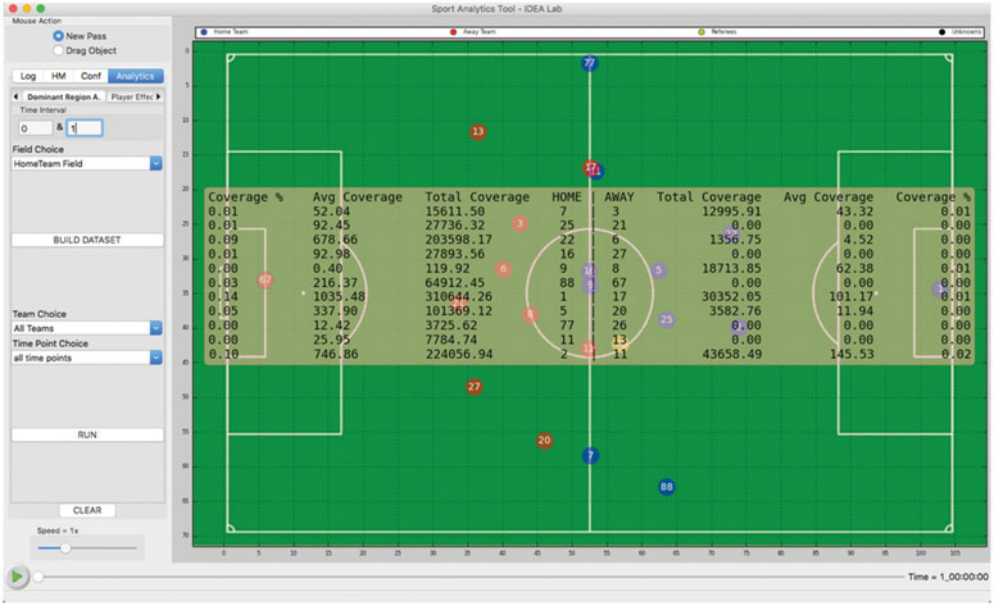

Figure 14. A sample dominant region analysis.

stopping times. Figure 15 illustrates a sample running distance analysis that covers all players from home and away teams without any filtering on the included movements.

5.2. Ball ownership and handling analysis

Another indicator of a player's performance in a game may be how the player handles the ball during the game. Our tool offers a quick ball ownership analysis, which reports duration that each player has the ball during the game. In addition, the number of times a player makes a pass, loses the ball, or steals the ball is also included. Figure 16 illustrates a sample ball ownership and handling analysis.

5.3. Player effectiveness analysis

A deeper analysis of how a player handles the ball would be analyzing the effectiveness of the passes the player made during a game. As discussed in previous sections, our tool quantitatively evaluates each pass for its effectiveness. Based on this pass evaluation infrastructure, our tool includes a high-level quick analysis tool that computes and reports the effectiveness of passes by each player. Figure 17 illustrates a sample player effectiveness analysis.

6. Evaluation

6.1. Accuracy evaluation

Computational modeling of passes between players, computing an effectiveness score for passes, and quantitative and heatmap-based visualization of the computed scores are

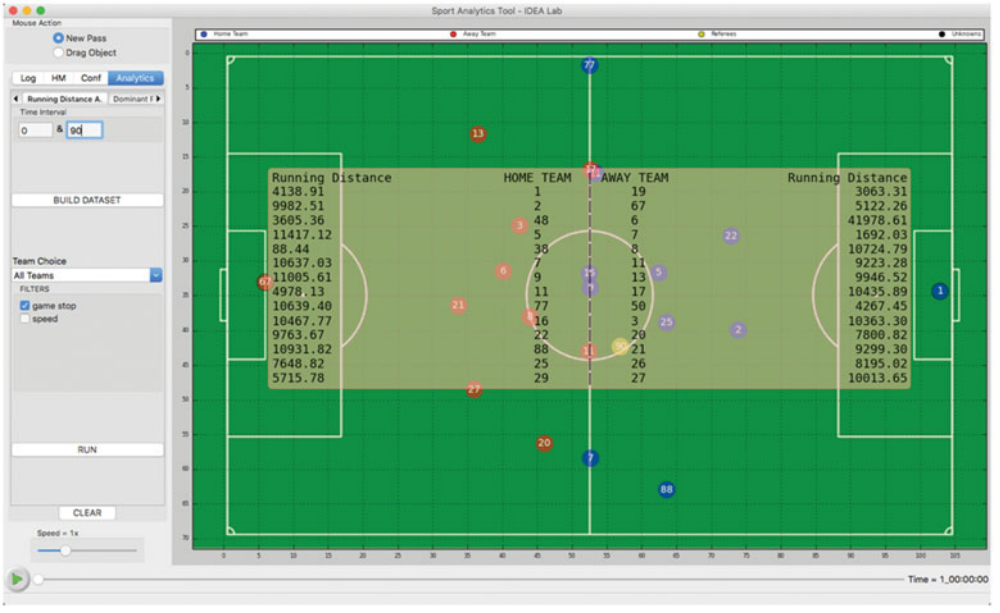

Figure 15. A sample running distance analysis that covers all players with no filters.

Figure 16. A sample ball ownership analysis.

essential components of the proposed software system. In order to evaluate the accuracy of the pass modeling, we performed a user-study with domain experts (i.e. professional soccer trainers and players). To this end, a survey (Pass Evaluation Survey 2016) with different passing scenarios are presented to the domain experts, and they are asked to rank the alternative passes in terms of their contribution to the passing team. Then, expert answers are compared to our tool's evaluations. Details of the user-study are included in another paper (Cakmak et al. 2018). Here, we summarize the results. We

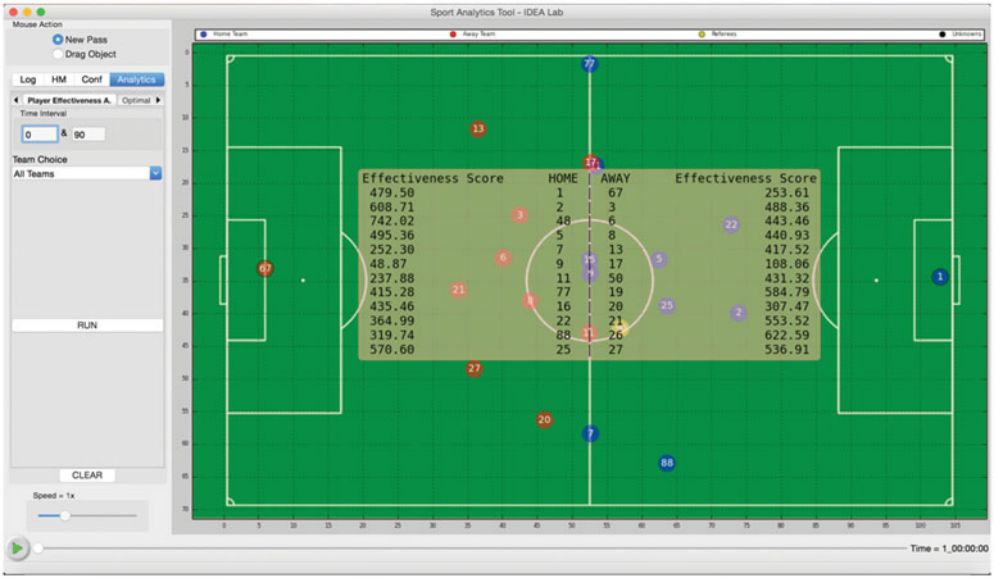

Figure 17. A sample player effectiveness analysis.

employed genetic optimization to estimate the best set of values for the weights in our pass model. Besides, we performed k-fold cross validation with leave-one-out strategy, and the accuracy of our tool in evaluating different pass scenarios is over 94%. We kindly refer the readers to Cakmak et al. (2018) for details.

6.2. Feature-based evaluation in comparison with the state of the art

In this section, we compare our analytics tool with the existing similar tools (both academic and commercial software) which is summarized in Section 2. Table 1 presents a featured-based comparison of our proposed tool with some of the leading tools in the field. Namely, we compare six sport analytics tools: Our proposed tool in this paper (IDEA Lab Sports Analytics Tool (SAT)), SoccerStories (Perin et al. 2013), SportVis (Cox and Stasko 2006), Sentio,[1] ProzoneSports[2] and PerformaSports.[3] The features are divided into three subcategories for improve the readability. The first subgroup of features are those that are unique to our tool. The second subgroup of features are those that are commonly offered by the majority of the tools, and the last subgroup of features are those that are not included by our tool. As the table illustrates clearly, our tool is unique in that it enables users to modify the positions, create passing scenarios, and allow them to run what-if scenarios. In addition, providing a platform for prediction-based analytics is just another unique feature that is further strengthened by the save/load functionality. Our tool stands out as the best one among not only academic prototypes, but also commercial products, since it provides useful unique features in addition to offering features that are already provided by the others.

7. Conclusion

The presented tool offers interactive data mining and machine learning features adapted for soccer data analytics, such as pass success prediction, optimal shooting prediction,

Table 1. Feature-based comparison with the state of the art.

		Academic studies			Commercial tools		
		IDEA Lab SAT	Soccer Stories	SportVis	Sentio	Prozone Sports	Performa Sports
Unique to our tool	pass success prediction	✓	✗	✗	✗	✗	✗
	ball ownership prediction	✓	✗	✗	✗	✗	✗
	interactive modification	✓	✗	✗	✗	✗	✗
	manual pass definition and modification	✓	✗	✗	✗	✗	✗
	pass effectiveness analysis	✓	✗	✗	✗	✗	✗
	save/load a particular position	✓	✗	✗	✗	✗	✗
	support for what-if scenarios	✓	✗	✗	✗	✗	✗
	optimal shooting point prediction	✓	✗	✗	✗	✗	✗
Common in most tools	interactive game play	✓	✗	✗	✓	✓	✓
	pass evaluation (success, failure)	✓	✓	✓	✓	✓	✓
	heatmaps and/or graphs	✓	✓	✓	✓	✓	✓
	dominant region analysis	✓	✓	✓	✓	✓	✓
	running distance computation	✓	✓	✓	✓	✓	✓
	open source	✓	✓	✓	✗	✗	✗
	import data using common formats (xml, json, etc.)	✓	✓	✓	✗	✗	✗
	player effectiveness analysis	✓	✓	✓	✓	✓	✓
	tracing	✓	✓	✓	✓	✓	✓
	ball ownership analysis	✓	✓	✓	✓	✓	✓
Missing from our tool	video analysis	✗	✗	✗	✓	✓	✓
	game pattern detection	✗	✓	✓	✗	✗	✗
	video exporting	✗	✗	✗	✓	✓	✓
	real-time analysis	✗	✗	✗	✓	✓	✓
	player tagging	✗	✗	✗	✓	✓	✓

etc. Moreover, our tool includes an extensive set of visual analytics features that allows visually investigation of the underlying soccer data ranging from interactive pass analytics to player position taking analysis with heat maps. Last but not the least, our tool offers high-level, quick data analysis capabilities such as running distance, ball handling, and player effectiveness computations and reports. Even though some of the above features are provided in other systems, our tool differentiates from other works in that it provides interactive analytics features, which allow it to run what-if scenarios in settings that are configured by users. This allows soccer professionals to analyze not only what has happened in a game, but what could happen as well.

Notes

1. http://www.sentiosports.com/
2. http://prozonesports.stats.com/
3. http://performasports.com/

ORCID

Ali Cakmak http://orcid.org/0000-0002-1382-6130

References

Akenhead R, Hayes PR, Thompson KG, French D. 2013. Diminutions of acceleration and deceleration output during professional football match play. J Sci Med Sport. 16(6): 556–561.

Aurenhammer F. 1991. Voronoi diagrams – A survey of a fundamental geometric data structure. ACM Comput Surv. 23(3):345–405.
Baumer B, Zimbalist A. 2014. The sabermetric revolution. Philadelphia: UPenn.
Cakmak A, Uzun A, Delibas E. 2018. Computational modeling of pass effectiveness in soccer. Adv Complex Syst (ACS). 21(03n04):1–28.
Castellano J, Alvarez-Pastor D, Bradley PS. 2014. Evaluation of research using computerised tracking systems (Amisco® and Prozone®) to analyse physical performance in elite soccer: a systematic review. Sports Med. 44(5):701–712. http://dx.doi.org/10.1007/s40279-014-0144-3
Cervone D, D'Amour A, Bornn L, Goldsberry K. 2016. A multiresolution stochastic process model for predicting basketball possession outcomes, J Am Stat Assoc. 111(514):585–599.
Cox A, Stasko J. 2006. Sportsvis: Discovering meaning in sports statistics through information visualization. In: Compendium of Symposium on Information Visualization. p. 114–115.
FIFA Stats. 2014. http://www.espnfc.com/fifa-world-cup/story/1888364/arjen-robben-of-netherlands-becomes-worlds-fastest-player, retrieved on April 24, 2017.
Goldsberry K, Weiss E. 2013. The Dwight effect: A new ensemble of interior defense analytics for the NBA. Sports Aptitude.
Gudmundsson J, Horton M. 2017. Spatio-temporal analysis of team sports. ACM Comput Surv. 50(2):1. 11
Gudmundsson J, Wolle T. 2014. Football analysis using spatio-temporal tools. Comput Environ Urban Syst. 47:16–27.
Kang CH, Hwang JR, Li KJ. 2006. (December). Trajectory analysis for soccer players. In: Sixth IEEE International Conference on Data Mining Workshops, 2006. ICDM Workshops 2006. IEEE. p. 377–381
Legg PA, Chung DH, Parry ML, Jones MW, Long R, Griffiths IW, Chen M. 2012. MatchPad: interactive glyph-based visualization for real-time sports performance analysis. Comput Graphics Forum. 31(3pt4):1255–1264.
Lucey P, Bialkowski A, Monfort M, Carr P, Matthews I. 2014. Quality vs quantity: improved shot prediction in soccer using strategic features from spatiotemporal data. In Proceedings of 8th Annual MIT Sloan Sports Analytics Conference, p. 1–9.
Maheswaran R, Chang Y-H, Henehan A, Danesis S. 2012. Deconstructing the rebound with optical tracking data. In MIT Sloan Sports Analytics Conference.
Miller A, Bornn L, Ryan PA, Goldsberry K. 2014. Factorized point process intensities: a spatial analysis of professional basketball. In ICML, p. 235–243.
Nakanishi R, Maeno J, Murakami K, Naruse T. 2009. An approximate computation of the dominant region diagram for the real-time analysis of group behaviors. In RoboCup p. 228–239.
Pass Evaluation Survey. 2016. Available online at https://www.dropbox.com/s/ssnvr68wtuhwivg/SurveyWithQs.pdf?dl=0
Perin C, Vuillemot R, Fekete JD. 2013. SoccerStories: a kick-off for visual soccer analysis. IEEE Trans Visual Comput Graphics. 19(12):2506–2515.
Scholkopf B, Smola AJ. 2001. Learning with kernels: support vector machines, regularization, optimization and beyond (Adaptive Computation and Machine Learning.). Cambridge: MIT Press.
Segaran T. 2007. Programming collective intelligence: building smart web 2.0 applications. O'Reilly Media, Inc.
Taki T, Hasegawa JI. 1998. (December). Dominant region: a basic feature for group motion analysis and its application to teamwork evaluation in soccer games. In Electronic Imaging'99 (pp. 48–57). International Society for Optics and Photonics.
Taki T, Hasegawa JI. 2000. Visualization of dominant region in team games and its application to teamwork analysis. In: Computer Graphics International, 2000. Proceedings. IEEE. p. 227–235.