SoccerStories: A Kick-off for Visual Soccer Analysis

Charles Perin, Romain Vuillemot, and Jean-Daniel Fekete, *Senior Member, IEEE*

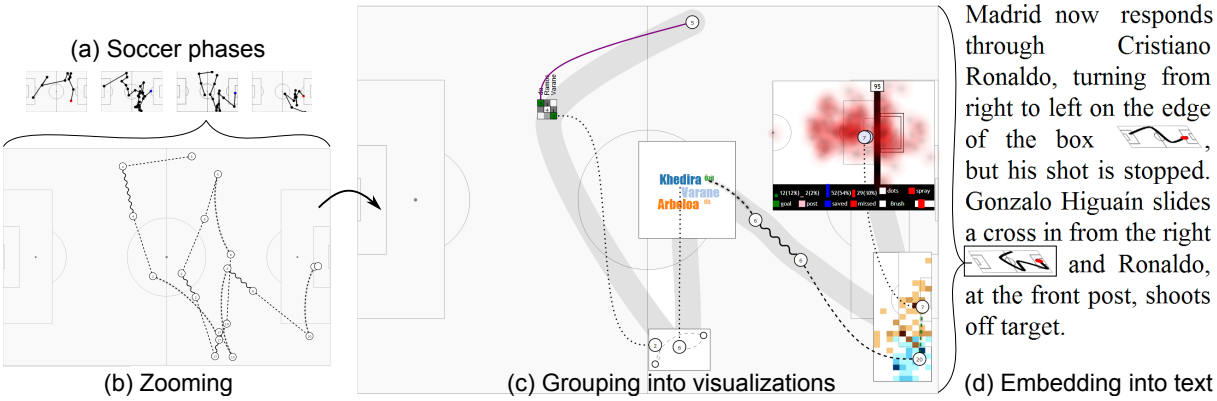

Fig. 1. Using *SoccerStories*: (a) navigating among soccer *phases* of a game; (b) mapping a phase on a focus soccer field; (c) exploring the phase by grouping actions into tailored visualizations; and (d) communicating using *SportLines* embed into text.

Abstract—This article presents *SoccerStories*, a visualization interface to support analysts in exploring soccer data and communicating interesting insights. Currently, most analyses on such data relate to statistics on individual players or teams. However, soccer analysts we collaborated with consider that quantitative analysis alone does not convey the right picture of the game, as context, player positions and phases of player actions are the most relevant aspects. We designed *SoccerStories* to support the current practice of soccer analysts and to enrich it, both in the analysis and communication stages. Our system provides an overview+detail interface of game phases, and their aggregation into a series of connected visualizations, each visualization being tailored for actions such as a series of passes or a goal attempt. To evaluate our tool, we ran two qualitative user studies on recent games using *SoccerStories* with data from one of the world's leading live sports data providers. The first study resulted in a series of four articles on soccer tactics, by a tactics analyst, who said he would not have been able to write these otherwise. The second study consisted in an exploratory follow-up to investigate design alternatives for embedding soccer phases into word-sized graphics. For both experiments, we received a very enthusiastic feedback and participants consider further use of *SoccerStories* to enhance their current workflow.

Index Terms—Visual knowledge discovery, visual knowledge representation, sport analytics, visual aggregation

1 Introduction

A new generation of soccer data is now available, as some companies [13] collect and provide extensive data covering almost all professional soccer championships, with a wealth of multivariate information related to time, player positions, and types of action, to name a few. Currently, most analysis on such data relate to statistics on individual players or teams. For instance, statistics on *"team ball possession"* and *"number of goal attempts for team A or B"* are popular on websites, TV and newspapers (Figure 5) and often accompanied by bar charts or plots on a soccer field. However, games are spatio-temporal data and each action has on average a dozen of attributes, which can be described in up to 50 different ways.

A soccer game is made of episodes—called *phases*—which, for example, start in the middle of the field, is followed by a kick to cross the ball on the other side to end up with a shot towards the goal (Figure 1). Such a phase is not fully captured by quantitative analysis: we interviewed four soccer experts who heavily rely on their own observations or reports of the games only first, and then use statistics or visualization to communicate their analysis. They acknowledge the need of a visual *big picture* of a game as a starting point for their analysis, as well as for further inclusion in their articles. From a thorough review of visualizations related to soccer, we observed there is no one-size-fits-all visualization for this application domain, as soccer is composed of multiple *phases*. Phases are sequences of actions from one team until it loses the ball and, according to our experts, provide the optimal semantic level to browse games. Keeping this level of abstraction is important because experts often write articles under strict time constraints and cannot explore the complete data space of a game. Browsing games with phases also enables experts to quickly find the key phase that explains the outcome of a game (*e. g.,* a red card, a goal) or the phase that is the flagship of a team's tactic.

We designed *SoccerStories*, a system for the visual exploration of soccer phases. It uses a series of soccer-related visualizations, called *faceted views*, that we selected—after a thorough review of current visualizations for soccer data —for each group of actions within a phase, connected and ordered on a soccer field . We provide design guidelines for each visualization and their layout in a spatio-temporal flow. We evaluated *SoccerStories* for exploring and communicating findings about games that would enrich experts in tactics current reports on blogs or newspapers. We also explored design alternatives for embedding soccer phases into word-sized graphics. In these experiments we received a very enthusiastic feedback and participants consider further use of *SoccerStories* to enhance their current workflow.

SoccerStories tackles the challenge of using visualization for complex data in an application domain which so far has been dominated by game statistics (i.e. numbers). Our contributions are the following:

- *Charles Perin is with INRIA and Universite Paris-Sud. E-mail: charles.perin@inria.fr.*
- *Romain Vuillemot is with INRIA. E-mail: romain.vuillemot@inria.fr.*
- *Jean-Daniel Fekete is with INRIA. E-mail: jean-daniel.fekete@inria.fr.*

Manuscript received 31 March 2013; accepted 1 August 2013; posted online 13 October 2013; mailed on 4 October 2013.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

1. A review of existing soccer visualizations and their matching with soccer's most important actions and phases;
2. *SoccerStories*, a system that combines several visualizations to explore and communicate game phases;
3. Two qualitative experiments that validated *SoccerStories* for exploring and reporting on tactics analyses, and the design alternatives of compact representation of games as word-sized graphics to embed in articles.

2 Problem Description

In this section we introduce soccer rules and provide a representative sample of qualitative and quantitative analysis of a game. We then present interviews with experts with whom we collaborated to understand their current workflow for game analysis.

2.1 Background on Soccer

A soccer field is divided into areas, visible with white landmarks: a center circle, penalty areas, corner quarter-circles. Games last 90 minutes divided in two half periods, and the winning team is the one that scores more goals than the other. Players have roles (goalkeeper, fullback, midfielder and forward) indicating their theoretical position on the field, but the actual position is adjusted according to tactics (*e. g.,* the most commonly used tactical *lineup* is called 4-4-2: 4 defenders, 4 midfielders, and 2 forwards).

Each interaction with the ball generates an *action*, and is qualified according to multiple criteria: the part of the body that touches the ball, the strength of the kick or header (striking the ball with the head), the direction or the outcome of the action. A series of actions by one team is called a *phase*. Phases are separated by transitions that occur whenever the other team touches the ball. Figure 1(b) shows actions as circles, plotted on the field and Table 1 enumerates the main actions in soccer and the field areas where they take place.

2.2 Example of Qualitative/Quantitative Analysis

We briefly introduce a running example we use throughout this paper: UEFA Champion's League 2012/2013's 1/8 final game between Real Madrid and Manchester United. The game ended in a 1-1 draw. *ZonalMarking* [26], often cited by experts as the standard for game analysis, provided the following qualitative insight: *"Ronaldo scored the equalizer with a superb header, it's fair to say United's approach against him worked reasonably well. Yes, he had a typical number of attempts from goal, but the majority were from long-range, with three attempts from free-kick situations. By forcing Ronaldo over to the opposite side, United had moved him away from his preferred position and he plays on the left because he's most prolific from that zone. His tendency to drift away from that flank was also helpful to Rafael, who had an extremely nervous period towards the end of the first half, but was rarely tested after half-time."* Here are some samples of quantitative analysis of this particular game: Real Madrid dominated with a 61% ball possession, and they roughly committed the same number of fouls. Real Madrid attempted to score nearly twice as much as Manchester United (28 goal attempts against 13 for Manchester) but nearly the same number were on target (8 against 6). A series of visualizations concludes the analysis with horizontal bar charts (Figure 5) for each statistic, players lineup on a soccer field and shots on goal distribution for each team.

2.3 Interviews with Experts

To better understand the mechanisms behind game analysis (*e. g.,* supporting tools, focus on specific parts of a game) we conducted interviews with four soccer experts of various types: two online journalists, one Opta Sports soccer specialist and one professional sports trainer.

The first expert—journalist #1—writes tactical analyses for different venues: a specialized blog and featured articles in newspapers. His workflow consists in watching a game while writing down his thoughts on it and then using his notes as basis for an article which will be finally published. While writing an article he often uses statistics and very simple visualizations to complement his notes and better support his reasoning, namely heatmaps to plot the average position of players, and an interactive slider to navigate through the game by specifying time intervals of particular interest. Journalist #1 was integrated at the early stages of the designing process of our system, following MILCs method [22]. We closely worked with him until the final evaluation.

The second expert—journalist #2—is the editor of an online soccer newspaper. While watching a game, he writes down highlights on a table with one column for each team and lines containing descriptions of the particular events, along with the time they took place and the players that were involved. If an event was particularly important (*e. g.,* goals, close calls, red cards) he highlights them through color or markings, with the final table being essentially a handmade visualization that shows an easy overview of the game. When reporting on a game, he looks for phases with a *story*, *e. g.,* an interesting beginning that leads to an important outcome. He is also interested in comparing the objective facts such as theoretical positions, to what the players really did during the game. He generally completes his findings with online tools dedicated to soccer data exploration. His idea of an ideal system would be a set of statistical tools that would allow professional analysts to better explore the data, with analysts then being able to use their findings in the stories they tell their audiences.

The third expert is working for Opta, one of the world leaders in live sports data collection [13]. The company trains experts around the globe to collect detailed data for major championships. This data is fundamental for fans or professionals who want to generate analytics before games for preparation, during games to support live comments or betting, and after games for performance reports. This expert provides the company's clients, which include sports media, game sponsors, and teams, with data through tables and simple statistics. He communicates the data this way because the people who use them are not necessarily trained in reading visualizations, while acknowledging it would better value the full potential and complexity of the data.

The fourth expert—sports trainer—is a former professional athlete and has experience in coaching team sports, including soccer. His approach in analyzing a game is to first look at the instructions that were given by the coach (*e. g.,* team compositions, individual roles, etc.). He then compares the actual positions of the players on the field to the ones they were supposed to assume in the theoretical position. In his analysis he also considers the context of the game in terms of its overall strategy, such as when the team should play more on the defense side instead of making scoring goals a priority. During the interview, he often used drawings to get his points across, as he does when coaching a team.

While experts have different perspectives in their analyses, they share the following elements in their workflow:

1. **Telling Stories:** Their job is not to provide an exhaustive list of statistics about games or players, but to tell stories and to express findings. They are particularly interested in phases and actions they believe would make good stories.
2. **Statistics Against Bias:** They all admitted having certain biases either against or for particular players. However, as professionals, they go back to statistics for factual information.
3. **Time Is Precious:** They all work on tight time constraints, and need supportive tools (*e. g.,* as said journalist #2, *"most journalists are lazy, and they need to be assisted in their work.").*

As far as we know there is no coherent tool that attempts to support those requirements and fully takes advantage of the data.

3 Related Work

We collected a series of visualizations and data graphics covering current best practices for communicating information about soccer games, but also visualizations related to team sports in general.

Football drawing [18] is a hand drawn visualization that shows an overview of the game by representing the continuous movements of the ball as lines on the field (Figure 2). The density of the lines can reveal trends in how the game was played, with a game's image becoming a unique representation of it, much like a person's fingerprints. The data displayed is persistent (*i. e.* once the ball moves from one point to another, the line stays forever in the visualization), and attributes such as what kind of action made the ball move are not taken into account.

Table 1. Summary of Main Actions Which Can Occur in a Soccer *Phase*

Action	Description	Area	Illustrative example
Long ball	*Attempt to make a long distance pass via a cross*		
Turning the ball	*Multiple players pass the ball to one another making short passes*		
Square ball	*A pass between teammates laterally, across the field*		
Long Run	*A player runs a long distance with the ball*		
Cross	*Ball delivery from either side of the field across to the front of the goal*		
Corner	*Kick taken from a corner towards a player/group of players*		
Shot	*A player hits the ball towards the goal, making an attempt to score*		

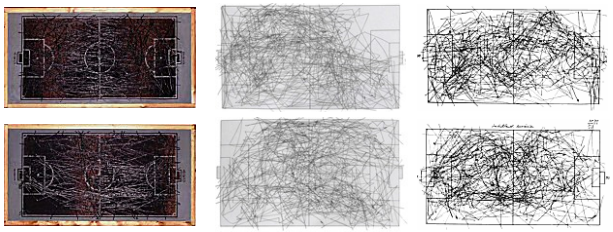

Fig. 2. Manual *football drawing* [18] of the continuous ball's movement seen from above. Image courtesy of Susken Rosenthal.

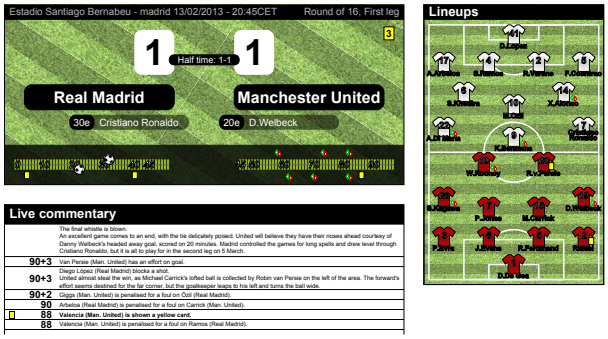

Fig. 3. Soccer game live cover [12]. Inspired by http://www.lequipe.fr.

Soccer team management simulators such as SEGA's Football Manager [9] provide automatically generated visualizations of entire games to help users make their decisions. These visualizations are very detailed, including features such as videos, statistics and textual transcripts. They are, however, based on simulated data, which contain more information about a game than a real dataset would.

When dealing with real data, *live covers* (Figure 3) are the most common approach, combining text and simple visualizations on websites that let the general public follow a game as it happens in almost real time. They can be used for different sports and are very similar, being comprised of multiple views with a timeline for the game overview and progress, team details on the sides, and a panel that textually shows events as they happen. Each text entry contains roughly a hundred characters, with an icon representing the action.

When following a soccer game it is important to understand how the different events that make up the game are distributed through time. Since games happen in predefined intervals of time it is also important to know at which point in time a game currently is (to know, for instance, how much time a team has left to score enough goals to win). An often used visualization that achieves that is the *timeline*.

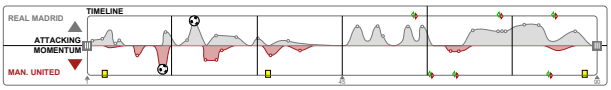

Fig. 4. UEFA Live cover timeline. Inspired by http://www.uefa.com.

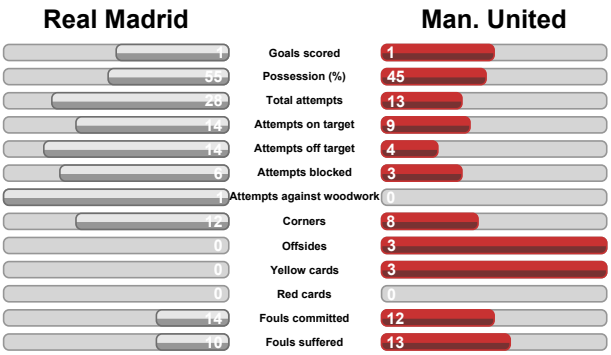

Fig. 5. Aggregated statistics as horizontal bar charts [24]. Inspired by http://www.uefa.com.

A timeline consists of mapping the interval the game takes place to a horizontal axis on which important events and the current point in time are indicated by icons and other markings. Icons tend to be analogous to real-life events (*e. g.,* a ball stands for a goal, a yellow rectangle stands for a yellow card). Figure 4 shows its use in a live cover. This timeline is also augmented with what is called the 'attacking momentum', a subjective measure expressing which team is dominating the game in terms of ball possession and goal attempts.

To understand a game it is also important to know a team's strategy and composition (which players assume which roles). This is usually done with a *team lineup* visualization, which consists of a representation of a soccer field on which icons standing for the players are placed according to their respective theoretical positions (*i. e.* roles). This visualization is typically used to introduce players on TV and in live covers (Figure 3).

Statistics are also effective at better understanding a game (*e. g.,* ball possession, number of fouls). All the different media (television, live covers, newspapers) communicate these, either by explicitly showing the numbers or through a simple visualization. For example, Figure 5 illustrates the standard way of representing the aggregated statistics of a game. However, the scale used for the width of each bar is different across statistics, and the color scale may not be well chosen.

A game shown on television might be too dynamic for an audience to be able to discern everything that is happening in real time; they might miss out on important events that can be revealing of a team's strategy. Experts, however, are trained to spot such important events

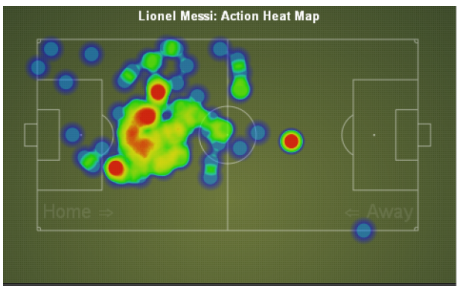

Fig. 6. Heatmap of Messi's position during a game. Image from http://chalkontheboots.wordpress.com/2012/09/, created on http://www.squawka.com/.

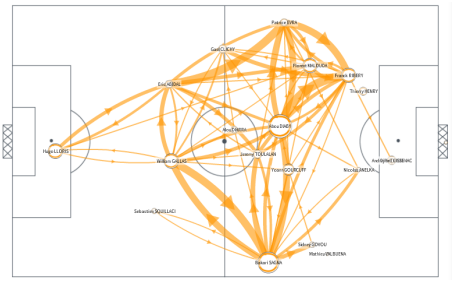

Fig. 7. Communication between players [5]. Image from http://www.footoscope.com/, courtesy of Fabien Girardin.

and their insight is used to produce augmented replays that are shown either during the break or after the game is over. An augmented replay consists of superimposing an animated diagram on a replay of the original video that is slowed down to emphasize particular moments while a narrator explains to the audience what can be seen. A static version of this approach is often used in blogs, with videos replaced by annotated images accompanied by text.

Deeper analysis of a game finds its way using non-soccer specific visualizations in the context of a soccer game. One such visualization is the *heatmap*, through which player's most frequent positions is displayed by density. For instance, Figure 6 reveals that *"Messi moves deeper now into a classic No. 10 position on the pitch and is more or less laterally aligned with the most advanced of Barcelonas midfielders."* [3]. Another frequently-used visualization is the *flow graph*, illustrated Figure 7, where the size of the nodes shows player's role in the game and the links show the connections between players.

Academia showed recent interests in sport visualizations [4, 14]. Soccer Scoop [19, 20] and MatchPad [11] use glyph-based visualizations respectively to compare soccer players, and to analyze performances during rugby games. CourtVision [6] and SnapShot [15], respectively designed for basketball and hockey, introduce specific types of heatmaps tailored to ball and puck shots. All those works acknowledge the important need, impact and potential of visualization systems applied to the characteristics of sport games and users.

4 Faceted Views for Soccer

In this section, we extend the current corpus of visualization of soccer we identified previously with standard visual representations. For each of the main actions in soccer (Table 1), we designed so-called *faceted views* based on the characteristics of each action and the literature in information visualization.

4.1 Corner Kicks and Crosses

Corner and crosses are short actions which originate from the side of the field, towards the goal. We represent corner and crosses by heatmaps, which show the part of the field where they start and end. The cropped and zoomed view on the right uses cyan and brown scales for the ball's position at the start and end of the action, respectively. An overlay shows the players involved and the path the ball follows from one to the other (Figure 8(a)).

4.2 Distribution of Shots

Shots towards the goal (Figure 8(b)) are the most important actions in soccer. Shots are represented as lines. We show their origins in the *top view* ((x_o, y_o) coordinates in the penalty area) and their destination in the *front view* ((y_d, z_d) coordinate, z_d being the height of the ball when it crosses the line beyond the goals). The different line colors stand for the different outcomes (missed shots, posts, saved shots and goals). The shot being analyzed is highlighted as a thicker line, so that it can be easily compared with the others.

The interactive legend shown on the bottom of the faceted view can be used to filter the shots by outcome. Spatial filtering can also be performed by brushing on either view (Figure 8(c)). Statistics (number of shots and number and percentage of each shot type) are updated according to the currently shown shots.

When too many shots are displayed at the same time, the view might get cluttered. To address this issue, the shots can be alternatively represented as heatmaps through the *spray* feature (Figure 8(d)). The shots' origins and destinations are encoded as circular color gradients centered at their positions, whose radii can be interactively set. Brushing and filtering remain unchanged.

4.3 Long Runs

Long runs occur when a player runs at least one fifth of the field's length while in possession of the ball. This action can happen either on the left side, the middle or the right side of the field, with the player's exact trajectory being of little importance. Long runs are shown as arrows across the field (Figure 8(e)). The thickness of each arrow indicates the frequency of a long run in a particular area by members of a given team (*i. e.* the more often a team's members perform long runs, the thicker the corresponding arrow will be). The arrow including the long run being currently analyzed is shown in gray.

4.4 Pass Clusters

The most common event in a game is a pass, and most of these passes are short in distance. Passes may happen in sequences, with a series of short passes called a *cluster*. In terms of effect, a cluster with the ball beginning from a player A and ending with a player B is roughly equivalent to a direct pass from A to B, making A and B the most important players in the cluster. However, it is sometimes interesting to know more specific information about a cluster, such as which players were involved in it and the order of the passes. Because clusters can assume many roles in a phase, we propose to visualize them in different ways and each faceted view associated to clusters has its pros and cons. For example, such tasks may depend on the spatial position of action, the chronological order of events, and the frequency at which a player appears within a series of passes.

In the full node-link diagram (Figure 8(g)), all players are shown as nodes and passes as links between them. Nodes are placed according to their respective player's theoretical position. If players are involved in the cluster, their respective nodes have a larger size, with the ones corresponding to the first and the last players even larger.

The compact node-link diagram (Figure 8(h)) is analogous to the full node-link diagram except that players not involved in a cluster are omitted, showing only the players who touch the ball during the action.

An adjacency matrix (Figure 8(i)) shows the players as both columns and rows. Each entry of the matrix shows how often the player in the row passes the ball to the player in the column, with a darker shade indicating a higher frequency of passes. Since a player cannot pass the ball to himself, the diagonal is used to show the player's identifying number. First and last players are shown in green and passes within the cluster are represented as dotted lines within the matrix. A cluster's typical matrix is very dark, indicating that the players involved often pass the ball to one another. A lighter matrix indicates a rarer cluster.

Hive Plot (Figure 8(j)) is a type of graph layout [10] where each axis is a player identified by its jersey number. A player touching

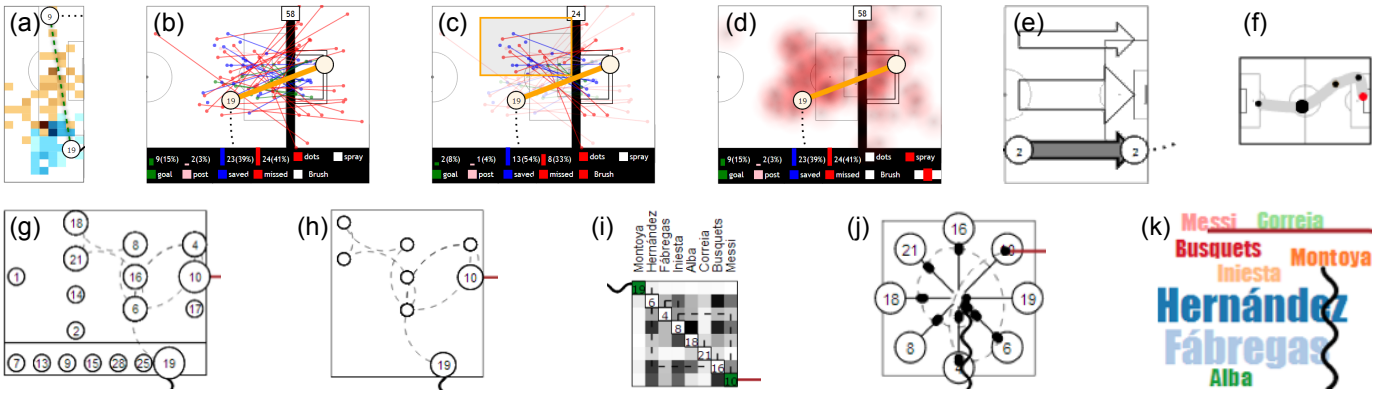

Fig. 8. Faceted views for soccer: (a) corner kicks and crosses, (b-d) shots, (e) long runs, (f) global flow, (g-k) different facets for pass clusters.

the ball is represented by a black dot and passes between players are links between these dots (from one axis, to another). The first pass of the cluster is represented by the closest black dot to the center of the Hive Plot and the last one by the furthest black dot from the center. If a player passes the ball several times, several dots are displayed on his corresponding axis. In this faceted view there is a trade-off of spatial information for the order of the passes in the cluster and this visualization is efficient to get the temporal order of a series of passes.

Very popular in many contexts other than soccer, the tag cloud (Figure 8(k)) depicts the names of the players involved in the cluster, with the font size being proportional to how often a player appears in the sequence. With a tag cloud visualization, estimating the importance of a player inside a pass cluster is immediate. Color and position may encode additional information, such as the players' theoretical position.

4.5 Global Flow of the Phase

A phase can be summarized by the line connecting the faceted views together. Each faceted view—corresponding to one or several actions—is aggregated into a point on this line and its size encodes the number of actions it contains (Figure 8(f)). As experts mentioned, only approximate positions of the actions are important and using the position of the faceted views instead of the positions of all actions helps users identify the *fingerprint* of the phase by preventing clutter.

4.6 Faceted Views Layout and Coordination

To support complex tasks, *i. e.* tasks involving multiple actions, several views previously introduced should be combined. However, we did not observe any particular consistency in our review of soccer visualization—apart from live coverage, dedicated to live data—for soccer views presentation. We experimented with several prototypes containing multiple coordinated views—where a selection or a focus in one view is propagated to others—that could be used to explore all facets of a soccer game (Figure 9). Journalist #1 found none of them suitable for his workflow. He added it was difficult to elaborate and validate hypotheses, as too many visualizations were available and he did not know where to start.

From our discussions with experts, we found that the soccer field is the primary object of observation and analysis in soccer. Analysts construct their mental model over the spatial arrangement of the team, and its motion, over time. In the next section, we introduce *SoccerStories* which uses the soccer field as a layout for faceted views to represent a series of actions.

5 SoccerStories

SoccerStories provides an overview+detail [21] interface of game phases, and their aggregation into a series of connected faceted views, where each faceted view is tailored for specific actions. The central compound of the interface is a soccer field—the zoom—and is surrounded by overviews and details panels.

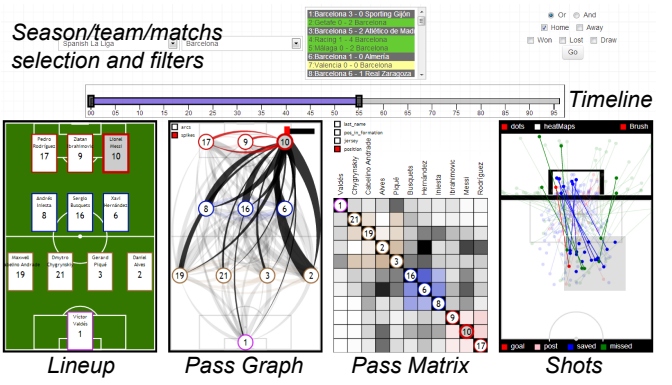

Fig. 9. An early prototype with synchronized faceted views in a grid layout, which was rejected by the experts.

5.1 Workflow

We designed *SoccerStories* advised by the four experts we collaborated with and around our existing collection of faceted views. The system works as follows (Figure 10):

- A game is picked from a list, and loaded in the interface (not visible in figure 10);
- A timeline and small multiples provide an overview of the game, to navigate into the phases of the game;
- The selected phase is displayed on the soccer field and is aggregated into a series of faceted views;
- Details are available on the side for selected players; the phase can be exported as word-sized graphics to embed into text.

5.2 Design Rationale

SoccerStories is built on the faceted views described in section 4. Because each faceted view focuses only on a particular action or group of actions, we propose to *connect* faceted views together to show the phase in its entirety. The faceted views are displayed on a representation of a soccer field on the central area of the screen—the *zoom*—around which are panels with a timeline and small multiples—the *overview* of the game—for a convenient navigation and a sidebar showing player statistics—the *details*—(Figure 10) as follows:

Soccer Field as Zoom. The central workspace is a soccer field, as experts often refer to it as the dominant way to display soccer data. A soccer phase is shown as a *node-link diagram* drawn over a soccer field, making this focus view a temporal zoom. Nodes represent players, with each node placed in the visualization in accordance to its respective player's position in the actual field at the moment of the represented action (Figure 1(b)). The visual encodings of nodes follow the way soccer tactics are displayed [25]. For example, a dashed line is a pass and a squiggly plain line is a player move (Figure 10.2).

Timeline as Overview. Phases are shown in chronological order

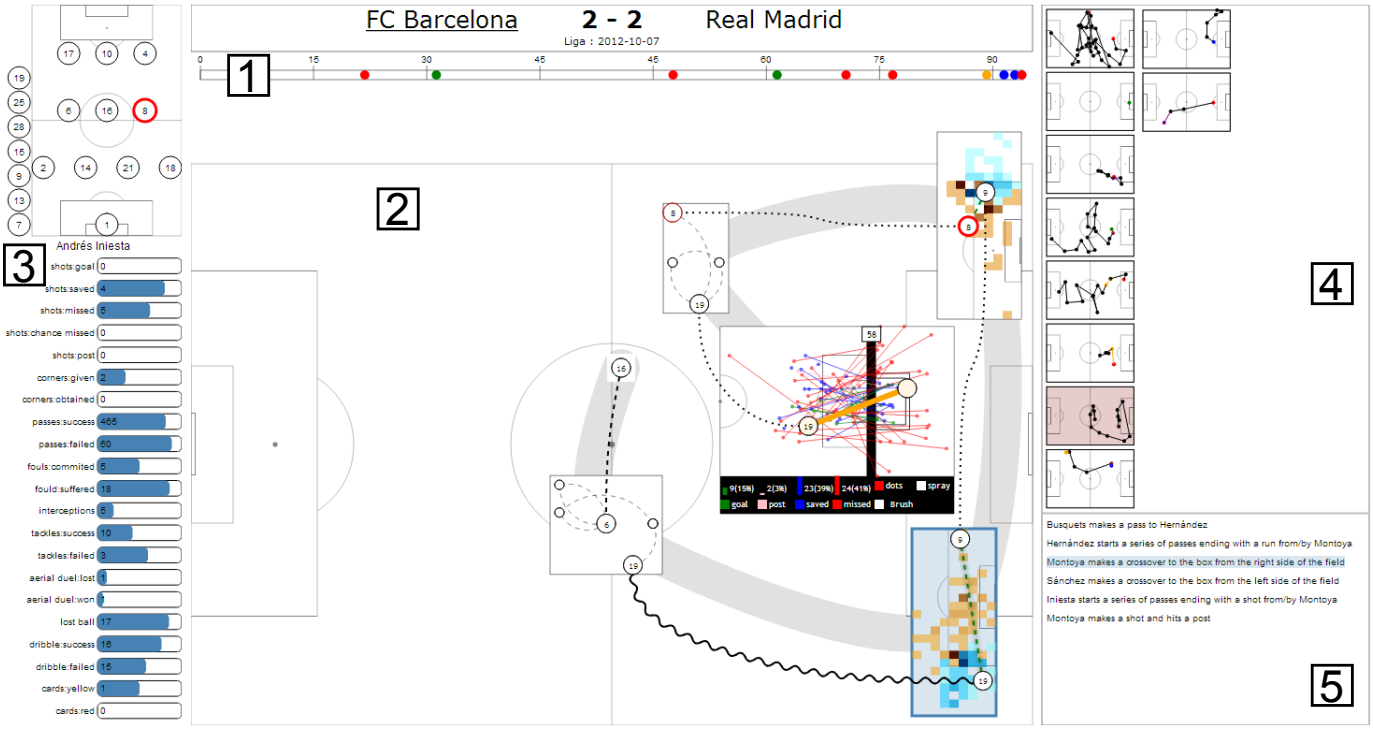

Fig. 10. *SoccerStories* user interface: (1) complete game overview as a timeline, (2) temporal zoom on a game phase and layout on a soccer field, (3) details on the side. After iterations, we added (4) the thumbnails and (5) generated text-annotations.

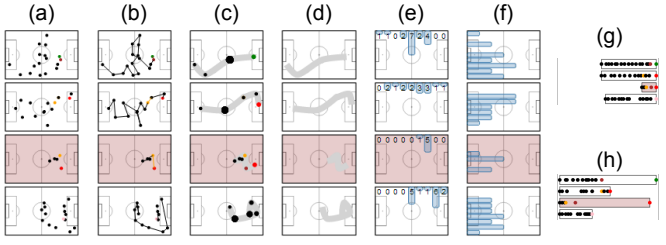

Fig. 11. Available small multiple views: (a) dots, (b) lines and dots, (c) flow and groups, (d) line, (e,f) x and y distribution histograms, (g) alignment by distance travelled, (h) alignment by duration.

on the timeline (Figure 10.1) as colored dots representing their outcomes, with red standing for missed shots, yellow for posts, blue for saved shots, and green for goals. When hovering over a dot, a thumbnail of the phase pops up below and displays the actions of the phase represented as nodes into a reduced soccer field as an overview. When clicking a dot, its corresponding nodes sequentially move from the pop-up window to their respective positions in the main soccer field representation, in chronological order.

Side Bar for Details-on-Demand. Detailed information about a selected player in the field panel is displayed on a sidebar (Figure 10.3). The side bar also contains the static visualization of the team using a soccer field as the layout for the lineup.

After iterations we added two additional interface features in the context of soccer data analysis:

1. **Phase Comparison:** The small multiples view (Figure 10.4) shows all the phases as small multiples, which provides an immediate overview of each phase and allows visual scanning and comparison of phases. Several small multiple views are available (*e. g.,* scaled down view of the focus view, line as a fingerprint of the phase, projection on the temporal axis or the distance the ball travelled). Each of these small multiple views is dedicated to a particular task (Figure 11).
2. **Automatic Text Generation:** Because analysts are used to text annotation (*e. g.,* in live covers) we generated very basic sentences containing the phases' main entities and actions (Figure 10.5). Example of these automatically generated texts are: `Messi makes a pass to Xavi, Xavi makes a cross to Iniesta.`

Before visually grouping actions into faceted views, a preliminary dataset preparation with extraction of phases is required (Section 5.3). We then describe the process of visual representation of these phases into connected faceted views (Section 5.4) and the associated animated transitions (Section 5.5).

5.3 Data Preparation and Phases Extraction

Data preparation involves two steps: first extracting the phases, then for each phase identifying groups of actions.

The extraction of phases is done by first detecting potential phase-ending events, such as shots, as in [8]. For each phase, contiguous preceding events are added until a phase-breaking event is reached (e.g., the ball leaves the field, the opposite team takes possession of it, etc.). A further filtering of the phases can be done according to many criteria, such as selecting phases ending with an interesting outcome. We chose to keep phases leading to shots, which are the most important events in a soccer game.

Based on our discussions with experts, we identified the set of events leading to a shot or scoring opportunity. This results in the following classification, where **F** indicates events which can be the first event of a phase; **f** events which can be first if the only event in the phase; **M**, the ones which can occur during the phase; and **L**, the ones which end the phase, *i. e.* the possible last events of the phase: **F:** interception, clearance, corner; **FM:** pass, take on, good skill, aerial duel, tackle, free kick; **fML:** shot:post, shot:saved; **fL:** shot:goal, shot:missed.

Other events are either events breaking the phase (*e. g.,* the opposite team gets the ball, the ball goes out) either ignored events (*e. g.,* an opponent missed a tackle or unsuccessfully tried to get the ball). We consider that this decomposition is standard, except for very particular

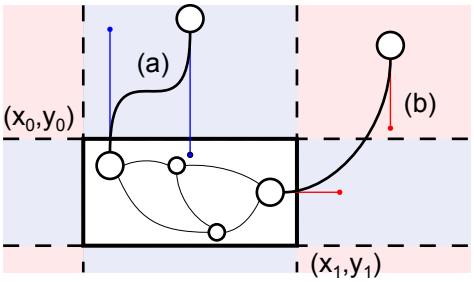

Fig. 12. Cubic Bezier Links drawing between faceted views. According to source and target nodes positions, control points (shown in color) are built (a) parallel or (b) perpendicular.

cases, for example when the ball goes several times and quickly from one team, to another.

The grouping of actions is algorithmically computed based on action attributes and following a set of soccer-specific heuristics. Each group is composed of 1 to n actions, resulting in 1 to n nodes and 0 to $n-1$ links grouped together. This naive approach may not lead to the optimal grouping, but was found satisfying by the experts.

5.4 Visualizing Facets and Phases

We now describe the process of representing a phase using standard soccer visualizations to represent faceted views of actions. *SoccerStories* displays each faceted view as a small box containing the node-link representation of the actions it contains, which is stylized according to the faceted view characteristics.

When the user clicks the "Cluster" button, the nodes and links of the focus phase are animated with respect to temporal order, transforming the phase into a series of faceted views connected by their first and last actions. The connections encode the type of action between the faceted views. For instance, an aerial pass is shadowed, a pass is dashed, a player moving with the ball is squiggly. To reduce overlap and facilitate the phase reading, links are cubic splines between the source and target nodes. The link leaves the faceted view almost perpendicularly and from the side of the faceted view where overlap is the lower. The path of each link is computed using (x_0,y_0) and (x_1,y_1), the upper-left and bottom-right corners of the faceted view. If the target node's x coordinate (respectively y) is in the range $[x_0,x_1]$ (respectively $[y_0,y_1]$), the bezier curve is a sigmoid (Figure 12(a), blue areas) with two parallel control points. If the target node's position is elsewhere, the control points are perpendicular (Figure 12(b), red areas). We used Bezier curves because of the visual abstraction they provide. Indeed, a straight line might be interpreted as the real trajectory of a player or of the ball. Instead, we use Bezier curves that look stylized and less realistic and have been extensively used in graph drawing [7, 17].

A particular type of link occurs when a node belonging to two groups is duplicated and the duplication is represented by a dotted link between the two faceted views. For example, when a player receives a cross ball and makes an immediate shot, he belongs both to the faceted view of the cross and the faceted view of the shot. Then, the node representing the player is duplicated and a dotted link is created between the two nodes, one belonging to each of the two faceted views.

5.5 Animated Transitions

When the user requests to group the focus phase, a four-step process preserves the continuity between actions and their grouping into faceted views (Figure 13): **T1** A convex hull visually groups the nodes. Nodes move from their initial position to their position in the faceted view being created. The links and the hull have their shapes updated as the nodes move. If a node belongs to two faceted views, it is duplicated and a dotted line links the two nodes; **T2** The convex hull takes the shape of the faceted view; **T3** The nodes and the links in the group are stylized according to the faceted view needs; **T4** The hull fades out while the faceted view fades in.

When a phase has been converted into faceted views, the user can select a faceted view and switch from it, to another, for the same group of nodes. The order of staged transitions for switching between faceted views is: **T1** the faceted view fades out and only the nodes and links remain visible; **T2** the nodes move from their position in the current faceted view to their position in the new faceted view and the links are updated accordingly; **T3** The nodes and the links in the group are stylized according to the new faceted view needs; **T4** The new faceted view fades in.

5.6 Implementation

SoccerStories is implemented using *D3* [1] and JQuery [16]. It runs on any modern web browser and dynamically loads `JSON` data processed on the server (Section 5.3) by querying a database. These queries return pre-computed phases and statistics (for the details panel). This way it is very fast to process queries and transfer the resulting data to the visual application.

One important implementation design choice was to enable an unlimited number of faceted views. As we designed the generic visual canvas described in Section 5.4, we also support their integration in a generic way to allow the future implementation of faceted views. Adding a new faceted view consists of taking as input the group of actions and a position. Some functions are handled by the super class (*e. g.,* `drag()`) while others remain very specific to the new faceted view (*e. g.,* `drawNodes()`, `drawLinks()`, `drawContext()`). The result is an `SVG` element that is displayed on top of the soccer field's element due to its higher level in the `SVG` scene graph.

The dataset we worked with was provided by Opta Sports [13]. It consists of `xml` files with entries for actions and their qualifiers, with a typical game consisting of about 1000 entries. There is an entry for each action, with subfields indicating what, when, where and how it happened and which player was its protagonist. Additional qualifiers describe the ways a player hits the ball. The datasets do not contain the position of players at all times, only when they are actively involved in an action (*e. g.,* hitting the ball or passing it to someone else).

Both the extraction of phases and the grouping of actions are predetermined in our implementation; the experts found our naive approach satisfying, probably because we selected a particular outcome for the phase. However, manual interaction to adjust the computation of phases as well as the set of rules applied for the grouping of actions offers promising perspective to give the user more freedom.

6 Evaluation

We presented *SoccerStories* to the four experts with whom we had collaborated (Section 2.3), for a one hour remote interview. They all found the layout of the interface very easy to understand and use. Even if they were not aware of the phase sequencing process, they knew a choice had been made and that phases were divided based on each team's periods of ball possession.

From their feedback we designed an experiment consisting of including *SoccerStories* into journalist #1's (our tactics analyst) workflow. We hypothesized it would assist him in gaining insight from the games and writing articles. He successfully used *SoccerStories* to write and illustrate four articles.

6.1 Experiment 1: Writing Articles

The first experiment was conducted to validate our hypothesis that *SoccerStories* can help strategy experts better illustrate their insights about a game. The participant was journalist #1, an expert in soccer tactics analysis. The experiment consisted of inserting *SoccerStories* in the analyst's workflow as he analyzed two games he had recently watched but not worked on (namely, the UEFA Champion's League games played between Real Madrid and Manchester United on 13 February and 5 March 2013). Aside from the addition of *SoccerStories*, the analyst's work environment was not altered (i.e., he was not limited to using our system, and also had access to the web and to game replays). The session lasted two hours, throughout which one of the *SoccerStories* designers was present to provide technical assistance.

During the session, the analyst began his work as he usually does, searching the web for other articles about the games. However, he very quickly switched to *SoccerStories*, which was from then on the

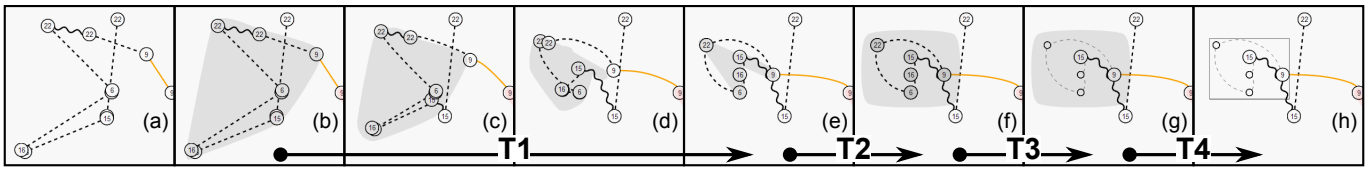

Fig. 13. Four-steps staged animation from a group of nodes, to a reduced node-link visualization of pass cluster.

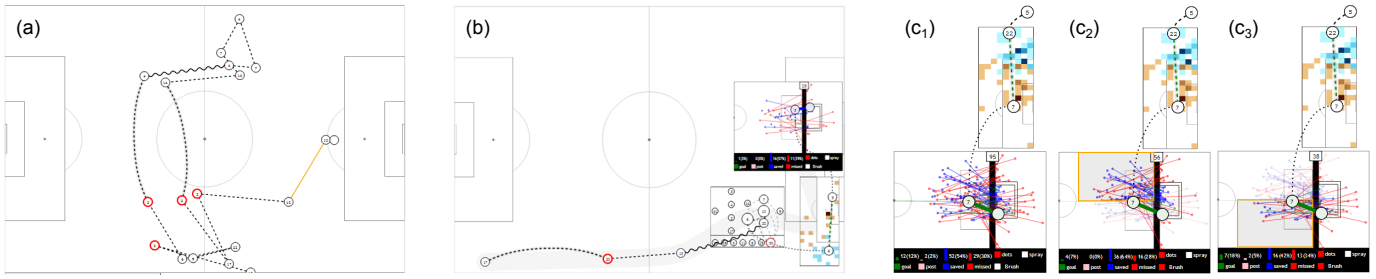

Fig. 14. Screenshots from the analyst for Real Madrid vs Manchester United game: (a) Varanne's actions (highlighted in red), (b) Higuain's role (highlighted in red), (c) information retrieval from the faceted view of shots: (c_1) all shots, (c_2) shots from the left side of the field, (c_3) from the right side of the field are scored goals for 12%, 7%, and 18%, respectively.

only tool he used to explore the games (the games were still fresh in his memory, so he did not feel a need to use the replays).

He started his use of *SoccerStories* by first rapidly navigating over all of the games' phases to get an overview. He then moved on to detailed views of the phases he found most interesting. This process resulted in four articles in which he used screen captures of *SoccerStories* to illustrate his text.

"The Offensive Defender": He began the first article during his initial exploration of all the phases when he was surprised to see that Real Madrid's defender Varanne (number 2) was, despite his nominal role, active in many offensive phases of the first game. To illustrate this, he selected Varanne to highlight his actions and took the screenshot shown in Figure 14(a). Proceeding with his analysis, he found out that this player was much less involved in offensive phases in the second game. He also compared Varanne's statistics in both games, which showed that the player made much more passes (48) in the second game than in the first (33). Based on what he found out with *SoccerStories* and his previous knowledge, he deduced that Varanne (and to some extent the whole Real Madrid team) performed this way due to the location of the games: when not playing at home they preferred to wait for the other team to make a risky move and then counter-attack.

"Rewarded Coaching": The analyst began working on his second article when he inspected the small multiples for the second game and saw that phases during its last third contained significantly more actions. This immediately reminded him that during this game Manchester United's player Nani received a red card around the 60th minute and that Real Madrid's coach opportunistically substituted defensive player Arbeloa for the offensive Modric. He explored the phases following the substitution and noticed that Modric became very present in both phases (6 of a total of 11 remaining phases) and actions (an average of 3.5 actions per phase). Modric also scored the goal that tied the game and was key to his team's second goal, which guaranteed its victory. Although no images of *SoccerStories* were used in the article, the story was based on what was discovered with it.

"Beyond Personal Bias for Players": In his third article, the participant wrote about two strikers of Real Madrid's team, Benzema and Higuain. Knowing that he is biased in favor of one of them, he used *SoccerStories* to balance this out by objectively analyzing the games. Based on what he saw by exploring the phases in *SoccerStories*, he was forced to admit that in the first game his least preferred player (Higuain) performed better than his favorite one (Benzema). In particular, he thinks that Higuain plays too much on the center of the field but for this game, he realized that this player made many actions on the right side of the field. He illustrated this with the screenshot shown in Figure 14(b). In the second game, though, he saw the players go back to what he considers their normal behavior, writing about his least preferred player: *"Higuain becomes himself again"*.

"Shoot More from the Right?": In his fourth and last article, journalist #1 described a team predilection for shooting from the left side of the field and how it surprises their opponents when they shoot from the right. He first noticed this when using the cross visualization heatmap to examine a phase that ended in a goal Figure 14(c_1). Filtering shots of this phase by origin and using heatmap brushing confirmed this as he observed that the team had only 7% of goals scored by shot from the left side of the field Figure 14(c_2) against 18% from the right side Figure 14(c_3). He also guessed that this unbalance happened due to the opponent's awareness of the team's predilection for one side.

Overall Observations: Journalist #1 wrote four stories based on a tactical analysis of the games. He created stories about facts that he did not know in advance, that he wanted to check, or that challenged his initial knowledge. On the other hand, even if he intentionally used complex and unusual visualizations—from his point of view—to achieve his analyses, he only used standard visualizations to illustrate his findings and communicate the story. He said *"My readers are not ready for such complex visualizations"*, raising the issue of his audience's lack of visual literacy. The interviews with the data expert from Opta confirm this. He clearly needed well-known visualizations as landmarks, such as the statistics on the left side. We realized that the acceptance of a new tool needs to include familiar visual anchors to make the tool appealing and consistent with the user's knowledge.

6.2 Experiment 2: Follow Up

The four experts heavily relied upon small multiples representing phases for soccer games overview, both for analysis and communication. We evaluated in a follow-up study different visual encoding of the phases' small multiples views in order to assess their design. We hypothesized that they could be used in articles similarly as sparklines [23]. The result is a ranking in a series of design variations, and a set recommendations for their export into word-sized graphics that we call *sportlines*. The study consisted of a 15-minute online experiment split into two parts: one for small multiples and one for export. We asked the four experts as well as soccer fans to complete the experiment; 13 people participated in total.

In the first part of the experiment we evaluated four representations of phases available in *SoccerStories* (Figure 15): (a) dots; (b) lines connecting dots; (c) flow visualization with a line and groups of ac-

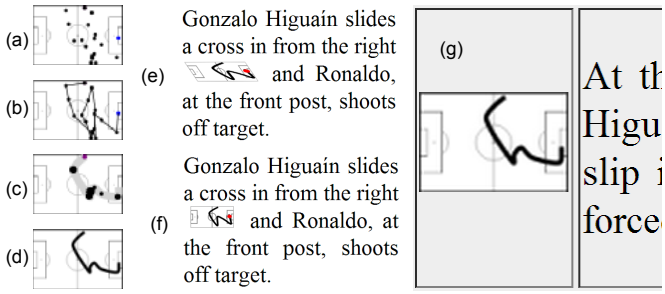

Fig. 15. (a-d) Small multiple views of a phase; (e,f) *SportLines* embedded in text; (g) small multiple in a live cover format.

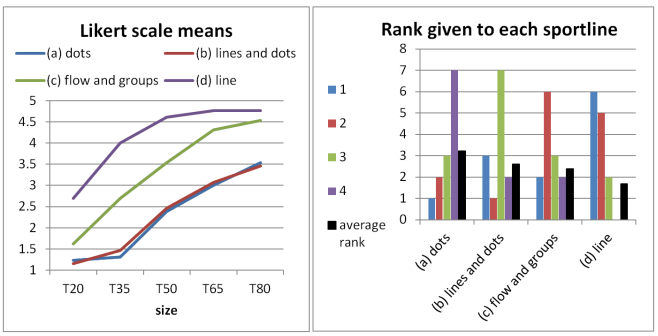

Fig. 16. Results for the evaluated representations for small multiples.

tions; and (d) a single line. Each *Representation* of small multiples had five different *Sizes*, in order to assess the potential of embedding small multiples into text: 80×60, 65×45, 50×37, 35×26 and 20×15px, the last one being close to the standard height of word-sized graphics (1em). Three phases were displayed for each *Representation* $\times$ *Size* combination, and were previously explained with a 400×270px image of the phase mapped on a soccer field as well as a text explanation.

Participants were asked to complete a questionnaire and evaluate each combination of *Representation* $\times$ *Size* on a Likert scale ranging from 1 to 5 (1 being the worst, 5 being the best). They were finally asked to rank the four different representations by order of preference, and provide some qualitative feedback in a plain text field. Figure 16 shows the results for this first part of the experiment. The tendencies show that the line is more robust to small scales, followed by flow and groups. The surprising result is that even as sizes increase, the ordering remains unchanged. The rankings reported by participants confirm that the two last representations are preferred, with the line representation being the favorite one.

In the second part of the experiment, we evaluated four different *Design* alternatives: a 400×270px image of the phase mapped on a soccer field (similar as those in in *SoccerStories*); a 65×45 miniature (*i. e.* close to the format used in live covers) of the phase with single line (Figure 15(g)), but colored in black; a 15px height miniature by scaling the soccer field to have a 30px width; a 15px height but with 3D perspective allowing a 45px width without distortion. The last two conditions were embedded with text explanation (Figure 1(e,f)) and their design was guided by Tufte's sparklines [23] and other word-sized graphics [2]. They consisted of a black line only, representing the aggregated flow of the phase, and a red point to highlight the last action of the phase. However, we kept the field landmarks to remind the spatio-temporal nature of the data.

We asked the participants how they consider the design variations adapted to communication in general, for integration in articles and to insert in game reports. They were also asked to rank the four designs. All designs received a 73% positive answer for communication. The integration in articles and game reports received mitigated answers, both obtaining approval rates close to 50%. The interesting result is that the live cover format was the preferred one, and that the non deformed version of the embedded sportlines was preferred to the 3D perspective one. Interestingly, while the embedded sportlines were ranked last overall—this may be due to their text-size, and is an asset that requires further design exploration—the participants commented that they provide an immediate overview of the phase.

7 Discussion and Conclusion

Our first experiment validated two aspects of *SoccerStories*. First, that with minimal training, an expert managed to find novel insights in soccer data by browsing games by phases. *SoccerStories* enabled him to support his existing workflow (*i. e.* relying on his intuition after watching a game), but also gave him a novel way to explore games, challenging his natural biases. Indeed, browsing games by representative phases helped him quickly find and compare phases, make quick hypotheses, and test them through iterative visual analysis. It also saved him time, a critical factor when writing articles. His output, as a series of four articles, are articles he would not have been able to write otherwise due to a lack of supporting information. Finally, *SoccerStories* successfully improved the expert's analysis, even if the analyst felt that the communication of his findings inserted in the text with small visualizations was not well-suited for his readers.

In the exploratory follow-up study, we investigated the design of word-sized exports of phase visualizations into text. The study confirmed that phases are an appropriate selection of time intervals, and that they implicitly conveyed a meaningful story, even when reduced to an aggregated trajectory of the ball without details about the actions. Based on the previous experiments, we conclude the following:

1. Phases are adequate for soccer exploration, as they are well understood, easy to extract, and convey more information than single events. They are the semantic level for browsing, analyzing and communicating soccer stories; the data provider expert was particularly enthusiastic about this level of abstraction, saying: 'You are definitely going in the right direction';
2. The soccer field is the workspace for understanding the phases and for communication;
3. Using advanced visualizations in a domain such as soccer requires particular care although their standardization is mandatory to increase the visual literacy of both sport analysts and their audience. However, some word-sized visualizations seem to be well-suited as they are easy to understand.
4. Spatio-temporal thumbnails or fingerprints are a promising alternative for timelines, but their design still needs further investigation as they are a compact representation.

Based on the feedback from experts, future work includes the manual selection of phases to explore; manual interaction to refine the automatic grouping within a phase; the synchronization between the interface and the video of the game; the adaptation of *SoccerStories* to live data streams; and a deeper exploration of *sportlines* design.

We consider *SoccerStories* as a *kick-off* for soccer analysis. As far as we know, using visualization to help analyze and communicate on soccer tactics is novel and our four experts acknowledged its usefulness and effectiveness. Even if some of the visualizations are deemed too complex by experts for communication purpose (due to the general public's lack of visual literacy), we have proposed a system to analyze soccer tactics beyond the standard statistics on players and teams. In addition to helping journalist write articles faster, we hope *SoccerStories* will add depth to readers' experience of soccer analysis by shortening the textual description of actions and offering compact yet expressive standard visualizations that support insightful narratives.

8 Acknowledgments

We are very thankful for Opta's involvement in the project, especially for the data they provided. We are also very grateful to the four experts we worked with. We thank Fanny Chevalier, Clément Leurent, and Jeremy Boy for their feedback during the project. We finally thank André Spritzer and Nadia Boukhelifa for their help proofreading the document, and Lora Oehlberg for her amazing voice-over in the video.

REFERENCES

[1] M. Bostock, V. Ogievetsky, and J. Heer. D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, Dec. 2011.

[2] U. Brandes and B. Nick. Asymmetric relations in longitudinal social networks. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2283–2290, Dec. 2011.

[3] E. C. Campista. Sevilla 2-3 Barcelona: Tactical Analysis., 2012. http://www.elcentrocampista.com/2012/10/questions-remain-for-barca-despite-perfect-start-sevilla-2-3-barcelona-tactical-analysis/.

[4] A. Cox and J. Stasko. Sportsvis: Discovering meaning in sports statistics through information visualization. In *Compendium of Symposium on Information Visualization*, pages 114–115, 2006.

[5] Footoscope. FIFA World Cup South Africa. http://www.footoscope.com/worldcup2010/.

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

[7] N. Henry, J.-D. Fekete, and M. J. McGuffin. Nodetrix: a hybrid visualization of social networks. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1302–1309, Nov. 2007.

[8] S. Hirano and S. Tsumoto. A clustering method for spatio-temporal data and its application to soccer game records. In *Proceedings of the 10th international conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing - Volume Part I*, RSFDGrC'05, pages 612–621, 2005.

[9] S. Interactive. Football Manager. http://www.footballmanager.com/.

[10] M. Krzywinski, I. Birol, S. J. M. Jones, and M. A. Marra. Hive plotsrational approach to visualizing networks. *Briefings in Bioinformatics*, 13(5):627–644, Sept. 2012.

[11] P. A. Legg, D. H. S. 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. *Comp. Graph. Forum*, 31(3pt4):1255–1264, June 2012.

[12] L'Equipe. http://www.lequipe.fr/.

[13] Opta. Last access: March 2013. http://www.optasports.com/.

[14] M. Page and A. V. Moere. Towards classifying visualization in team sports. In *Proceedings of the International Conference on Computer Graphics, Imaging and Visualisation*, CGIV '06, pages 24–29, 2006.

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

[16] J. Resig. JQuery. http://jquery.com/.

[17] N. H. Riche, T. Dwyer, B. Lee, and S. Carpendale. Exploring the design space of interactive link curvature in network diagrams. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, pages 506–513, 2012.

[18] S. Rosenthal. Football Drawings. http://www.susken-rosenthal.de/fussballbilder/indexen.html.

[19] A. Rusu, D. Stoica, and E. Burns. Analyzing soccer goalkeeper performance using a metaphor-based visualization. In *Proceedings of the 2011 15th International Conference on Information Visualisation*, IV '11, pages 194–199, 2011.

[20] 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, 2010.

[21] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, pages 336–, 1996.

[22] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, BELIV '06, pages 1–7, 2006.

[23] E. R. Tufte. *Beautiful evidence*. Graphics Press, Cheshire (Conn.), 2006.

[24] UEFA. UEFA Champion's League Live Text Coverage. http://www.uefa.com/uefachampionsleague/season=2013/matches/round=2000348/match=2009591/postmatch/statistics/index.html#1/2013/2000348/2009591/pitch-view.

[25] Upward. Coach playbook, 2009. http://www.upward.org/uploadedFiles/Coaches_and_Referees/SOL-CoachPlaybook-08-09.pdf.

[26] ZonalMarking., 2013. http://www.zonalmarking.net/.