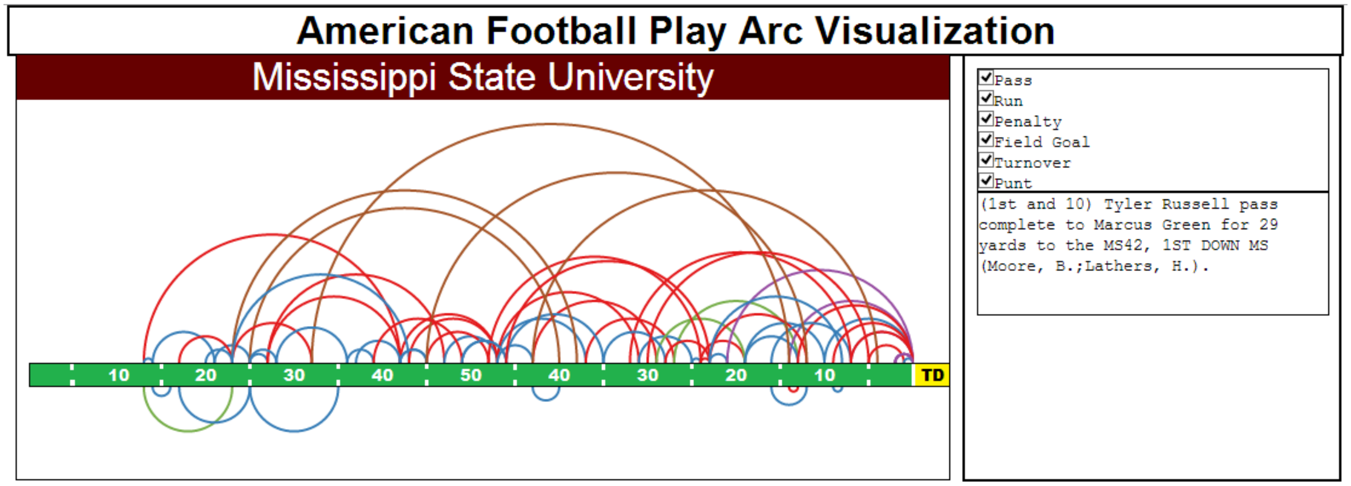
Visualizations for Exploration of American Football Season and Play Data

Sean Gabriel Owens and T.J. Jankun-Kelly, *Senior Member, IEEE*



American Football Play Arc Visualization

Mississippi State University

☑Pass
☑Run
☑Penalty
☑Field Goal
☑Turnover
☑Punt

```
(1st and 10) Tyler Russell pass
complete to Marcus Green for 29
yards to the MS42, 1ST DOWN MS
(Moore, B.;Lathers, H.).
```

Fig. 1. Overview of the game summary arc visualization. Each arc represents a different play completed by the team during the game. Arcs are colored based on the type of play that was run. These play types can be filtered using the selection box to the right. Hovering over an arc will provide extra information about that play in the play details box in the lower right of the visualization.

Abstract— Sports visualization is an emerging area of research. As the sports' industries are quickly rising into the billions of dollars in value, the need for performance evaluation is great. While, many sports such as baseball and soccer have developed fairly extensive visualization platforms for viewing player and team performance, one sport that has been lacking in meaningful visualizations is American football. The sport poses a challenge because its field only has one meaningful axis, therefore, spatial mapping of data onto a 2-dimensional field is both invalid and misleading. This paper presents two different visualizations for use in the analysis of American collegiate football data. The first provides an analysis of season-long data on a parallel coordinates chart, and the second presents a novel method of mapping football's 1-dimensional system using an arc diagram.

Index Terms—Visualization, american collegiate football, sports visualization, arc diagram, spatial mapping

1 Introduction

Although the field of sports visualization as a whole is becoming a much larger area of interest, there are very few visualizations that specifically address the sport of American football. Almost all of the visualizations currently in use display live, single-game data which is inadequate when looking for performance trends that only manifest over full seasons. One possible reason for this is the sport's lack of a meaningful two-dimensional mapping. Because the sport is played in a non-continuous manner, each team's position is discretely maintained play-by-play along the field's longer dimension. During this process, however, each team's position along the shorter dimension at the start of each play is irrelevant. This one-dimensional bias creates a difficult environment to map data onto.

American football is a very lucrative industry. Even at the collegiate level, many programs are worth over $100 million [3]. As these elite programs win more prestigious games, they are awarded larger payouts and endorsements. This increase in worth translates into better facilities and ultimately better recruiting, which in turn leads to more wins in the future. Therefore, to remain a viable contender in the league, any legal competitive advantage is crucial.

A key, untapped area that can provide this type of advantage is analysis through interactive data visualization. This paper presents two methods for analyzing the sport of American football. The first is a visualization that utilizes parallel coordinates to provide a visual overview of an entire season. The second is a novel method for viewing American football game data using arc structures spatially mapped to a one-dimensional grid. This method could potentially provide a new perspective from which to analyze opponents' play type trends.

These visualizations target any football analysts who would like to explore the collegiate American football domain. In this case, the term analyst can be used to define different levels of users. On an amateur level, there are average fans and amateur analysts who wish to track a team's performance over time or participate in statistical games such as fantasy football or sports betting. On a professional level, the main target users are professional analysts who can work for independent firms, independent team's, or league administrators. Other users on the professional level are coaches and trainers who can use the visualization to analyze their or their opponents' game plan.

- *Mr. Owens and Dr. Jankun-Kelly are with the Department of Computer Science and Engineering, Mississippi State University*
- *Email: owens.sean@outlook.com,tjk@acm.org*

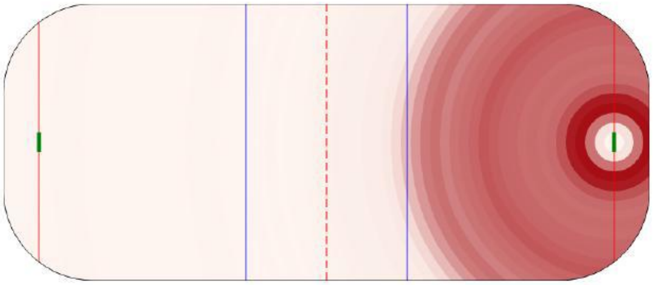


Fig. 2. Hockey rink used as a spatial substrate. Used in the Snapshot visualization [4].

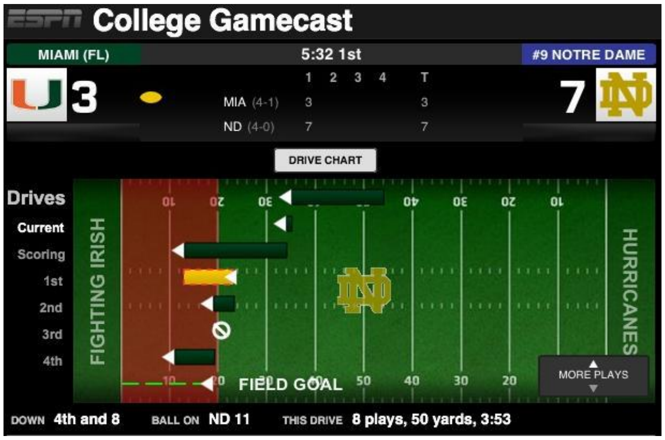


Fig. 3. ESPN's Gamecast Visualization showing a single drive mapped to a virtual football field. The implied movement from top to bottom is both inaccurate and misleading [10].

2 Related Work

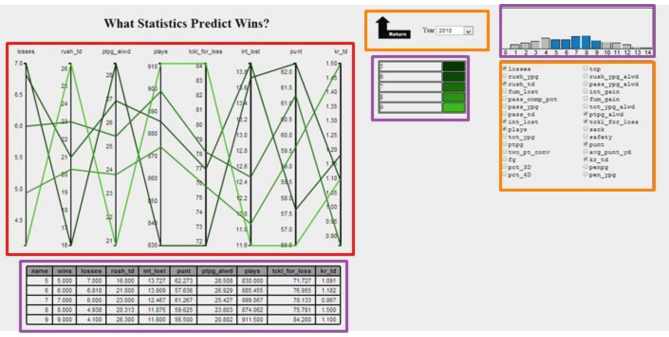
While sports data has been recorded for a long time, the widespread use of that data to analyze performance is a relatively recent development. Initially, statistics were analyzed to determine the potential performance of individual athletes. A special field of mathematics, Sabermetrics, was developed to analyze baseball performance [5]. Now, the area of sports analytics is rapidly advancing. Recently, information visualization techniques have been developed to aid in visual analysis. Initially, scatter plots and bar charts were used to analyze potential correlations, such as the baseball analysis by Cox and Stasko [1]. More recently, there has been a push towards visualizations that map data onto a meaningful spatial substrate. An example of this kind of visualization is the Snapshot hockey visualization presented by Pileggi et al [4] and shown in Figure 2.

This trend, however, cannot be easily translated to American football. As mentioned previously the difference between American football and other sports such as hockey, football (known as soccer in the United States), and basketball, is that American football only has one important spatial dimension which has led to a deficit of American football visualizations that use the field as a spatial substrate. An example of the most common football visualization is shown in Figure 3. It attempts a spatial mapping to the sports playing surface but implies movement in a direction (down) that is irrelevant to the sport.

3 Season Overview Visualization

In the next two sections, two visualizations will be presented. The first allows the user to view an entire seasons worth of data. The second, in Section 4, lets the user focus on the data from a single game.

The season overview visualization utilizes the standard parallel coordinates system introduced by Inselberg [2] and implemented using as a web-based visualization using the d3 javascript library present by Bostock [6] and found at [9]. It can be viewed online at [13].



☑ losses
☐ rush_ypg
☑ rush_td
☐ fum_lost
☐ pass_comp_pct
☐ pass_ypg
☐ pass_td
☑ int_lost
☑ plays
☐ tot_ypg
☐ ptpg
☐ two_pt_conv
☐ fg
☐ pct_3D
☐ pct_4D
☐ top
☐ rush_ypg_alwd
☐ pass_ypg_alwd
☐ int_gain
☐ fum_gain
☐ tot_ypg_alwd
☑ ptpg_alwd
☑ tckl_for_loss
☐ sack
☐ safety
☑ punt
☐ avg_punt_yd
☑ kr_td
☐ penpg
☐ pen_ypg

name	wins	losses	rush_td	int_lost	punt	ptpg_alwd	plays	tckl_for_loss	kr_td
5	5.000	7.000	16.000	13.727	62.273	28.508	830.000	71.727	1.091
6	6.000	6.818	21.000	13.909	57.636	26.929	885.455	76.955	1.182
7	7.000	6.000	23.000	12.467	61.267	25.427	899.067	78.133	0.867
8	8.000	4.938	20.313	11.875	59.625	23.803	874.062	75.781	1.500
9	9.000	4.100	26.300	11.600	56.500	20.802	911.500	84.200	1.100

Fig. 4. Overall view of the season overview visualization with the different usage area highlighted in different colors. The red box represents the main display area. The orange boxes contain selector objects including the categories and year selectors and the level of detail return button. The purple boxes contain legend and data information.

3.1 Data

The data used for this visualization is season-long football data for every team in the NCAA FBS division (the highest level of collegiate competition) over 31 data categories. A list of these categories is provided in the Appendix. While there are possibly more categories that could be included in this set, these 31 were chosen to provide a demonstrative range of categories over the whole domain.

The data used in this visualization was compiled by the National Collegiate Athletic Association (NCAA) and can be found at the NCAA's website [14]. The data in this set ranges from 2005 through 2011. This range includes all seasons that were completed at the time of data collection as well as had valid data for every team for all of the 31 categories.

3.2 Visualization Usage

Three main usage cases were defined for this visualization. The first user task is to observe which data categories correlate with higher number of win paths at the top of the axis and lower win paths at the bottom. If the data appears opposite of this, it is possible that a particular category is inversely correlated with success. The second usage case is comparing n-win teams against m-win teams. This case allows the user to determine how correlated a category may be with number of wins by showing the chart at a higher level of detail. The final case is comparing a team to others. This task would appeal to fans of a specific team or an analyst who is looking for a particular pattern for a single subject.

3.3 Visualization Design

This visualization has three main parts as shown in Figure 4: the display (outlined in red), the selectors (orange), and the legends (purple). The display area contains the parallel coordinates chart. Each of the 31 categories can be represented as an axis on the chart and individual teams are represented as paths across it. For this visualization, the user is comparing teams based on their number of wins. Therefore, an additional visual variable was needed to indicate this value. Each path in the chart is colored based on their number of wins using the color scale shown in Figure 5. The scale is linear along the chroma-lightness plane in the HCL color space to maintain cohesion across three levels of detail. Less wins are mapped to the darker/desaturated end of the scale while more wins are maped to the lighter/saturated end. Because the visualization has to distinguish up to fifteen different interval values, a green hue was chosen for its high response. Because there are over 100 teams in the data set, including all of them initially creates a chart that is too

Fig. 5. The color scale used to represent number of wins in the season overview visualization.

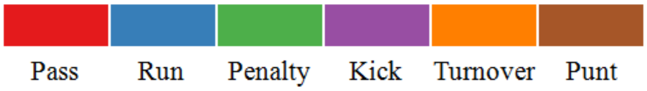


Fig. 6. The qualitative color scale used to color the arcs in the game summary visualization.

noisy to utilize. Therefore, the teams were grouped by their number of wins and the ability to zoom in and out based on this number included. There are three levels of detail. The highest level gives three groupings: 0-4, 5-9, and 10+ wins. Clicking on any one of these paths will expand into the second level of detail. At this level, each path represents a single number of wins (e.g. for the 0-4 group, a path would be drawn for 0, for 1, for 2, etc.). Clicking on one of these paths will load the final level of detail. At this level, each path represents a single team whose win value equals the path chosen in the previous level (e.g. all teams with 6 wins).

The second set of areas are the selectors. These allow the user to control what data is displayed on the chart. The return button allows the user to return to a higher range of wins. The year selector allows the user to choose which season data to view between 2005 and 2011. The third selector area is the category selector. This set of checkboxes allows the user to customize the subset of data categories to analyze on the parallel coordinates chart.

The final set of areas are the legends. These areas aid the user in determining what data is being shown on the chart as well as provide additional information to augment the visualization. The legend to the immediate right of the chart tells the user what color mapping is currently being used on the chart (i.e. what number of wins corresponds to which green value). The bar chart in the top right of the visualization provides a histogram of the number of teams at each number of wins. The bars are colored to show which set of wins are being shown on the chart in order to provide a way for the user to always know what level of detail they are currently viewing. Finally, the data table at the bottom gives the detailed numerical data for the data in the chart for reference in cases of close comparisons or exact information requests.

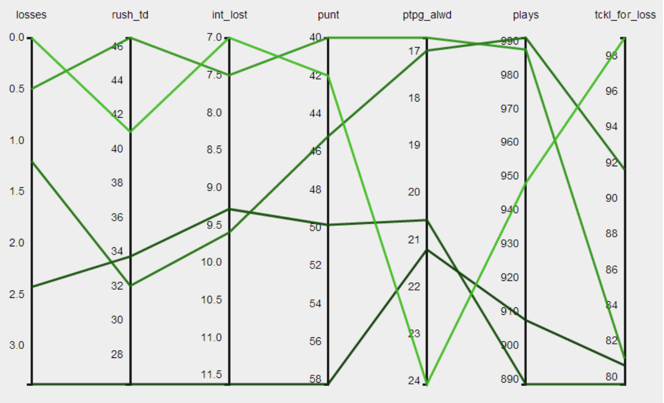


Fig. 7. The parallel coordinates chart shows how certain data categories can demonstrate correlation. The lightest green represents teams with 9 wins, all the way down to the darkest green representing teams with 5 wins. For all of the data categories in the chart, if the paths order themselves from lightest to darkest from top to bottom, this is an indication of possible correlation.

3.4 Visualization Analysis

As can be seen in Figure 7, analyzing the parallel coordinates chart, a user can quickly determine that the data categories on axes 1, 3, and 7 all have possible correlations that could be further investigated. Axis 5 exhibits the trait of an inversely correlated category but on closer inspection may not prove conclusive. The other axes (2, 4, and 6) clearly show no correlation properties and can quickly be dismissed.

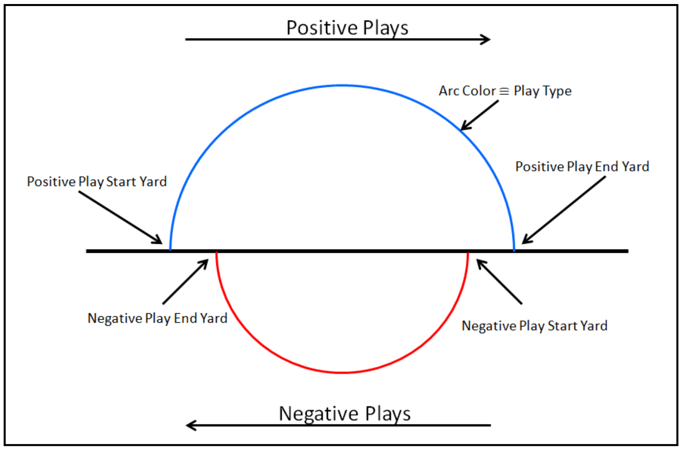


Fig. 8. This diagram shows the construction process for the arcs used in the game summary visualization

4 Game Summary Visualization

The game summary visualization provides an overview of the data associated with all of the plays (discrete actions) that occur in a single football game. This visualization uses the arc diagram structure originally developed by Wattenberg [7]. It has also been developed as a web-based visualization that again utilizes the d3 library at [9] and can be demoed at [12].

4.1 Data

The data set used for this visualization method consists of every play from a single collegiate American football game. Each play has nine unique parameters. They are as follows: play identifier, team identifier, drive number, play in drive, down, starting yard, ending yard, play type, and play description. The play identifier is a unique value that distinguishes each play from one another. The team identifier field indicates which team executed the play. The drive number field indicates which drive a given play was part of for grouping purposes. The play in drive field gives the order in which plays were run in a given drive. The down field gives on what down the play was executed. The starting and ending yard fields define where on the field the play took place. The play type field indicates which of the following six types of plays was run on a given play: pass, run, penalty, kick, punt, or turnover. Finally, the play description field is a string that contains the official record of the play including down and distance, involved players, and description of the action.

The data for the examples shown in this paper was obtained from the same website as the previous data set [14]. This sample data is taken from the game between Mississippi State University and the University of Tennessee on October 13, 2012. While the choice of a sample data set is arbitrary, it is useful that this set has multiple elements for every defined play type.

4.2 Visualization Tasks

In order to determine the necessary interactions for the visualization, its purpose must be established. The first and main goal of the season overview visualization is to provide the user with a unique summary of the action that occurred during a single football game. The second goal of the visualization is to allow users to analyze play data for trends comparing play types to location on the field. In order to be able to perform this kind of analysis, the user needs to remove the noisy data (unwanted play type arcs) from the

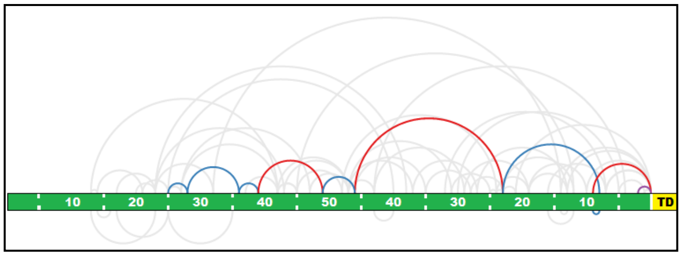


Fig. 9. An use of the drive selection functionality. In this example, it can be seen that the team moved steadily at first, running three times and then passing before running again. Then there were three large plays that moved them the rest of the way to the goal.

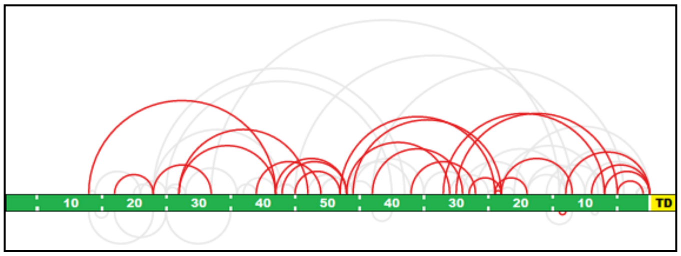


Fig. 10. An example of using the play type filter to perform a location versus play type analysis. In this case, it would appear that the team prefers to pass the ball more often in the middle of the field as opposed to near the endzone. Furthermore, a small negative passing play is visible on the underside of field line

field. Another analysis track that is popular among football analysts is the order of play types over an entire drive.

4.3 Visualization Design

This visualization technique is constructed by displaying arcs corresponding to single plays on a horizontal line that corresponds to the full 100 yard playing field. The arcs endpoints are placed on the line corresponding to where the play began and ended. Every arc is a regular semicircle. Therefore, taller arcs represent longer, and thereby more successful, plays. Any plays that result in a net loss of yardage are displayed on the underside of the horizontal line. The arcs are colored based on the play type of the corresponding play. A simple ordinal color scale was chosen for the arc colors and generated using Colorbrewer [8]. Figure 6 shows the color scale used to color the arcs and the associated play types for each color. Figure 8 shows a diagram of how the arcs are laid out on the line.

A key part of creating this overview is providing the user with a sense of the temporal ordering of the data. While other football visualizations use a second dimension to achieve this, as is shown in Figure 3, the data mapping is a false correlation to the spatial grid it is displayed on and therefore has the potential to lead to confusion. The approach that has been proposed to accomplish this is an animation that will systematically draw the arcs one at a time in temporal order creating the effect that the user is watching the action unfold live. This animation also has the added benefit of reinforcing the direction in with the play arcs are flowing. Furthermore, if the user hovers over a play in the field, a text box in the legend provides the official box record for that particular play. In order for the user to be able to analyze different play types, a filtering system was created to allow the user to select any subset of the play types at a time, while any deselected types are desaturated. Figure 10 shows an example case where it is obvious that team prefers to pass the ball in the middle of the field as opposed to close to end zone.

To accomplish the goal of allowing the user to view individual drives in the game, an interaction was included to allow the user to click on a play and have the drive that it is part of be highlighted while the rest of the arcs on the field are desaturated. This effect is shown in Figure 9. Furthermore, a motion animation was proposed for this scenario to allow the user to see how the sequence of plays occurred temporally. This animation is similar to that found in [11] where the illusion of motion is created by running a contrasting object along the arcs.

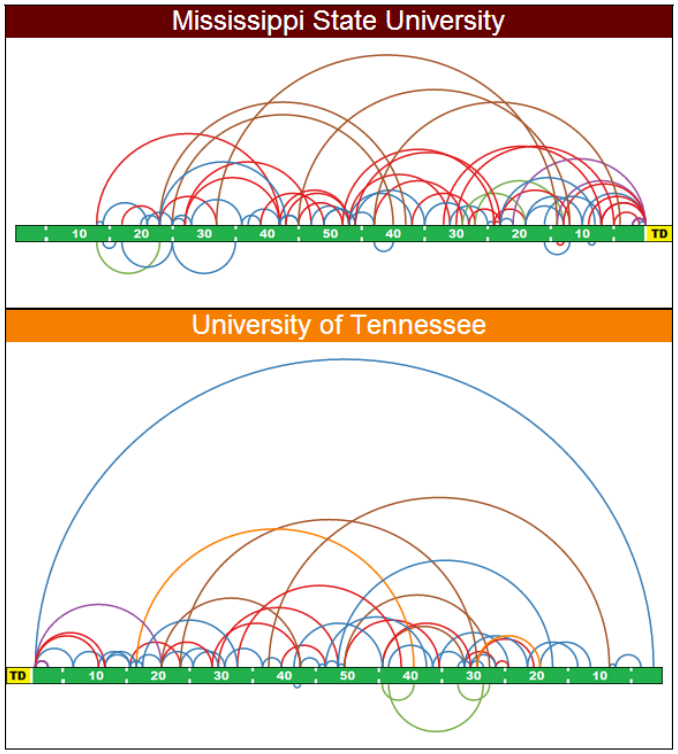


Fig. 11. A full view of the stacked approach for viewing both teams' play data at the same time. The top team's plays are traveling from left to right, while the bottom team's plays are moving from right to left

In order to display an entire game, both team's plays should be represented. Therefore, two different grids are stacked one on top of the other as shown in Figure 11. Each team's name and primary color are used as labels to separate them. To more closely mimic the flow of the sport, the direction in which the arcs flow is mirrored for the second grid (i.e. the top team's positive arcs flow from left to right and the bottom team's positive arcs flow from right to left.) Another visual affect that further emphasizes the direction of flow can be created by having each arc's width be larger at the starting endpoint and gradually shrink to a point at the ending endpoint. This effect hasn't been included at this development stage because of the difficulty to implement it and its minimal impact for providing a concept method.

4.4 Visualization Analysis

As can be seen from the images presented in this paper, the game summary visualization provides a user with an at-a-glance view of the type of action that occurred during a game of football. The larger arcs, as seen in Figure 11, show where and how large gains were made (e.g. the Tennessee team ran back a kickoff for 98 yards which is shown as the largest arc in the visualization). Furthermore, users can quickly determine if a team prefers running or passing (e.g. in Figure 1, the team appeared to throw the ball more often). Using the filters, information about what types of plays are run at different locations on the field can be gleaned. For instance, from Figures 1 and 10, it's clear that this team prefers to run as it gets closer to the end zone.

5 Future Work and Conclusion

As these visualizations are works in progress, there are many additional features that could be included to make them more useful. The first extension would be to allow for different data sets that the user can choose between (e.g. different games, teams, or even leagues). Moreover, the ability to view multiple games of data for the same team in the game summary visualization could allow the user to analyze trends over longer times. For the season overview visualization, including more coordination between the legend areas and chart and creating a separate selector for choosing which win ranges/totals to display would allow the user to more precisely control what data to look at and/or highlight.

In a similar manner, adding a range selection system for the arc visualization would allow the user to select smaller sections of the playing field to view data corresponding to plays that occur (start, end, or both) inside the range. Another useful addition to the game summary visualization would be a pane in the legend that provides basic information over the game for each team such as the total number of each type of play, average play distance, etc. Presenting this information could allow the user to quickly determine what the distribution of play types are for a given team in a given game.

In conclusion, while the visualizations presented in this paper are still works-in-progress, they present new methods for viewing data for the sport of American football. Specifically, the game summary visualization demonstrates the potential usefulness of using an arc diagram structure as a spatial mapping method.

Appendix A

List of statistic categories used in the visualization:
Wins
Losses
Rushing Yards/Game
Rushing TDs
Fumbles Lost
Pass Completion Pct.
Passing Yards/Game
Passing TDs
Interceptions Lost
Plays
Total Yards/Game
Points/Game
Two-Point Conversions
Field Goals
Third Down Pct.
Fourth Down Pct.
Time of Possession
Rushing Yards/Game Allowed
Passing Yards/Game Allowed
Interceptions Gained
Fumbles Gained
Total Yards/Game Allowed
Points/Game Allowed
Tackles for Loss
Sacks
Safeties
Punts
Average Punt Yards
Kick Return TDs
Penalties/Game
Penalty Yards/Game

References

[1] A. Cox and J. Stasko. "Sportsvis: Discovering meaning in sports statistics through information visualization." *Compendium of Symposium on Information Visualization*. 2006.

[2] A. Inselberg. "The plane with parallel coordinates." *The Visual Computer 1* (1985), 69-91.

[3] C. Smith. "College Football's Most Valuable Teams." *Forbes*. Forbes Magazine, 22 Dec. 2011. Web. 10 Dec. 2012. <http://www.forbes.com/sites/chrissmith/2011/12/22/college-footballs-most-valuable-teams/>.

[4] H. Pileggi, et al. "SnapShot: Visualization to Propel Ice Hockey Analytics." *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012): 2819-2828.

[5] J. Albert. "An introduction to sabermetrics." Bowling Green State University (http://www-math.bgsu.edu/~albert/papers/saber.html) (1997).

[6] M. Bostock, V. Ogievetsky, and J. Heer. "D^3 data-driven documents." IEEE Transactions on Visualization and Computer Graphics, 17(12):2301–2309, Dec. 2011.

[7] M. Wattenberg. "Arc Diagrams: Visualizing Structure in Strings." Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02), p.110, October 28-29, 2002.

[8] http://www.colorbrewer2.org

[9] http://d3.js

[10] http://digitalvideospace.blogspot.com/2012/10/second-screen-and-college-football.html

[11] http://hint.fm/wind/

[12] http://seangabrielowens.com/arc_viz/index.html

[13] http://seangabrielowens.com/parallel_viz/index.html

[14] http://web1.ncaa.org/mfb/mainpage.jsp