UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

JOSÉ LUIS SOTOMAYOR MALQUI

A Visual Analytics Approach for Passing Strategies Analysis in Soccer using Geometric Features

Thesis presented in partial fulfillment
of the requirements for the degree of
Master of Computer Science

Advisor: Prof. Dr. João Luiz Dihl Comba

Porto Alegre
April 2017

CIP – CATALOGING-IN-PUBLICATION

Sotomayor Malqui, José Luis

A Visual Analytics Approach for Passing Strategies Analysis in Soccer using Geometric Features / José Luis Sotomayor Malqui. – Porto Alegre: PPGC da UFRGS, 2017.

64 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2017. Advisor: João Luiz Dihl Comba.

1. Visual analytics. 2. Visual knowledge discovery. 3. Sport analytics. 4. Pattern recognition. I. Comba, João Luiz Dihl. II. Título.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. Rui Vicente Oppermann
Vice-Reitora: Prof. Jane Fraga Tutikian
Pró-Reitor de Pós-Graduação: Prof. Celso Giannetti Loureiro Chaves
Diretora do Instituto de Informática: Prof. Carla Maria Dal Sasso Freitas
Coordenador do PPGC: Prof. João Luiz Dihl Comba
Bibliotecária-chefe do Instituto de Informática: Beatriz Regina Bastos Haro

“The most incomprehensible thing about the universe
is that it is comprehensible”
— JOHN C. LENNOX

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my master advisor, Jõao Luiz Dihl Comba for his dedication, direction and discussion. I would also like to thank the the National Council for Scientific and Technological Development (CNPq) for their financial support and the Institute of Informatics (INF) at Federal University of Rio Grande Do Sul (UFRGS) for taking a chance on me. We are especially grateful to Opta Sports for generously providing us with the soccer data set.

During the period of almost two years, my friends were essential for my adaptation to a new city and country. I am very thankful to all my colleagues and friends which shared my journey to become a master, specially to Lizeth Castellanos, Gerson Groth, Marcelo Oliveira, Vinicius Machado, Guilherme Oliveira, Wagner Schimitt, Fabian Colque, Jorge Chamby, Julio Toss and Cícero Pahins.

Last, but not least, I would like to thank to my family members, especially my parents, Rafael Sotomayor Estrada and Maria Salomé Malqui Ortega. Without them I would have never made it this far in life. They always have been there for me to support every step of the way and encouraged me through all of my tough decisions.

ABSTRACT

Passing strategies analysis has always been of interest for soccer research. Since the beginning of soccer, managers have used scouting, video footage, training drills and data feeds to collect information about tactics and player performance. However, the dynamic nature of passing strategies is complex enough to reflect what is happening in the game and makes it hard to understand its dynamics. Furthermore, there exists a growing demand for pattern detection and passing sequence analysis popularized by FC Barcelona's tiki-taka. We propose an approach to abstract passing strategies and group them based on the geometry of the ball trajectory. To analyse passing sequences, we introduce a interactive visualization scheme to explore the frequency of usage, spatial location and time occurrence of the sequences. The frequency stripes visualization provide, an overview of passing groups frequency on three pitch regions: defense, middle, attack. A trajectory heatmap coordinated with a passing timeline allow, for the exploration of most recurrent passing shapes in temporal and spatial domains. Results show eight common ball trajectories for three-long passing sequences which depend on players positioning and on the angle of the pass. We demonstrate the potential of our approach with data from the Brazilian league under several case studies, and report feedback from a soccer expert.

Keywords: Visual analytics. visual knowledge discovery. sport analytics. pattern recognition.

RESUMO

As estrategias de passes têm sido sempre de interesse para a pesquisa de futebol. Desde os inícios do futebol, os técnicos tem usado olheiros, gravações de vídeo, exercícios de treinamento e feeds de dados para coletar informações sobre as táticas e desempenho dos jogadores. No entanto, a natureza dinâmica das estratégias de passes são bastante complexas para refletir o que está acontecendo dentro do campo e torna difícil o entendimento do jogo. Além disso, existe uma demanda crecente pela deteção de padrões e analise de estrategias de passes popularizado pelo *tiki-taka* utilizado pelo FC. Barcelona. Neste trabalho, propomos uma abordagem para abstrair as sequências de pases e agrupá-las baseadas na geometria da trajetória da bola. Para analizar as estratégias de passes, apresentamos um esquema de visualização interátiva para explorar a frequência de uso, a localização espacial e ocorrência temporal das sequências. A visualização Frequency Stripes fornece uma visão geral da frequencia dos grupos achados em tres regiões do campo: defesa, meio e ataque. O heatmap de trajetórias coordenado com a timeline de passes permite a exploração das formas mais recorrentes no espaço e tempo. Os resultados demostram oito trajetórias comunes da bola para sequências de três pases as quais dependem da posição dos jogadores e os ângulos de passe. Demonstramos o potencial da nossa abordagem com utilizando dados de várias partidas do Campeonato Brasileiro sob diferentes casos de estudo, e reportamos os comentários de especialistas em futebol.

Palavras-chave: Visual analytics, visual knowledge discovery, sport analytics, pattern recognition.

LIST OF ABBREVIATIONS AND ACRONYMS

F24 *Opta feed with match events data*

SSE *Sum of Squared Error*

FSV *Frequency Stripes Visualization*

FIFA *Federation Internationale de Football Association*

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

Soccer has always been a widely popular sport in the world and among the global sports events market, it is the sport with more revenues. Besides the huge investment, it is well known that soccer is the world's premier sport when it comes to popularity amongst its devotees (GIULIANOTTI, 1991). Fans are not the unique group of people supporting soccer. The performance of a team is also followed by managers, advertisers and club owners. In contrast with other sports, the low probabilities of scoring and the strategies applied by players add to the complexity of game analysis. This complexity increments due to the rules involved and external variables like weather, teams playing home or away, players health, and the alignment used by trainers during a match. The analysis of soccer matches is useful because it allows a team to learn about their errors and to study the adversary to take advantage of its weaknesses. Strategic analysis has been widely studied since the beginning, of the sport. An example of this are the different strategies used in the beginning which included all players in the attack position and no goalkeeper, passing to more elaborated strategies like the Italian "Catenaccio" or the Dutch "Total Football", which dominated the scene for years (WILSON, 2013).

Recently, research in sports analytics has increased with the help of technological advancements with GPS trackers and video processing. Behind the promise that the movement of the ball around the field and the positioning of players reflect a visible and measurable behavior of the team, clubs and companies have enabled the acquisition of new kinds of data using GPS devices and analysis of video footage. The output of such techniques consists in a set of logs that support managers with the analysis of events that occurred during a match.

Match statistics benefits both coaches and players by adding performance information to their knowledge. However, apart from key events in a soccer match such as shots, goals, fouls and number of passes, there exists an interest from the research community to understand the dynamic aspect of the game. To deal with the complexity of the game events, previous work (LINK; WEBER, 2015; SHAO et al., 2016; GYARMATI; ANGUERA, 2015) study particular sequences of events that might result in good scoring opportunities. These events could be player positions, shots or the positioning of the ball during a passing sequence. Hughes et al. (HUGHES; FRANKS, 2005) shows that, for successful teams, longer ball possessions were confirmed to produce more goals than shorter passing sequences. Simultaneously, the prevention of ball loss reduces the proba-

bility of taking a goal because of counter-attacks and blocks the ball possession from the rival.

The overall goal of our work is to investigate the most common ball trajectories created by pass sequences and how they relate to strategies on a specific period or pitch region. Defining a similarity trajectory distance we found eight common ball trajectories formed by three long passing sequences. Geometry shapes were always present in a soccer game. In passing, triangular formations create oportunities for offensive and defensive plays. Players spread out on the pitch allows a team to take advantage of more space effectively and move the ball all over the length of the field. Good team passing permits to maintain possession longer and create plays. In addition to ball control and dribbling, team formations also play an important role in passing strategies. This formations selected by a soccer trainer, modify the spatial distribution of players and the ball trajectory, which usually interferes with the quality of the passing. Given that the ability to maintain possession or turn over the balll is strongly related to the quality of passing, we observe that moving with or without the ball create shapes which help in building the passing support network, finally creating more chances to score.

One unexplored research field is the geometry of the trajectory of the ball during pass events and its relationship with players involved in those passing sequences. We use geometric passes features as an additive factor to support the analysis of passing strategies. Specifically, due to the correlation between player positions during a tactical movement and the direction and position of a pass destination. Although relevant approaches have been proposed to detect strategies in the last years, our proposal identifies playing patterns preferred by a team, which players execute them, and correlate their frequency with parts of the pitch and times of the game to provide a deeper understanding of the ball trajectory made of passes.

The main research contributions of our work are as follows:

- Provide an unsupervised approach to discover shape patterns in passing strategies based on clustering by geometrical similarity.
- A multi-faceted analysis of passing sequences using trajectory heatmaps coordinated with frequency stripes for interactive analysis of specific players and regions of the pitch;
- A case study of passing strategy analysis using data from Brazilian Serie A 2015 dataset, with feedback from a soccer expert.

Visual designs were integrated on an interactive analysis tool using multiple coordinated views which allows users to select subsets of the data using spatial and temporal filters and examine details on demand. Furthermore, our solution is designed as a web application for easy exploration of passing strategies from raw soccer data.

1.1 Design Requirements

Regarding player movement and performance, some challenging questions often faced by coaches are: Is the team applying the designated tactic scheme? Which player is frequently involved in dangerous situations? Is a certain player compatible with the team structure? Which pattern of passes our players prefer?. We used soccer analysis research questions to formulate tasks that should be supported by our geometry abstraction and visualization techniques. The visualization designs and encodings described in Chapter 4 are based on the requirements below:

- **R1: Space and Time exploration.** Search of individual and sequences of passes based on: pitch region in which they occurred and period of time. These features allow a tool to report the most common ball and player movement patterns between distinct pitch regions.
- **R2: Overview and detail of multiple matches.** Understand how the distribution of passing sequences changes through a match and during a tournament, allowing to compare the dynamics of the passing strategies and request details on demand.
- **R3: Rapid interaction and summarization.** Visual analytics tools must facilitate managers and scouts work without disrupting the flexibility for exploration. Visualization interfaces should be simple, interconnected and the data preprocessing method should be reproducible on any programming environment. Summarization helps to simplify the amount of data without loss of significant information and simplify the correlation of perspectives with time and space attributes.

1.2 Document organization

This thesis is organized as follows. Chapter 2, discusses existing work in related areas to soccer analytics and visualization. The third chapter introduces the details of the data preprocessing and feature extraction. In Chapter 4 we present the methodology

for shape discovering. Analysis and reports using real and recent dataset are presented in Chapter 5. Finally, we conclude this work, summarizing results and presenting future work ideas.

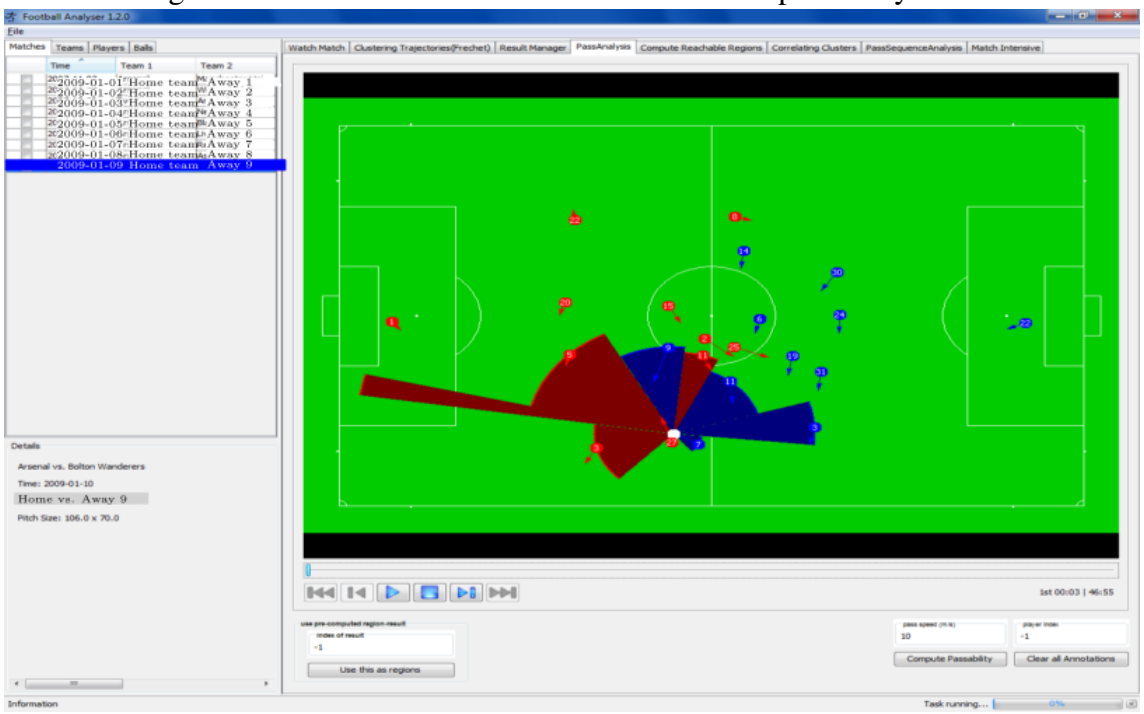
2 RELATED WORK

In this chapter, we describe related work to our approach. We divided previous work in two parts: statistics and data mining oriented work and visualization techniques in soccer and tactics analysis.

2.1 Statistics and Data mining

Regarding statistic analysis, data mining and information discovering research we identified the following related work. Gudmundsson and Wolle (GUDMUNDSSON; WOLLE, 2014) developed tools to cluster passes and movement of individual players. They calculate all possible passing alternatives in a given time and compute the most frequent pass sequences. Additionally, they computed correlations between sub-trajectory clusters computed from players movement as an evaluation of frequent actions. An interesting aspect of this work is the use of *dominant regions* which is defined as the geometric region around a player in which that player could receive a pass.

Figure 2.1: Gudmundsson and Wolle GUI for the pass analysis tool

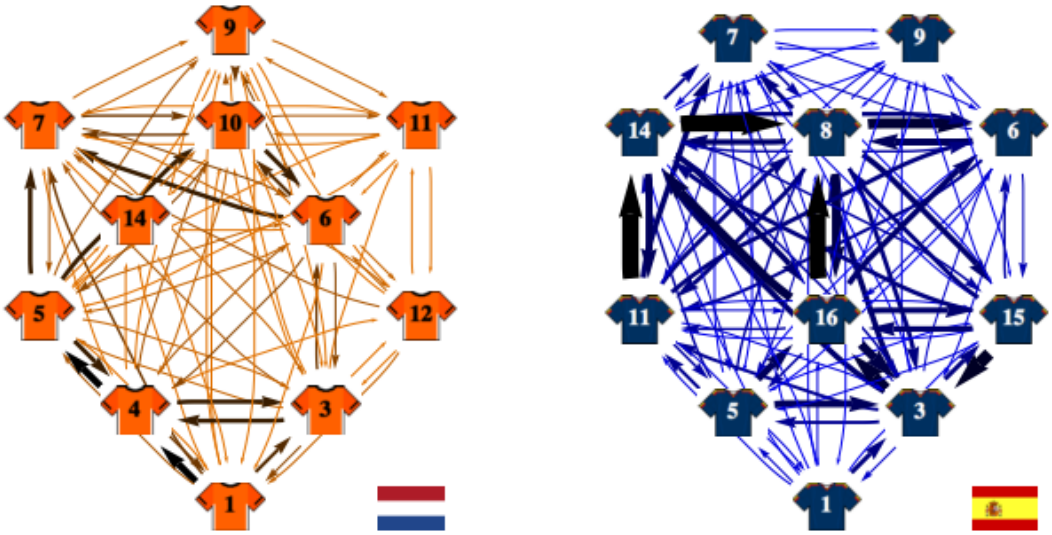


Illustrates the possible passes for a given passing speed. A wider cone indicates a simpler pass, while a narrow cone indicates that the direction of the pass has to be very accurate to be a successful pass. Source: (GUDMUNDSSON; WOLLE, 2014).

Lucey et al. (LUCEY et al., 2013) highlighted the problem of alignment when

dealing with multi-agent trajectories and presented a representation based on player "role" instead of one based on player "identity". They showed an effective way of discovering team formation and plays using the proposed role representation. Regarding the analysis of team behavior in soccer, Pena and Touchette (PEñA; TOUCHETTE, 2012) used tools from network theory to describe team strategies. They defined a passing network with players as nodes and edges weighted by the successful number of passes completed between them. The passing networks provide a visual summary or 'snapshot' of a football team's style. To validate their approach they used data made available by FIFA during the 2010 World Cup. From the resulting network they identify play patterns, determine key events on the play and potential weaknesses. To do this, they define the popularity or relevance of a player following different parameters. Although their approach is static, they complement measuring the individual contribution of a player using local network invariants of the passing network. The three measures used by (PEñA; TOUCHETTE, 2012) are: network closeness, betwenness and Pagerank centrality.

Figure 2.2: Passing networks for Netherlands and Spain

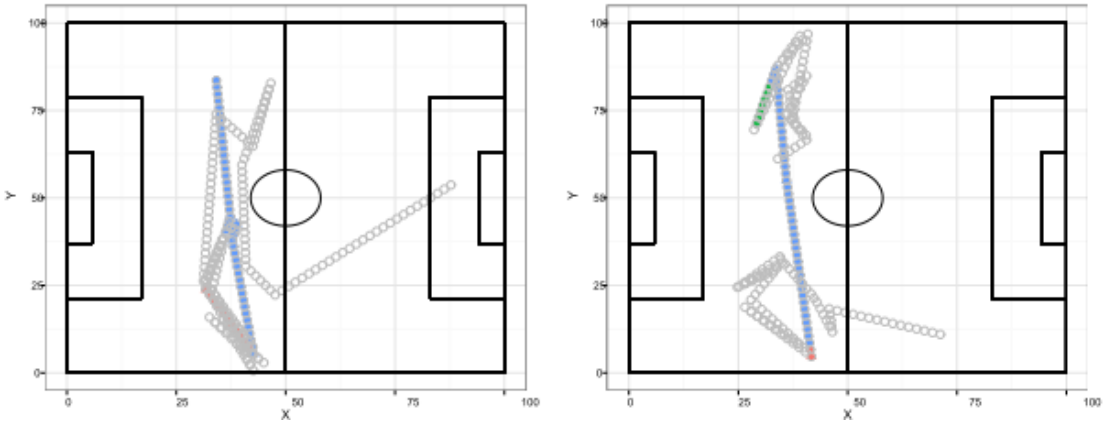


Visualization drawn before the final game, using the passing data. They use a passing network as a tool for visualizing a team's strategy by fixing its nodes in positions corresponding to players' formation on pitch. Source: (PEñA; TOUCHETTE, 2012).

Wei (WEI et al., 2013) take advantage of "role-representation" and used a feature reduction strategy to create a compact spatiotemporal representation. They found the most likely formation patterns of a team by clustering plays associated with a particular event (such as shots, corners, free-kicks). Using the spatiotemporal representation they show the capability to segment a match into game phases and distinctive hihglights without manual intervention.

Gyarmati and Anguera (GYARMATI; ANGUERA, 2015) propose a method based on Dynamic Time Warping to reveal the tactics of a team through the analysis of repeating time series. They define sequences of passes as time series and use a DTW based technique for time series matching. The algorithm finds the start and end locations of two subsequences automatically; however, the pattern may not start or terminate at an endpoint of the pass. Finally, the output of the tecnique is a list of the most common sequences of passes. The basic idea of extracting reoccurring pass patterns is to obtain a match between two subsequences of passes if they are close enough. Using this approach, they were able to automatically select all the recurring passes of a season and reveal insights on the strategy of a team. The analysis of the presented patterns reveal underlying strategies of teams such as: counter attacks with or without side changes, balanced ball possession and crosses. They evaluate the proposed method with one season of the Spanish first division dataset. The evaluation investigated the prevalence of the designed passing sequences of individual teams.

Figure 2.3: Gyarmati et. al. Dynamic Time Warping approach



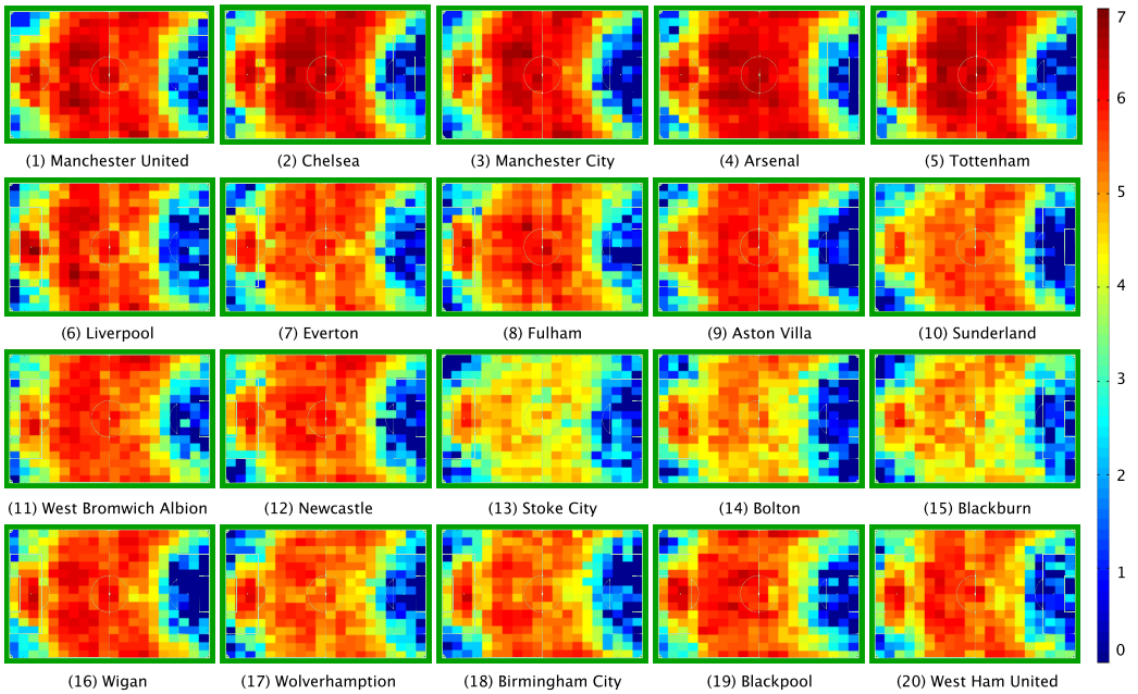
Illustrates the procedure of the Dynamic Time Warping scheme. They extract the reocurring pattern (blue circles) from the whole pass sequences of a team (grey circles). Source: (GYARMATI; ANGUERA, 2015).

More recently, Wang et al. (WANG et al., 2014) developed a model for learning tactical patterns named as Team Tactic Topic Model (T^3M). Withing the proposed model, they consider that each team has several tactical patterns. Each pattern consists of a passing pattern and a team positioning. Additionally, a visualization of tactical patterns over multiple matches is presented. Different colors were mapped to different found patterns. They demonstrated meaningful tactical patterns and showed tactics that are more likely to score a goal.

As an alternative to trajectory analysis, Lucey et al. (LUCEY et al., 2013) used

Occupancy maps to make comparisons between team's style of play. Their approach consist in partitioning the ball tracking data in possession strings and quantizing the field into bins. Subsequently, a distribution is built to characterize the expected behaviour in each location. The top ranked teams presented higher entropy over most of the field compared to the lower ranked teams, which gives an indication that these teams are less predictable. They visualized the difference of occupancy between home and away matches and provide a method of automatically flagging behavioural differences.

Figure 2.4: Occupancy Maps Visualization



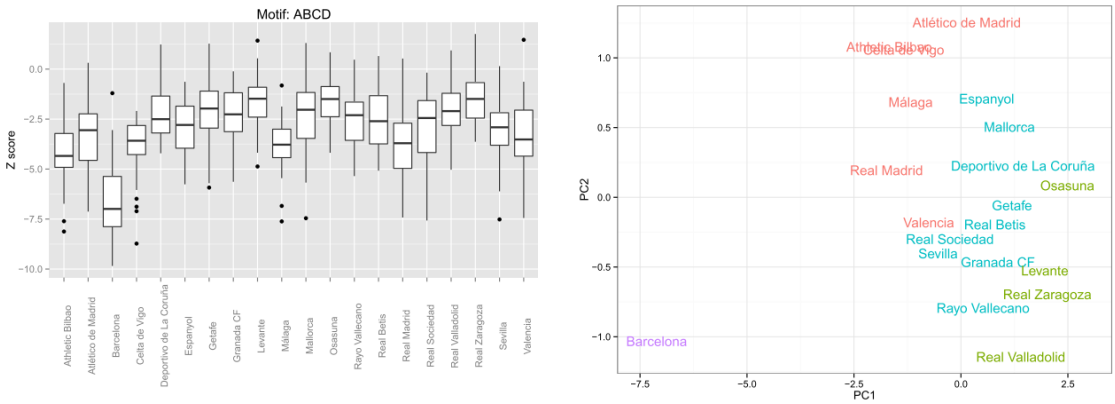
Mean occupancy maps using entropy to describe each pitch area. The visualization can give a indication of redundant patterns. The maps were normalized for teams attacking from left to right. Red and blue areas were mapped to high variability and predicatable behaviour respectively. Source: (LUCEY et al., 2013).

This was followed by (BIALKOWSKI et al., 2014), which utilized a formation descriptor to determine the identity of a team. They wanted to discover the team-based features that discriminate a team's behavior. To do so, they minimize the entropy of role-specific occupancy maps, projecting them into a low-dimensional discriminative feature space using linear discriminative analysis (LDA). Their approach presented the capability to characterize individual team behaviour three times better that other match descriptors. The presented descriptor was validated using an entire season of player tracking data provided by Prozone. (MILO et al., 2002) introduced the concept of network motif. Passing networks can be reduced to complex networks which Milo et al. studied in order to find structure similarities and perform information processing. To quantify soccer

changes in (LUCEY et al.,), Lucey et al. presented a method to estimate the likelihood of chances of scoring. They trained a logistic regressor with strategic features such as a defender proximity, speed of play and defensive formation. In terms of large scale analysis, Bialkowski (BIALKOWSKI et al.,) worked with large datasets and presented a method to conduct both individual player and team analysis. They discovered player roles from data by utilizing a minimum entropy data partitioning method and automatically detected formations.

Gyramati et al. (GYARMATI; KWAK; RODRIGUEZ, 2014) propose to quantify the motif characteristics of soccer teams using "flow motifs". The proposed "flow motis" consist of a given number of consecutive passes executed in a time window. They relax the identity of the involved players and focus on the structure of the passes. Five distinct motifs were presented: ABAB, ABAC, ABCA, ABCB and ABCD while analyzing three-pass long motifs. After having the motifs identified in the passing network, they quantify the prevalence of the motifs using the z-score by comparing the original and constructed random passing networks. As a result, they try to find similarities and disparities between teams using data from Spanish and other European Leagues. Although most teams tend to apply similar style, they showed that a unique strategy of soccer exists, applied by FC Barcelona.

Figure 2.5: Searching for a unique style in soccer using flow motifs.



Left: Z-scores of the ABCD motif. FC Barcelona uses the ABCD less often than the other teams. Right: K-means clustering + PCA of the teams in the Spanish League. FC Barcelona present a unique style based on its passing motifs. Source: (GYARMATI; KWAK; RODRIGUEZ, 2014).

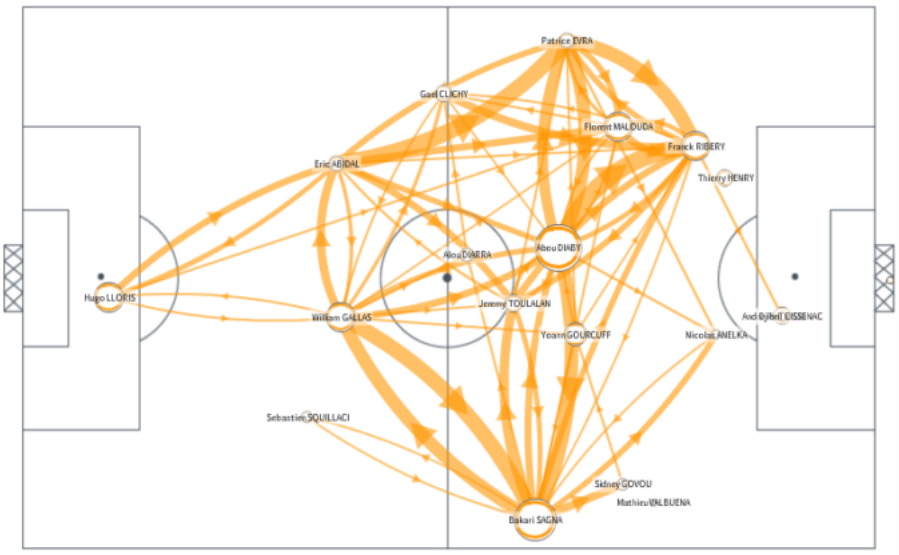
We expand on their motif analysis by considering the relation with the geometric clusters we identify in the passing sequences.

2.2 Visual Analytics

In combination with statistics, visualization and analytics techniques are used to extract insights from sports data. A popular visual design are heatmaps (SPORTS, 2016), which map player's most frequent positions to colors. The heat map can be used to illustrate several aspects of the performance of a player or a team. For example the movement of the ball, quantity of engagement and player positioning.

Another one is the flow graph (FOOTOSCOPE. . . , 2010) where the team is represented as a graph, with players as nodes and the links show the connections between players. Footoscope provides a perspective on the morphology and tactics of a football team using raw data about passing events transformed into indicators and visualizations. Their tool was tested with a soccer expert based on the raw statistics of the World Cup in South Africa accessible on the FIFA Web site. The results discuss, for instance, the key role of player Schweinsteiger in Germany's midfield that other players such as Stankovic failed to reproduce.

Figure 2.6: Footoscope



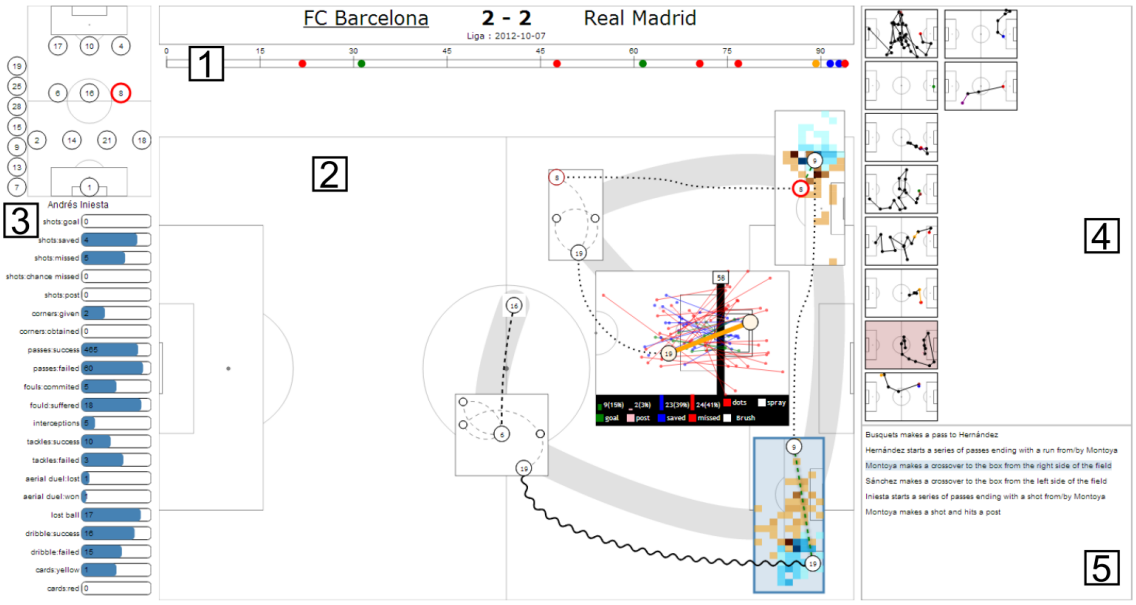
Communication between players, Image from *http://www.footoscope.com/*, courtesy of Fabien Girardin. Source: (FOOTOSCOPE. . . , 2010).

There are several visual analytics systems for sports analytics. Soccer Scoop (RUSU et al., 2010) and MatchPad (LEGG et al., 2012) use glyph-based visualizations to compare soccer players and analyze performances during games. CourtVision (GOLDSBERRY, 2012) and SnapShot (PILEGGI et al., 2012), respectively designed for basketball and hockey, introduce specific types of heatmaps focused on the ball and puck shots. Legg et

al. (LEGG et al., 2013) describes a visual search system for Rugby matches. They used a sketch-based interface to perform a search without semantic annotation.

Perin et al. (PERIN; VUILLEMOT; FEKETE, 2013) developed the tool SoccerStories, a visualization interface to support analysts in exploring soccer data and communicating interesting insights. Their tool offers different views on soccer match data for event comparison and generating automatic reports. The system provides an overview+detail interface to study game phases and their aggregation into a series of connected visualizations. The main visualization interface presents a game list to choose from, allows navigation into game phases using a timeline and small multiples, selection of phases and aggregation into faceted views and finally, details are provided on the side for selected players. Additionally they present the option to export the phase as word-sized graphics to embed into text.

Figure 2.7: SoccerStories user interface

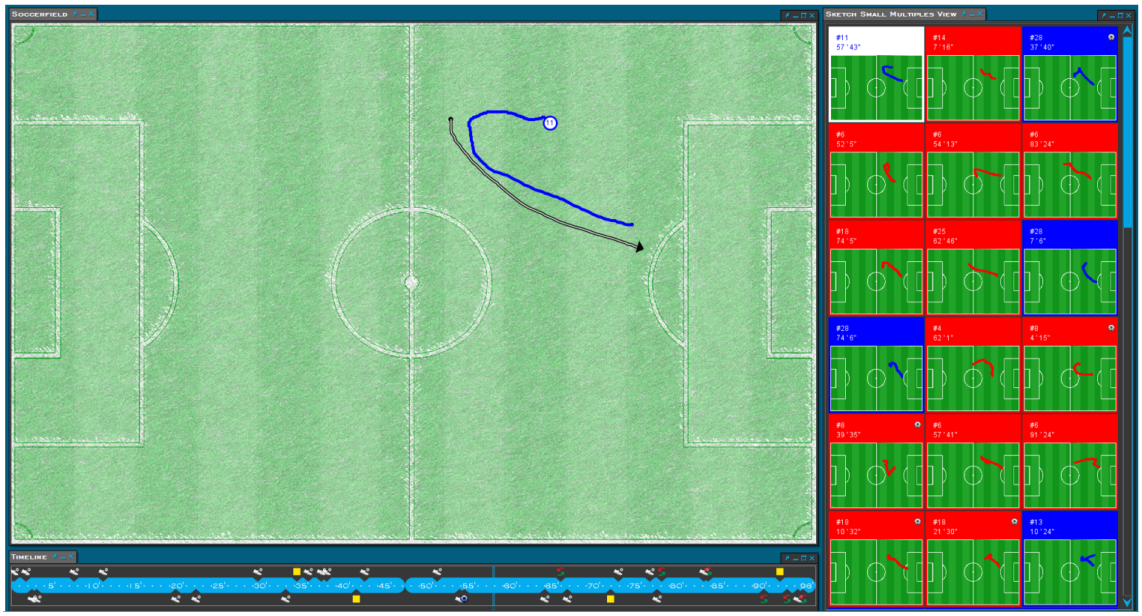


(1) Complete game overiew as a timeline. (2) temporal zoom on a game phase and layout on a soccer field. (3) additional details. (4) thumbnails and (5) text annotations. Source: (PERIN; VUILLEMOT; FEKETE, 2013).

Janetwzko et al. (JANETZKO et al., 2014) detected relevant events and phases semi-automatically by integrating statistical features. Soccer drawn is a visualization that presents an analysis of a soccer game representing continuous movements of the ball as lines (ROSENTHAL,). The position of the lines in the same part of the pitch reveals trends in how the game was played, offering a brief summary. Soccer simulators provide automatically visualizations of matches to help managers process games simultaneously and improving the decision making.

Regarding soccer trajectories, Shao et al. (SHAO et al., 2016) proposes a novel approach for searching trajectory data in soccer matches in which the user sketchs a situation of interest based on two different similarity measures. Furthermore, they use a domain specific prefiltering process to extract a set of movement segments, which are similar to a given sketch. To do this, a single-trajectory, multi-trajectory and event-specific search functions were used based on two different similarity measures. The similarity search is designed to find the exact position, rotation and scaling of the trajectory (player and ball). They propose an analysis workflow to enable the analyst to define a query object directly on the soccer pitch by applying a sketch of the desired trajectory at the desired place. Then, the analyst may inspect the result set and to further refine th analysis they can add additional filters, constraints or trajectory sketches.

Figure 2.8: Searching trajectory data in soccer

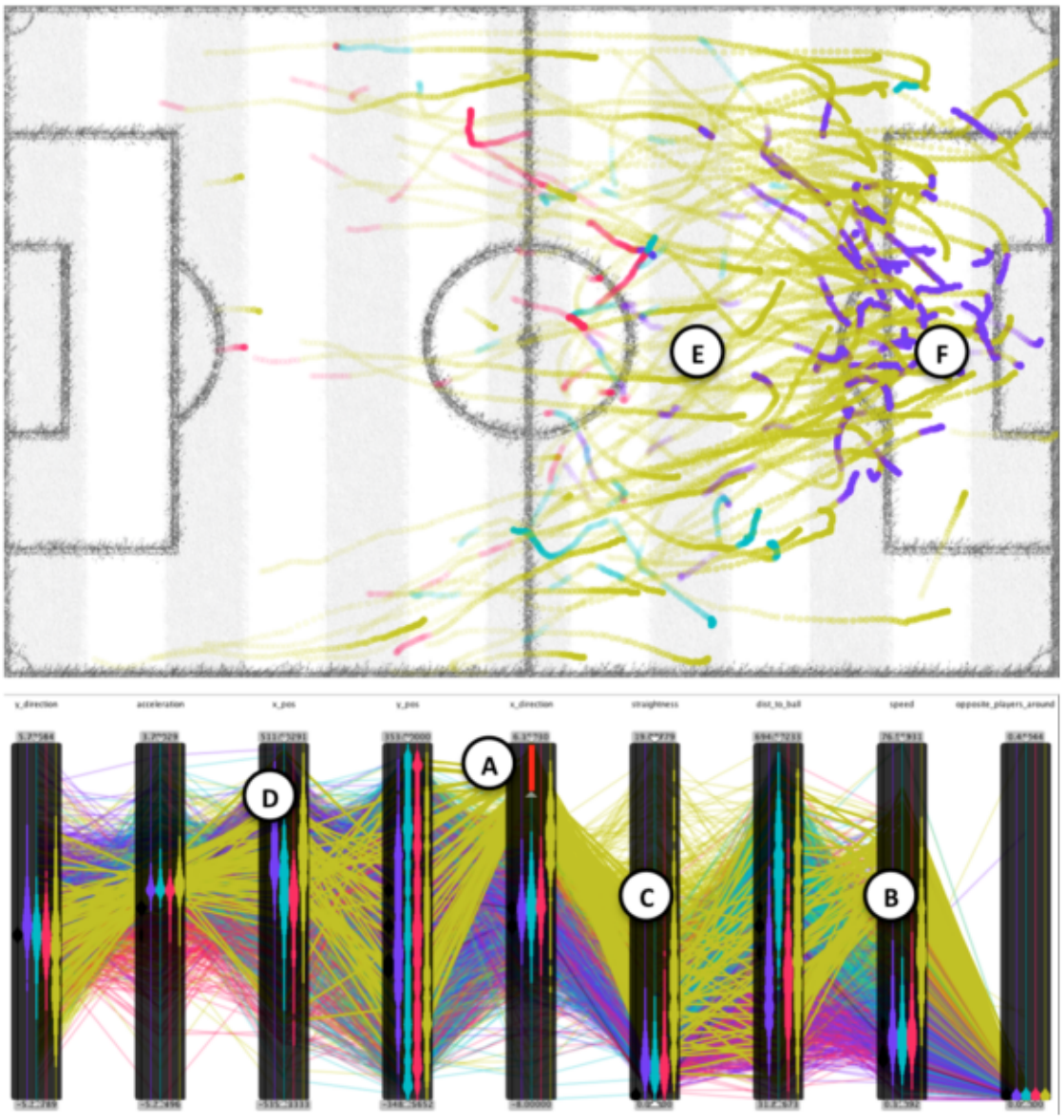


The prototype show a single-trajectory search by applying a spatial distribution approach. In this illustration, they search for movement patterns that start from the outer mildfield and run into direction of right penalty area. The best matching results are shown in the right panel. Source: (SHAO et al., 2016).

Stein et al. (STEIN; SACHA, 2016) uses a parallel coordinate plot for statistical analysis. They include a density distribution representation of clustered data for activity phases of professional soccer players. In this work, the authors visualize the frequency distribution of clusters along a dimension axis. They implemented stacked bar charts and violin plots to show a kernel density estimation, commonly used in Statistics. For further statistical measures, they integrate box plots to measure median,quartiles and outliers. The applicability of the enhancements were shown by analyzing recorded soccer move-

ments. The data consist of overall 66 professional soccer matches. For each of the 22 players two-dimensional position data were available with a temporal resolution of 100 milliseconds.

Figure 2.9: Enhanced Parallel Coordinate plot for soccer data



Top: Spatial trajectory visualization. Bottom: Parallel coordinates enhanced with violin plots. Colors represents the actiity phases of players. Yellow back run phases are followed by purple ending phases near the penalty area of the own goal. Source: (STEIN; SACHA, 2016).

Our proposal differs from previous work (Table 2.1) in that we allow the analysis of passing sequences using the shape of the ball trajectories in the spatio-temporal domain. Our visualizations propose an overview to analyze multiple matches along with individual player participation within the passing strategies.

Table 2.1: Related work comparison

Related Work	**Technique**	**Spatial Analysis**	**Time Analysis**	**Player Analysis**	**Within Analytical Abstraction**	**Within Visualization Abstraction**
(GUDMUNDSSON; WOLLE, 2014)	Passing analysis, dominant regions	✓	✓		Fréchet distance, clustering	
(LUCEY et al., 2013)	Role-based detection	✓			Bilinear basis model, Hungarian algorithm	
(PEñA; TOUCHETTE, 2012)	Graphs			✓	Graph Theory	Graph layout
(WEI et al., 2013)	Decision-forest classifier	✓	✓		Decision Tree	
(GYARMATI; ANGUERA, 2015)	Passing sequences comparison with DTW	✓			Dynamic Time Warping	
(WANG et al., 2014)	Team Tactic Topic Model for tactical learning	✓			Team Tactic Topic Model	Matrix Visualization
(LUCEY et al., 2013)	Occupancy Maps	✓			Mean, median, mode, entropy	Heatmap
(BIALKOWSKI et al., 2014)	Team identity prediction	✓			LDA	Heatmap
(LUCEY et al.,)	Expected Goal Value method	✓	✓		Likelihood, hungarian algorithm	
(BIALKOWSKI et al.,)	Role-based representation	✓	✓	✓	Expectation Maximization	
(GYARMATI; KWAK; RODRIGUEZ, 2014)	Flow motifs frequency analysis	✓			Z-score and PCA	
(FOOTOSCOPE. . . , 2010)	Morphology evaluation			✓		Graph layout
(STEIN; SACHA, 2016)	Enhanced Parallel Coordinates for Player Analysis	✓		✓		Parallel Coordinates, Bar charts, Violin and Box Plot
(PERIN; VUILLEMOT; FEKETE, 2013)	Soccer stories visualization	✓	✓	✓		Overview + detail multifaceted visualizations
(SHAO et al., 2016)	Trajectory search	✓	✓		Grid, edit distance	Search Interface
(JANETZKO et al., 2014)	Events Visual Analytics System	✓	✓	✓	LMT,LibSVM classifiers	Horizon graphs
Proposal	**Passing sub-sequences trajectory analysis**	✓	✓	✓	**K-means and Spectral Clustering**	**Frequency Stripes Visualization (FSV)**

3 GEOMETRICAL ABSTRACTION OF PASSING STRATEGIES

Gyarmati et al. (GYARMATI; KWAK; RODRIGUEZ, 2014) introduce the concept of "flow motifs" to characterize the statistically significant of pass sequence patterns. To do this, they use the z-score to measure highly significant subgraphs that usually consist of three or four nodes. In a generalization to the soccer motifs analysis developed by Gyarmati et al. (GYARMATI; KWAK; RODRIGUEZ, 2014), we developed a clustering algorithm method to group passing sequences by shape similarity. Figure 3.1 shows the pipeline of our approach. A pre-processing step re-samples passing sequences and apply invariant transformations. We highlight the need of invariant transformations because we are only interested in the shape of the ball trajectory. The similarity of passing sub-sequences is defined based on an algorithm for gesture recognition (LI, 2010). The clustering algorithm uses a K-means and a spectral clustering algorithm in sequence. Finally, the clustering results are displayed using edge-bundling to reduce cluttering and visually identify main structures from each cluster.

All the sub-possessions are formed by four players who participated in the sequence of passes and the sub-possession trajectory is composed of the position of the player where they made the pass. The input for our proposed approach are pass events taken from soccer logs, in our case provided by Opta. The main attributes needed from those logs are x and y positions, match time and player identity. We refer to the four-point trajectory as the shape of the sub-possession or soccer motif.

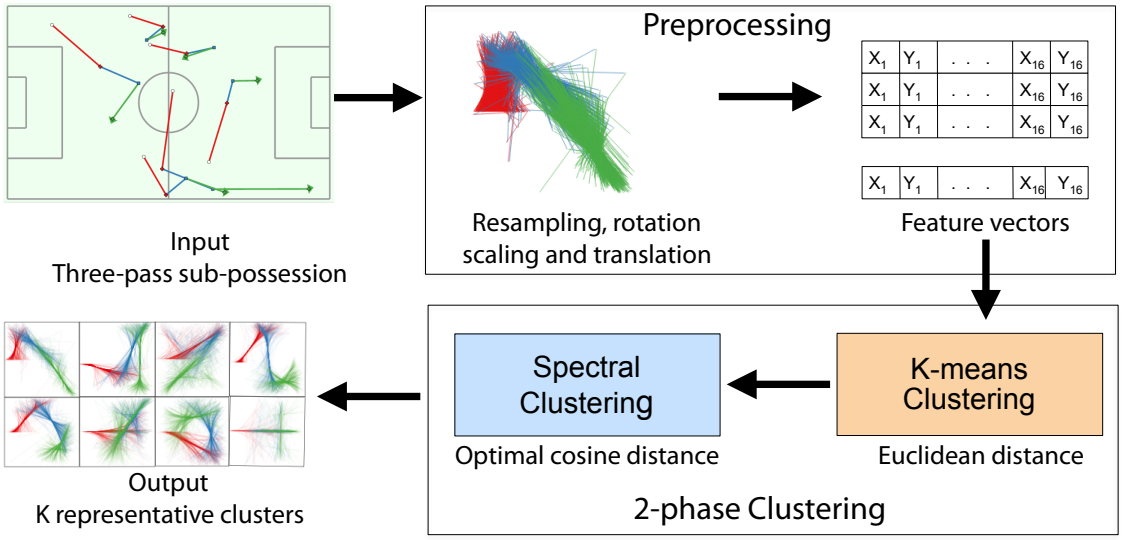
3.1 Soccer data description

The data used in this work is the 2015 Brazilian Serie A season F24 Opta feed (OPTA, 2017). The F24 Opta feed consists of XML files with event tags, describing specific match situations along their position. Some events include passes, fouls, tackles and corners. Due to data restrictions, only the second half of the tournament was considered for our analysis. The dataset consists of 180 games with more than 300,000 events. Our analysis used 18 games per team, which corresponds to matches from the second half of the year.

The relevant fields for our approach are pass events, player identity and event position. The format of a pass event p_n is

$$p_n =< id, x, y, player_i, player_j, t(n) >$$

Figure 3.1: Clustering Pipeline



Input: three-pass sub-possessions are first resampled and transformed to be invariant to orientation and location. Transformed passing sequences are processed by a k-means and an spectral clustering. In the output, we generate clusters that encode the shapes most used in passing strategies

where

n : Index of the pass in the possession chain

id : Identifies an event within the entire dataset.

x : Pitch x-coordinate of the pass.

y : Pitch y-coordinate of the pass.

$player_i$: Identifies the player who performed the pass.

$player_j$: Identifies the player who recieved the pass.

$t(n)$: The time instance when the pass was executed.

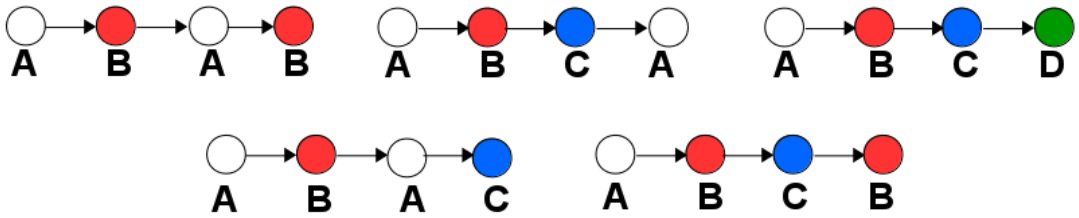
To identify a pass sub-sequence it is necessary to establish all the ball possessions that a team had. Ball possession is formally defined by Gyarmati et al. (GYARMATI; KWAK; RODRIGUEZ, 2014) as a sequence of passes $< p_1, p_2, , p_n >$ that fulfill two constraints:

$$player_j(m) = player_i(m+1), \forall m \in \{1, \ldots, n-1\}$$

$$t(m+1) - t(m) \leq T_{max}, \forall m \in \{1, \ldots, n-1\}$$

Similar to (GYARMATI; KWAK; RODRIGUEZ, 2014), we used the time $T_{max} = 5sec$ to determine if two passes belong to the same possession. The contraints shown assure that the passes are consecutive and not having major breaks. Then, we extract all

Figure 3.2: The five distinct motif structures: ABAB, ABAC, ABCA, ABCB and ABCD.



Players identifiers are converted into the appropriate A, B, C and D labels to assemble the motifs.

the three-pass long sub-possessiosn and convert the player identifiers to the labels A, B, C or D.

3.2 Soccer Motifs

We propose an spatio-temporal strategy analysis at a higher abstraction level than single pass events, using three-pass sub-possessions. Gyarmati et al. (GYARMATI; KWAK; RODRIGUEZ, 2014) shown that three-pass sub-possessions have discriminative value to identify different styles of play. However, their approach does not consider the identity of the involved players and focuses on the structure of the passes. As an example we consider the sub-possession ABAC which represents a sequence of passes with three different players whose identity is unknown. The first motif is formed by Player A passing to player B, player B back pass to player A and finally, player A passes to player C.

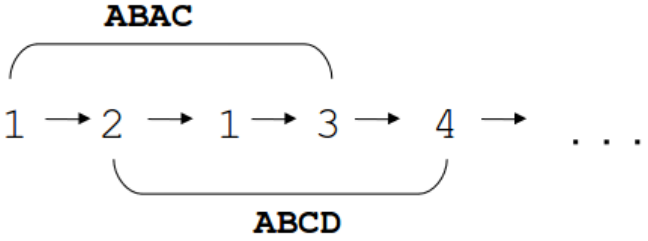
The four letter sub-possession is called a soccer motif. For three-pass sequences, there exist five motifs presented in Figure 3.2.

Our work differs from (GYARMATI; KWAK; RODRIGUEZ, 2014) in that we explore the players' involvement in the motifs, we focus on the geometric shape of the soccer motif and finally, propose a novel visualization design to support the study of passing sequences.

3.3 Preprocessing

We compute ball possessions for each match using a threshold $Tmax = 5$ seconds. That means that a pass is considered part of the possession if and only if it was made between the first five seconds after receiving the ball from other player. With this approach we try to focus on fast, accurate and structured passing. For each ball posses-

Figure 3.3: Example of a three-pass long structure analysis.



Ball possession where the ball moves between players $1 \rightarrow 2 \rightarrow 1 \rightarrow 3 \rightarrow 4$ translates into two motifs: ABAC and ABCD.

sion, we select all three-pass sub-possessions (see Figure 3.3).

A ball trajectory is composed of the four points which represents the players position on the pitch when they passed the ball. Since we are interested in trajectory shape similarity, we use a comparison metric which try to find the best alignment among sets of points. The sub-possession preprocessing consists of four operations: resampling, rotation, scaling and translation. To increase the accuracy of similarity comparisons, we resample the initial four point trajectory to $n = 16$. The parameter selection was based on a gesture recognition work, which proved to be efficient and accurate (WOBBROCK; WILSON; LI, 2007).

First, we sum the distances between each player and divide this length by $n - 1$ to obtain the increment *dx* between each resampled points. Then, the trajectory is iterated from the first pass position with a step of a *dx*, adding new points based on linear interpolation. As our objective is to group trajectories by shape similarity, we need a metric that is orientation-invariant, so we rotate the resampled shape around its centroid by its "indicative angle". For the rotation operation, the "indicative angle" approach (WOBBROCK; WILSON; LI, 2007) was used to approximate the best rotation angle that aligns two trajectories.

Depending on the pass distance, ball trajectories may vary in size. To overcome this, we apply a non-uniform scale transformation to a square domain for uniform comparison. To avoid horizontal and vertical lines distortion by a non-uniform scaling, we identify the horizontal and vertical pass trajectories by testing if the smaller dimension of the trajectory bounding box exceeds a threshold. If it does, we scale them uniformly. Finally, the shape is translated to a reference point (0,0). All trajectories are represented as a feature vector $v = [x_1, y_1, x_2, y_2, \ldots, x_n, y_n]$.

3.4 Geometric Clustering of Passing Sequences

3.4.1 Similarity Metric

A similarity metric is needed for ball trajectory shape comparisons. Given the previous alignment, the problem is to rotate two vector shapes v_t and v_g in an angle in which the distance between each point is minimized (θ). To solve this problem, we use the approach for gesture recognition described by (LI, 2010). It uses the *optimal angular distance*, which computes the cosine distance to find the angle between two vectors in a high-dimensional space. This allows to find the optimal angle much faster, thus improving the number of computations. Previous work, solved the optimization problem using a closed-form, solving the following equation:

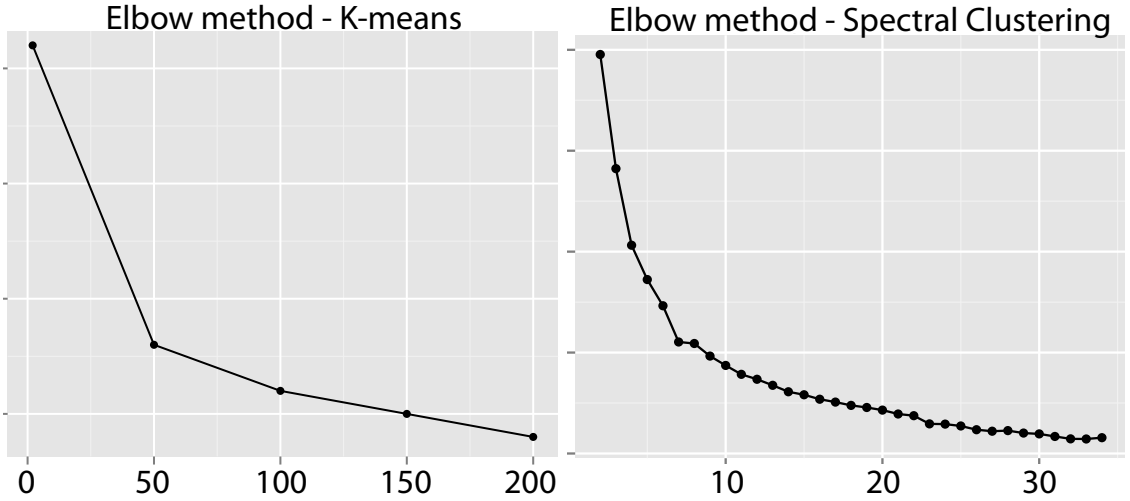
$$\theta_{optimal} = \underset{-\pi \leq \theta \leq \pi}{\arg\min} \Big(\arccos \frac{v_t(\theta) \cdot v_g}{|v_t(\theta)||v_g|} \Big)$$

where $v_t(\theta)$ is the vector obtained after rotating v_t by θ. We used the *optimal angular distance* for comparing sub-possessions.

3.4.2 K-means and Spectral Clustering

The discovery stage is based on a two-stage clustering algorithm. First, we apply a K-means algorithm with the Euclidean distance to reduce the number of vector shapes and generate a representative subset. The K-means clustering with euclidean distance does not allow to discriminate shapes by orientation. To overcome this need, we propose a second clustering stage with a similarity distance based on the rotation angle needed to minimize the distance between trajectories. Hence, a graph-based clustering approach was preferred over a centroid-based approach for the second clustering stage. Spectral Clustering deals with clustering as a graph partition problem. Given data point $(x_1, \ldots, x_n)$ pairwise affinities, the algorithm builds a similarity graph in which vertices correspond to data points and edges represent the distance between data points. The spectrum of the similarity matrix is used to perform a reduction of dimensionality. Finally, the clustering result corresponds to finding a cut through the graph (PLANCK; LUXBURG, 2006). The implementation used in our experiments was from (UW et al., 2001) due to its simplicity, good results and proved stability. Spectral clustering often outperform

Figure 3.4: Determining the number of clusters



Sum of squared error (SSE) for different K values using the elbow method. We used $K = 50$ and $K = 8$ for the K-means and spectral clustering, where the SSE decreased abruptly.

traditional approaches and can be solved by standard linear algebra methods.

We use spectral clustering as it does not rely on parametric density assumptions and avoid multiple restarts from traditional clustering methods. Additionally, a centroid-based clustering algorithm would lead us to misleading results due to the fact that we expect to obtain different orientation shapes in each cluster. To select the correct number of clusters and to prevent loss of information we run the clustering algorithms several times and select the number of clusters using the elbow method (KODINARIYA; MAKWANA, 2013). The selection of clustering parameters using the elbow method is presented in Figure 3.4. The elbow method looks at the amount of variance explained as a function of the number of clusters. An ideal number of clusters is the one that adding another cluster does not improve significantly the modeling of the data. Empirical evidence over our data set shows that the elbow method correctly chooses the number of clusters. The first stage K-means algorithm allowed us to reduce the domain to 50 representative clusters from almost 70 000 trajectories. The 50 clusters obtained were then subject to the spectral clustering, which were re-organized into final 8 clusters.

3.5 Displaying Geometric Clusters using Edge Bundling

The visualization of the passing sequences within each one of the 8 obtained clusters gives an intuition of the shape of the cluster. In the left of Figure 3.5, we display all passing sequences using red, blue, and green lines corresponding to the first, second, and

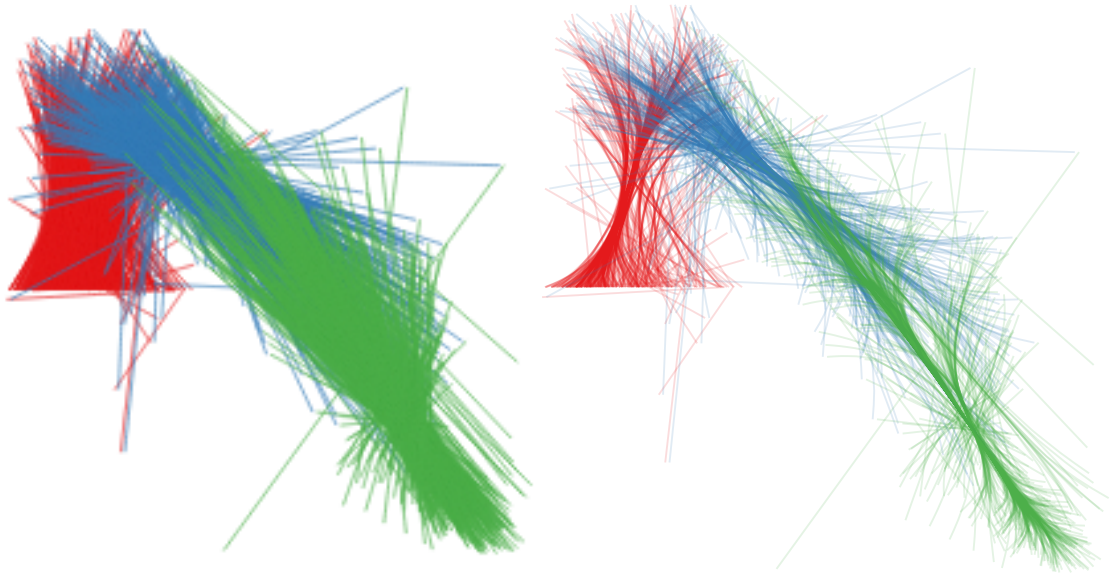
third pass from the sub-possession. All passing sequences are centered at a reference point (0,0) per cluster for visualization purposes. Due to the cluttering in the display of a large number of passes, we used a force-directed edge bundling algorithm (HOLTEN; WIJK, 2009) to summarize the high-level trajectory. A straightforward way to bundle edges together in a general graph would be to first create a hierarchy and then use a Hierarchical Edge Bundling. However, in our case with a general graph, creating such a hierarchy would not be trivial. It is not evident which hierarchical-clustering scheme or spanning tree generation method to use. The advantages of using a force-directed approach are: the behavior is easy to understand following the straightforward physics model, the algorithm can be implemented in a reduced number of lines and it can be extended depending on the layout.

In the right of Figure 3.5 we show the clutter reduction in lines which reveal the shape trajectory. Each colored line intersection represent a player position where they did the corresponding pass; however, we need to highlight that the sub-possessions are displayed after preprocessing to found the most common shapes.

To inspect the detected passing sequences shapes in real context, using this approach we visualize the resulting eight clusters in Figure 3.6 along representative passing sequences displayed in their actual location on the pitch. For simplicity, we used a nickname for each cluster that reminds of the shape of the cluster.

Trajectories in clusters 1 and 3 have three colinear players and a single pass with a divergent angle. While in cluster 3 both sides have similar length (creating a peak), one side of cluster 1 is shorter than the other (swoosh). Cluster 5 has a bowl shape due to two consecutive passes with similar right angle. The bowl shape allows to move the ball between one point to another using two players as intermediaries. Clusters 2 and 4 present a zig-zag pattern. However, the angles between ball trajectories in cluster 4 are similar. A interesting fact about the zizag shape is its composition of two parallel lines. From the point of view of symmetry and spatial distribution the zigzag is an interesting shape in soccer analysis. While cluster 2 presents a zigzag pattern, it presents a smaller acute angle between a pair of passes. The acute angle suggests a small player movement due to a wall pass. In the other hand, clusters 6 and 7 represent the passes that form a closed circuit, mainly returning the ball near to the original position with a crossing (cluster 6) or exactly to the original point (cluster 8). Finally, Cluster 8 contains passes where the four players are arranged on a straight horizontal or vertical line. In the next chapter we study the presence of the clusters using visual analytics techniques to summarize the passing

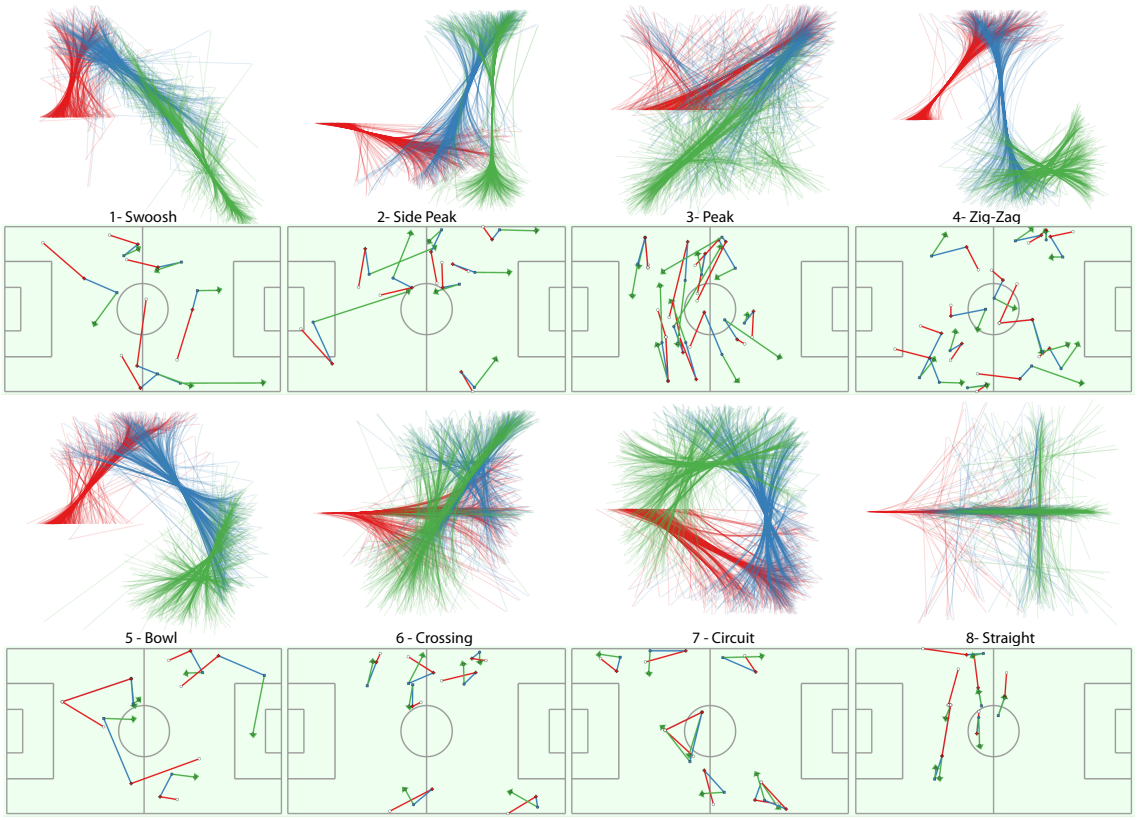
Figure 3.5: Example of ball trajectory bundling in cluster 1 (Swoosh).



Red, blue and green lines represent the first, second and last pass respectively, in four players sub-possessions. The ball trajectories are visualized after the preprocessing step and are centered on a reference point (0,0) for visualization purposes. Left: Passing sequences without bundling. Right: Passing sequences after force-directed edge bundling.

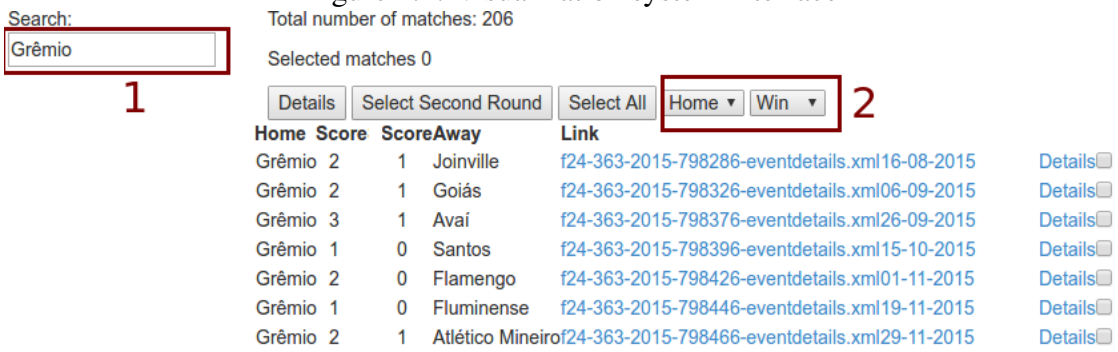
strategies that occurred during a tournament.

Figure 3.6: Geometric clusters visualization



Each cluster is composed of three passes colored with red, blue and green (in this order). Example of on pitch visualization with the original passing sequences that occurred during matches. Clusters 1,3 and 8 are composed mainly of straight lines. Clusters 6 and 7 are similar in the way they move the ball over the pitch but returns to the first passing point.

Figure 4.1: Visualization system interface



Illustrates the implemented web-based soccer analytics tool. Multiple soccer data format file reading is supported. The main selection match view consist of two filters. 1. Team name filter. 2. Match status filter: home or away condition plus match status: win, lose or draw. The user can select a single or multiple match for comparison purposes between teams and rivals.

4 VISUALIZATION DESIGNS AND ENCODINGS

To study the occurrence of the eight passing clusters, we designed an interactive visualization system (Figure 4.1) to support single and multiple matches analysis. The implemented prototype allows the selection of single match and multiple matches for team analysis. Additionally, we can filter the matches by result (win, lose or draw) and whether a team is playing at home or away. The filtering and selection operations support the analysis of strategies and team behaviour explained in Chapter 5.

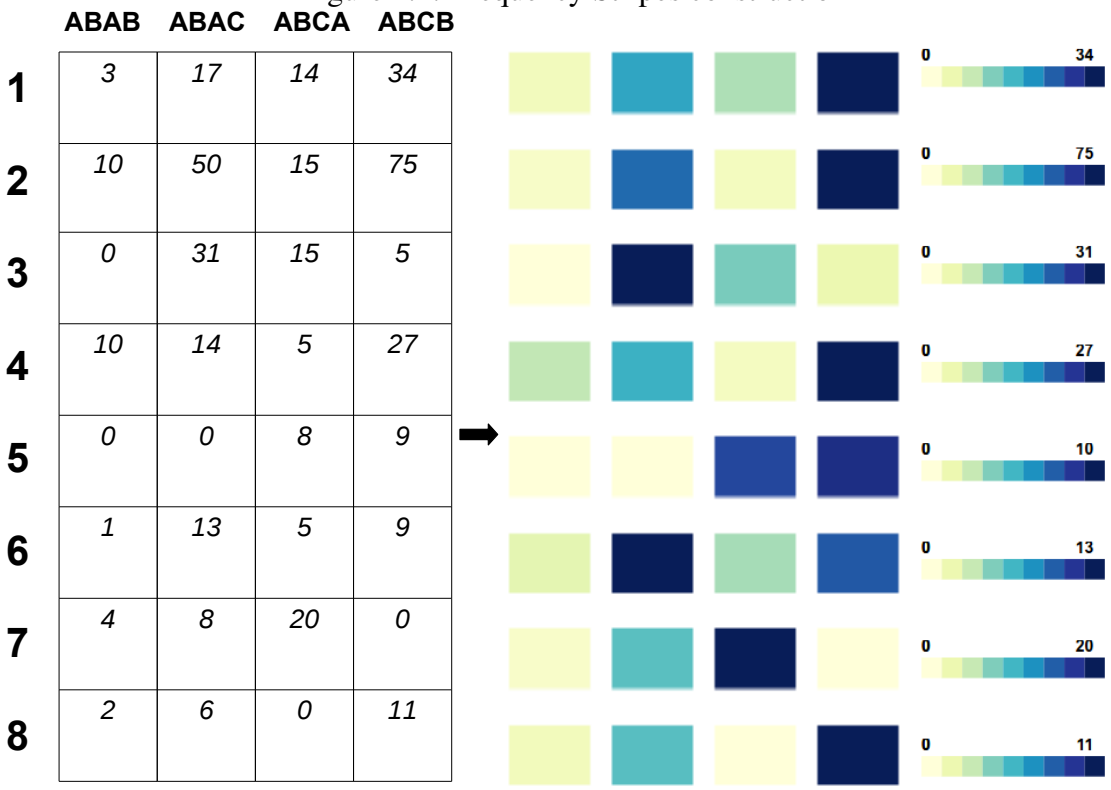
We propose a new visualization scheme to explore the number of structured passing sequences. The Frequency Stripes visualization presents an overview of the usage of clusters divided per cluster and regions of the field on which they occurred. By coordinating a modified timeline and trajectory heatmaps we build a tool which allows overview and specific details with spatial and temporal queries on demand.

The following sections describe the individual components of our visualization approach, and explain the usage and functionality to answer the requirements presented in Chapter 1.

4.1 Frequency Stripes

Following design requirement **R2** associated with strategy overview, we propose the use of frequency stripes to map the frequency of motifs usage on a given cluster of passes. The visualization was named as Frequency Stripes, because it encodes informa-

Figure 4.2: Frequency Stripes construction

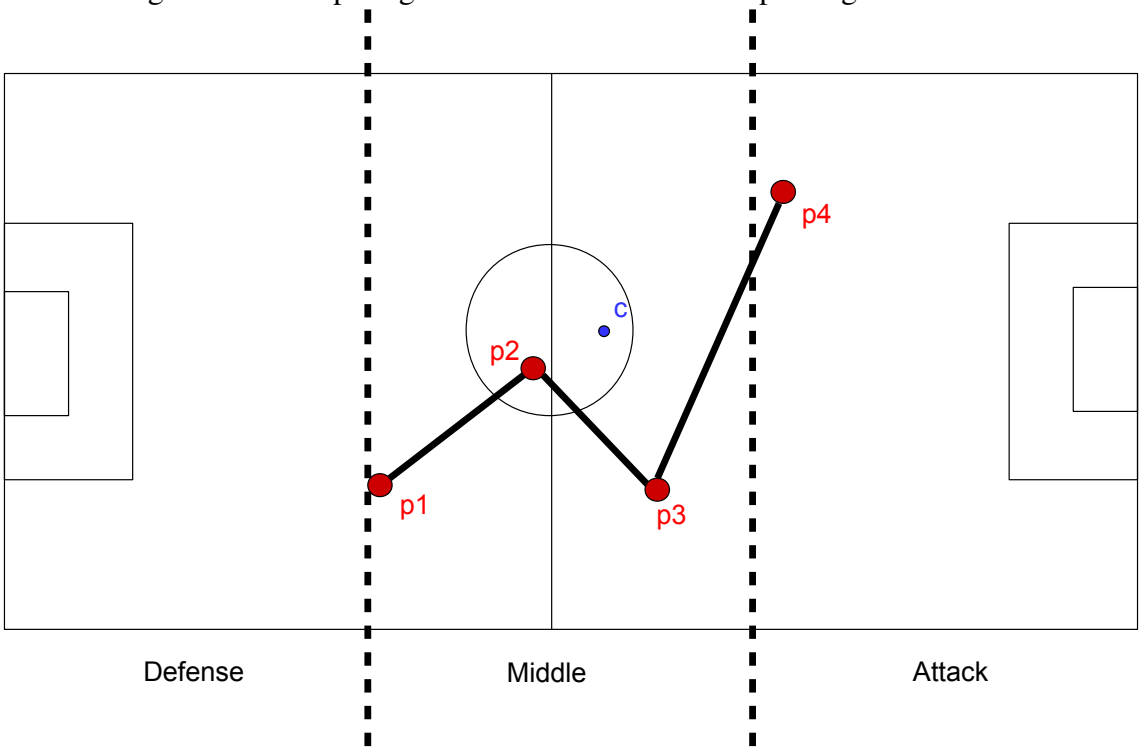


For each cluster (rows) we get the frequency of different motif structures (columns). Frequency Stripes example visualization for all Corinthians matches. Columns represent the motifs structures: ABAB, ABAC, ABCA, ABCB. The motifs count is associated with colors from the respective pallete.

tion about the number of all possible subposessions in the selected matches and groups them per team. We used a matrix representations, as to our knowledge, takes best advantage from screen size.

Frequency Stripes Visualization (FSV) consists on multiple matrices which relate motif structure patterns with the geometric clusters. Each matrix presents eight rows and four columns, which corresponds to the eight found clusters and four of the five possible motifs depending on player intervention: ABAB, ABAC, ABCA, ABCB. The motif ABCD was excluded from our analysis to allow us to focus on structured patterns involving one or two repeated players. The matrix cells represent the number of times a passing sequence was used by a team in the given cluster and motif structure. Colors are mapped to represent the frequency of the passing sequence as seen in Figure 4.2. For simplicity, colors are normalized by row, which makes easier the comparison between teams per cluster. For instance, the darkest blue on the first row is mapped to a diferent number from the darkest blue on the second row. We chose to do this type of normalization due to the

Figure 4.3: Computing soccer motifs centroid for pith regions division



Zigzag shaped subpossession. Our approach computes the centroid for each subpossession and categorize them according to their x-coordinate with three spatial classes: defenive, middle, offensive. The example shows a soccer motif labeled as a middle field motif.

high frequency variance between clusters.

Using our approach, we can extract different insights summarizing the passing behaviour of all matches in one single matrix. An average team performs 150 subpossesions on a five second time window. Given a tournament of 20 teams, where each team plays 38 games, a frequency stripe matrix codifies 5700 passing sequences. Additionally, it supports the analysis offering the distribution by cluster and motif structure.

Our design also allows to create a Frequency Stripe visualization per pitch region. To categorize each three long passing subsequence, we compute the centroid of each four point trajectory and label the sequence according to the position of the centroid in the pitch. Each visualization is normalized to show the direction of attack from left to right, so the x coordinate is partitioned in three regions. We separate the visualization depending on the pitch region the passing sequences occur: defense, middle and offense (Figure 4.3). Following this simplification, if a pass begins in the defense region, crosses the midfield and ends in the attack region, depending on the intermediate players position the motif centroid might correspond to the middle field region.

Another posibility is to change the order in which the matrix appear, which is useful for team analysis. We can rank the teams and keep with the top or bottom teams following a certain passing criteria. It is possible to reorder the teams matrices by pitch region, motif structure and cluster. In Chapter 5, we present a use case of Frequency stripes allowing comparison among teams by placing matrices side-by-side.

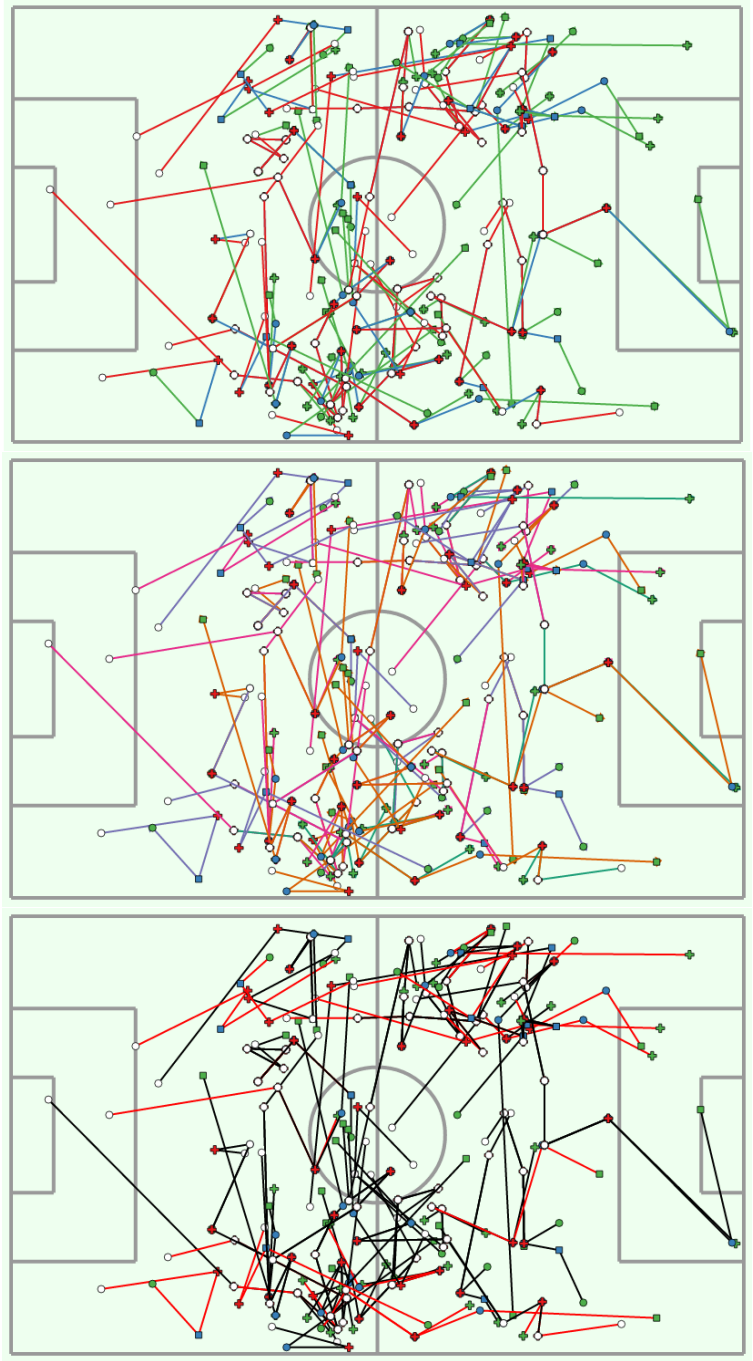
4.2 Sub-possessions on Pitch

To show the spatial attribute of the passes subsequence and the ball trajectory guided by requirement R1, we propose the pitch visualization which represents each pass with a line on the pitch. We investigate where do the passes occur and which shape do they compose in a sub-possession visualizing them on pitch. Being soccer a strategy game, coaches frequently design their teams formations on paper. They draw the possible player movements and pass sequences that lead to open plays whose main objective is to score goals. This visualization was inspired by coaching strategy drafts.

The three passes of a subsequence are drawn using the red, blue and green convention. An additional feature is the ability to draw lines in different palettes in real time depending on specific features of the passing sequence. Depending on the analysis driven by a specialist this could reveal interesting patterns if they want to detect a group of horizontal or vertical passes to separate zone crossing sequences, detect which of the sequences were wall passes to identify players associatons or examine the length of the three segments and compare them between clusters. Figure 4.4 shows three different color palettes for the pass line visualization.

The first coloring scheme is the convention red, blue and green for each segment respectively. Then, in the second scheme we map each pass color to the type of motif structure (ABAB, ABAC, ABCA, ABCB) and finally we present the coloring by bounding box. We used a simple and efficient way to separate the ball trajectory in horizontal or vertical possessions. We computed the bounding box for each trajectory and compared the width w and height h of the respective rectangles. We apply the following rules to classify a subpossesion in two classes: vertical or horizontal.

Figure 4.4: Sub-possession on Pitch visualization pallete examples



Our tool provides three different color palettes which depend on some attribute of the sub-possession. Each color palette helps to extract different insights about the ball trajectory and team ball movement. Top: illustrates a coloring by segment. We follow the coloring convention red, blue and green for the first, second and third passes. We observe that the first passes (red lines) are longer than the second and last passes. Center: a coloring by motif structure was used. We observe three motifs of type ABCB that start near the area, probably started by the goalkeeper. One ABAC motif was used near the rival area. Bottom: soccer motifs were colored by orientation. The horizontal subpossessions are clearly used by sides.

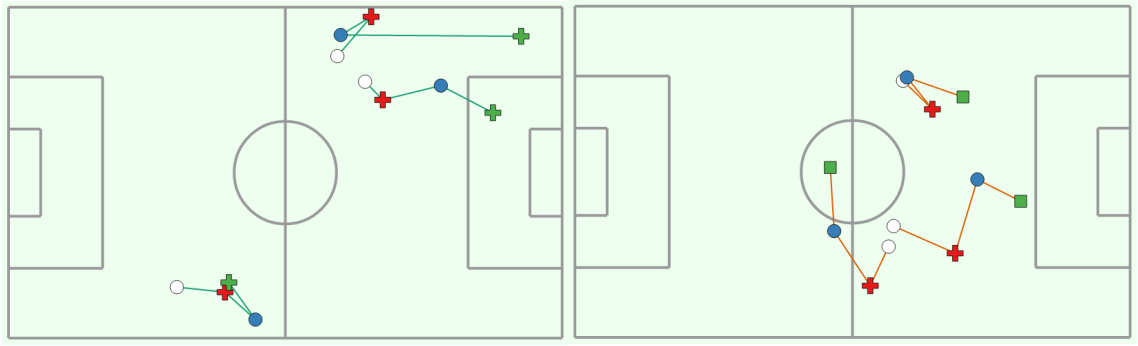
$$BB(w,h) = \begin{cases} \text{vertical}, & \text{if } w < h \\ \text{horizontal}, & \text{if } w > h \\ \text{omitted}, & \text{otherwise} \end{cases}$$

4.2.1 Players position glyphs

To be able to analyze and understand the role of players on the passing sequences, we draw glyphs in the corresponding player position. The glyphs represent information about the type of role in the structured motif. For this purpose, we use four colored geometric shapes: a circle, cross, square and a triangle which correspond to the role A, B, C or D respectively. The coloring scheme for the glyphs is applied following the rules: the first position always is colored with white. The second, third and last player position always follow the red, blue and green convention used for the pass trajectory visualization. In Figure 4.5 we present a subset of subpossesions from the match Atlético Mineiro vs Corinthians. We divided the subpossessions by motif structure to highlight the differences between glyphs depending on the role of the player. The white circle identifies the beginning of the passing sequence. We can observe that mainly all the passing sequences are from left to right, with the exception of some sequences in ABAC and ABCA types. Combined with the clustering labels given by our approach, the visualization helps us to identify which passing sequences returns to the same place, which of them go forward and the ones whose objective is just mantain the possession without attacking the rival. Additionally, the applied color mapping allow us to observe where the teams begins a passing sequence, analyze the intermediate positions or check to which pitch region the final passes belong.

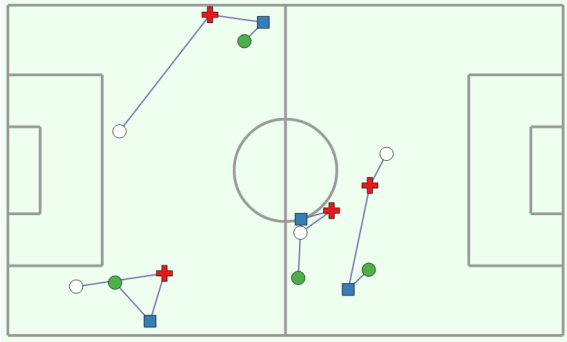
Alternatively, we present an aditional coloring scheme which highlights the positions of a selected player. The functionality helps the coaches to inspect how was a player behaviour in relation with the functionality of the passing network of the overall team. A use case of this visualization is shown in Chapter 5.

Figure 4.5: Player position glyphs example for each of the five structured motifs

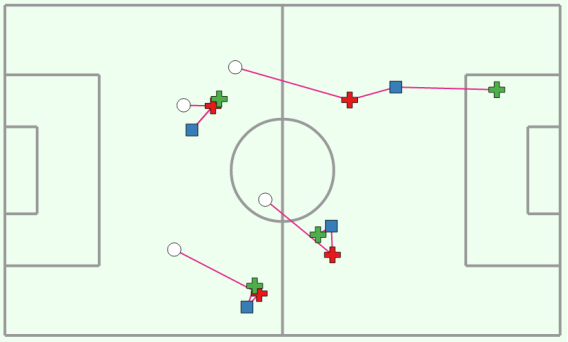


(a) ABAB

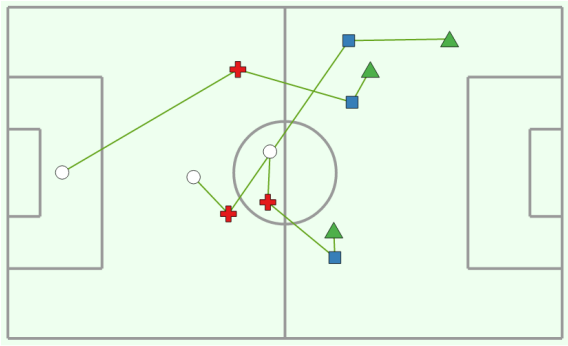
(b) ABAC



(c) ABCA



(d) ABCB



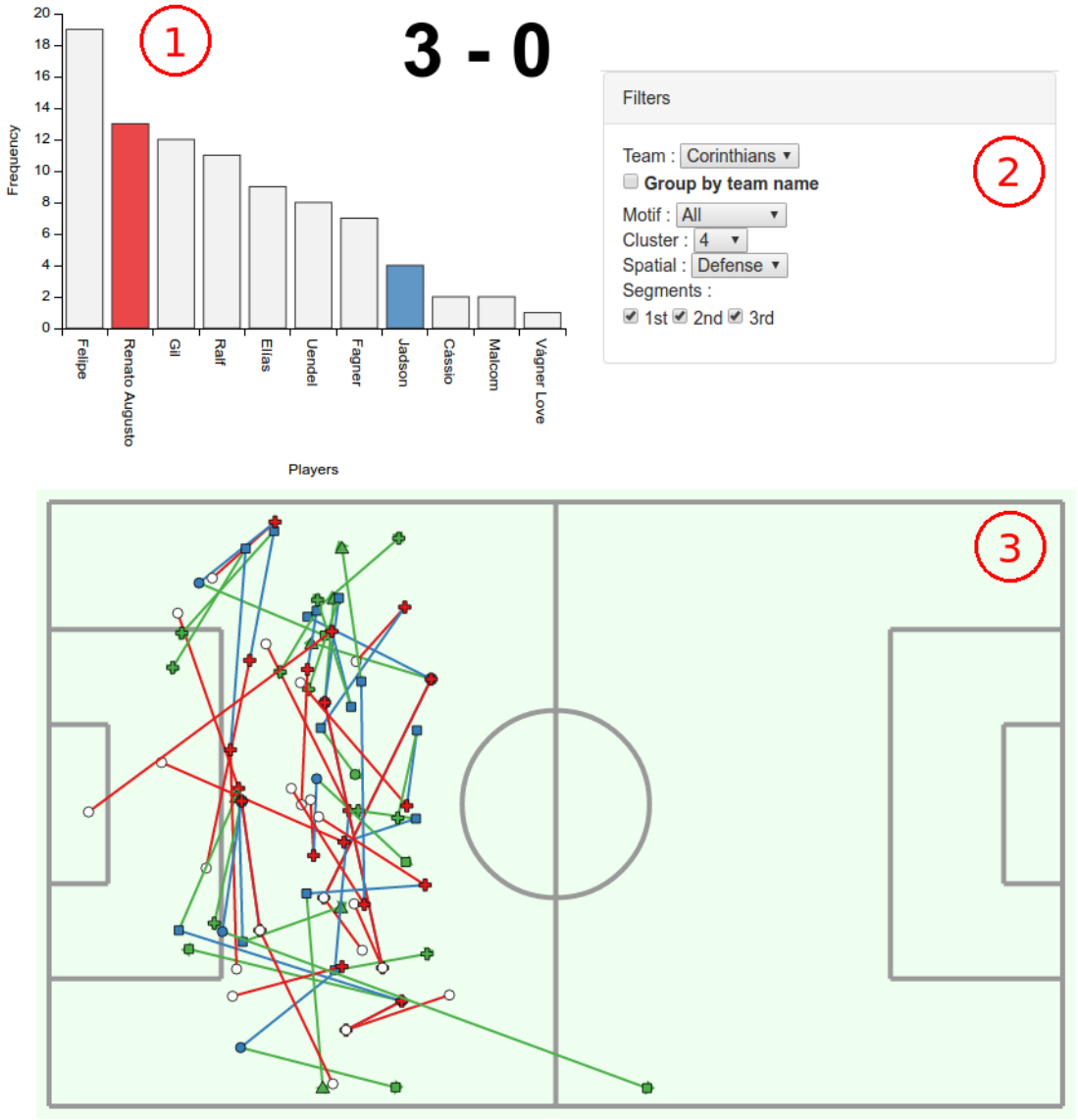
(e) ABCD

The circle, cross, square and a triangle shapes which correspond to the role A, B, C or D respectively. The standard coloring follows the convention of white, red, blue and green. If an specific player is selected on the histogram, the visualization will only show the position of the selected player.

4.3 Player Histograms

We extend a histogram visualization to display information about players participation on the soccer motifs that are currently displayed on pitch. Our system automatically selects the top 3 players that participated more on the soccer motifs of the match and colored them with red, blue and green respectively for the first, second and third place in the ranking. We added an additional feature to the player histogram which allows click-

Figure 4.6: Player histogram component coordinated with subpossession on pitch visualization. Query Example

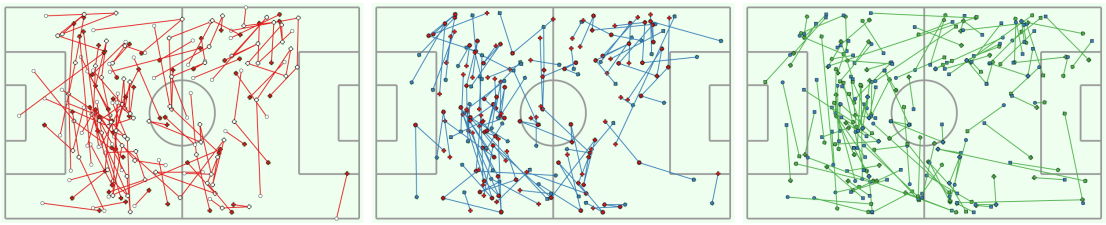


1. Player histogram with colors mapped to the top 3 most relevant players. 2. Filtering panel allows a real time selection and reduction of the dataset. The query could be read as display all structured motifs from Corinthians of type "zig-zag" which appear in the defensive section. 3. The Pitch subpossession visualization

ing on the bars from the plot. The user can highlight the position of the clicked player on the pitch and check where they appear on specific subpossessions. Figure 4.6 shows the interface for soccer motifs analysis. Filters panel allow the user to select a team, select the structure type, the cluster and the spatial division. Additionally, there exists the option to hide one or more passes that compose the subpossession.

In Figure 4.7, we separate the three passes from a soccer motif per pitch. The illustration presents the match Corinthians vs Joinville, where we clearly observe that

Figure 4.7: Sub-possession on pitch segment visualization



A soccer motif is composed of three passes. The line colors are mapped to the segment position on the sequence. Depending on the aim of the analysis the user can deactivate a specific segment of the sequence. The illustration shows the soccer motifs for Corinthians when they played against Joinville.

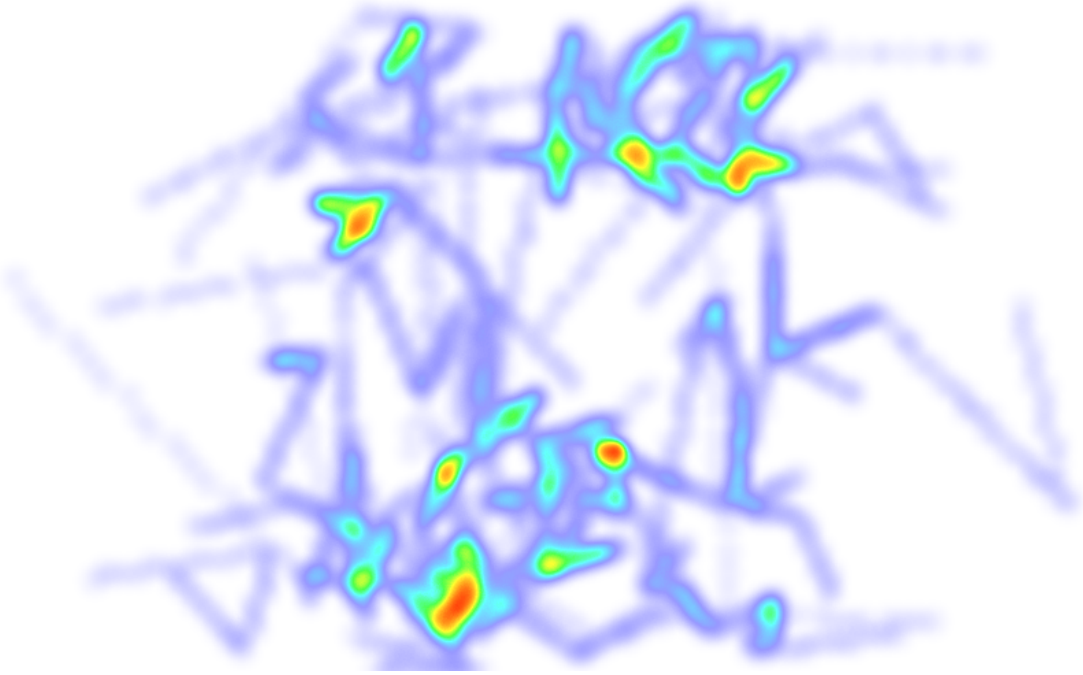
the possession from Corinthians was mainly defensive. The pattern on the three cases is similar: a high presence of structure passing near the goalkeeper but the third pass contains longer passes that go through the middle field and reach the rival goal primarily on sides. The final score was 3-0 for Corinthians victory. Analyzing just the last pass of the soccer motif we found that Corinthians passing strategy was to maintain possession combined with probably rapid long passes.

4.4 Sub-possessions Heatmap

The location that a player mantain during a match has been an important research topic in soccer analytics. The player positions are often displayed on websites and media with a heatmap visualization. Generally, the heatmap shows every place that a player were standing during the entire math which summarizes the player behaviors. However, a heatmap on its own is difficult to understand if we want to consider the space and time attributes. An interesting approach useful for soccer analysts is how the positional behavior evolves during the match. Thus, finer inspection would allow coaches to do specific studies in short time of periods and compare them between matches.

The requirement of summarizing passing events guided our exploration of spatial attributes with time filtering (design requirement **R3**). We used a trajectory heatmap as an alternative view to the pitch visualization. Instead of using the traditional player positioning, we differ in that we use the heatmap to highlight the regions of the pitch that had more ball movement. To create the sub-possessions heatmap, we divide the pitch in finite bins and for each trajectory we increase the count of the divisions that intersect with the trajectory. The trajectory heatmap is used to represent the trajectory of the movement of the ball during the sub-possession on the given time window. Our tool

Figure 4.8: Sub-possessions Heatmaps



Trajectory heatmap for Corinthians pitch visualization in Figure 4.4. Indicates that the team moved the ball more on the right side before the field half. The ball trajectory heatmap corresponds to all the 3 long passing sequences applied in a time window of five seconds. The heatmap uses a rainbow palette: the lower values are in the blue range and higher values in the reds. In between values pass through light blue, yellow and orange.

integrates the pitch visualization and trajectory heatmap with a time selection widget. Trajectory heatmaps solve the problem of occlusion while visualizing multiple motifs for one game or while analyzing multiple games. Colors are mapped to the number of times the ball passed through a given a point on the pitch. Thus, yellow and red colors represent parts of the game where the team had more control. (Figure 4.8)

4.5 Subpossessions TimeLine

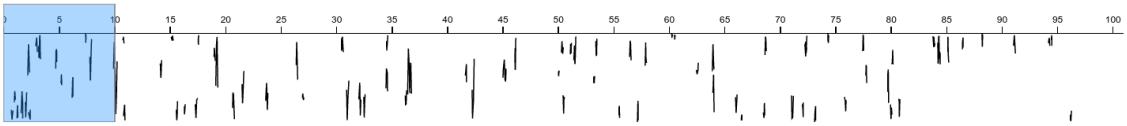
Different studies document the movement patterns of soccer players. Generally, the distances covered, movement speeds and directional changes are soccer statistics fairly well understood. However, we observed a lack in knowledge about the characteristics of ball possessions and specifically, data about short possessions that occured during a time window. In (ANDERSON; SALLY, 2013), Anderson documented that from the total distance covered by professional players just 1.7% was covered with the ball. This translates into about 1 minute of the match spent in possession of the ball. During each possession, a player covered 3 to 5 meters and averaged two touches per possession,

showing that interesting soccer moments can be really short and difficult to capture.

The main idea behind the subpossessions timeline is to present a visualization to improve the search for subpossessions on a timeline while analyzing multiple matches. We think that abstracting only one coordinate of the subpossessions in the visualization might reveal interesting patterns concerning the direction of play (left or right).

As ways of visualizing motifs and filtering by time we propose the subpossessions timeline. A widget that allows to select a time period and the same time displays the subpossessions during the match. The sub-possessions timeline supports the motifs on pitch visualization giving the capabilities to answer time queries. We incorporated a mini visualization of the motifs using the "y" spatial attribute of the sequence of passes. Each pass sequence appears in the timeline as a mini drawing, appearing at the top if the pass sequence was performed on top region of the pitch and similarly for the bottom region. As we use the "x" attribute to map the time of the game, the "x"-coordinate of the soccer motif is omitted. The subpossessions timeline allows the user to have an overview of the density of the motifs and their distribution along the "y" axis of the pitch. Furthermore, the visualization can be used to compare behavior of teams along time and see if they preferred the right or left side of the pitch for an offensive or defensive sequence of passes. The offensive or defensive sequences are filtered using selectors implemented in the web interface. In Figure 4.9 we observe that in the last ten minutes of all mathces of Atlético Paranense, the team had a preference for the left side for the "circuit" motifs.

Figure 4.9: Subpossession timeline example for Atlético Paranaense



Motifs timeline. The visualization admits and interactive brush for time period selection. Addtionally, it maps the y-coordinate of each soccer motif. The illustration shows the subpossession timeline filtered by cluster 7 (circuit) for all games from Atlético Paranaense.

5 VISUALIZATION RESULTS AND ANALYSIS

In this chapter, we present the results obtained with our visual analytics tool. The tool combines the previously presented visualization designs and allows flexible analysis of multiple matches. First, we study the presence of the soccer motif clusters in the tournament. Then, based on motifs structure and cluster, we characterize each team using the Frequency Stripe visualization and detect the teams that prefer defensive or offensive passing strategies. Finally, we use trajectory heatmaps to summarize the behavior of all teams and focus on individual player analysis for two important teams in the Brazilian Serie A tournament. We end the chapter showing an expertise feedback about the relevance of the tool.

The prototype was developed using web-based technologies, specifically it used Javascript with help from the D3 library. To run the system, the user only needs a web browser and a localhost server. Our tool reads the data from JSON files which are previously preprocessed. For initial statistics and data preparation, we used the R programming language.

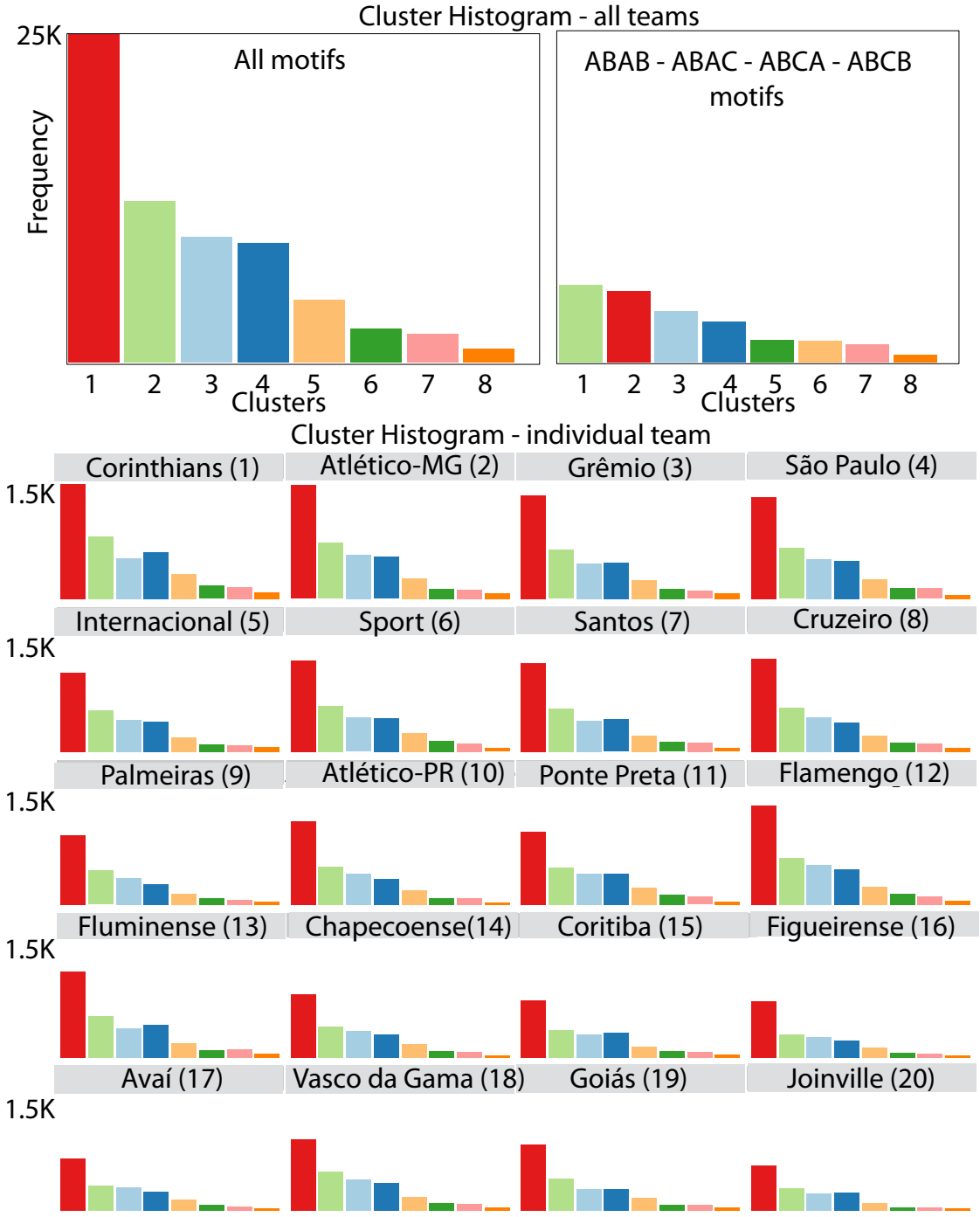
5.1 Passing strategies analysis

We built a case study to validate our approach over the 2015 Brazilian Serie A dataset. We describe below the analyses performed with our visualization designs over the dataset. We present three kinds of analysis: a team characterization by clusters usage with histograms, a comparison between teams visualizing passing clusters on pitch and finally, an evaluation of player presence on those motifs.

5.1.1 Overall Analysis of Clusters Frequency

To give an overall insight of the matches, we show the frequency of the eight discovered shapes in Figure 5.1. The passing sequences from Cluster 1 were the most used during the second half of the season. In contrast, we can observe that the frequency of trajectories inside Cluster 2 and 3 are almost half the number of Cluster 1. We plot the same histogram excluding the motif structure ABCD. The histogram shows a change in the order between cluster 1 and 2 as the most applied sequence and a similar distribu-

Figure 5.1: Soccer motifs frequency analysis with histograms



Soccer motifs frequency for Brazilian Serie A data set. Corinthians and Atlético Mineiro were the two teams that applied more consecutive passing. The order of histograms per team correspond to the tournament ranking.

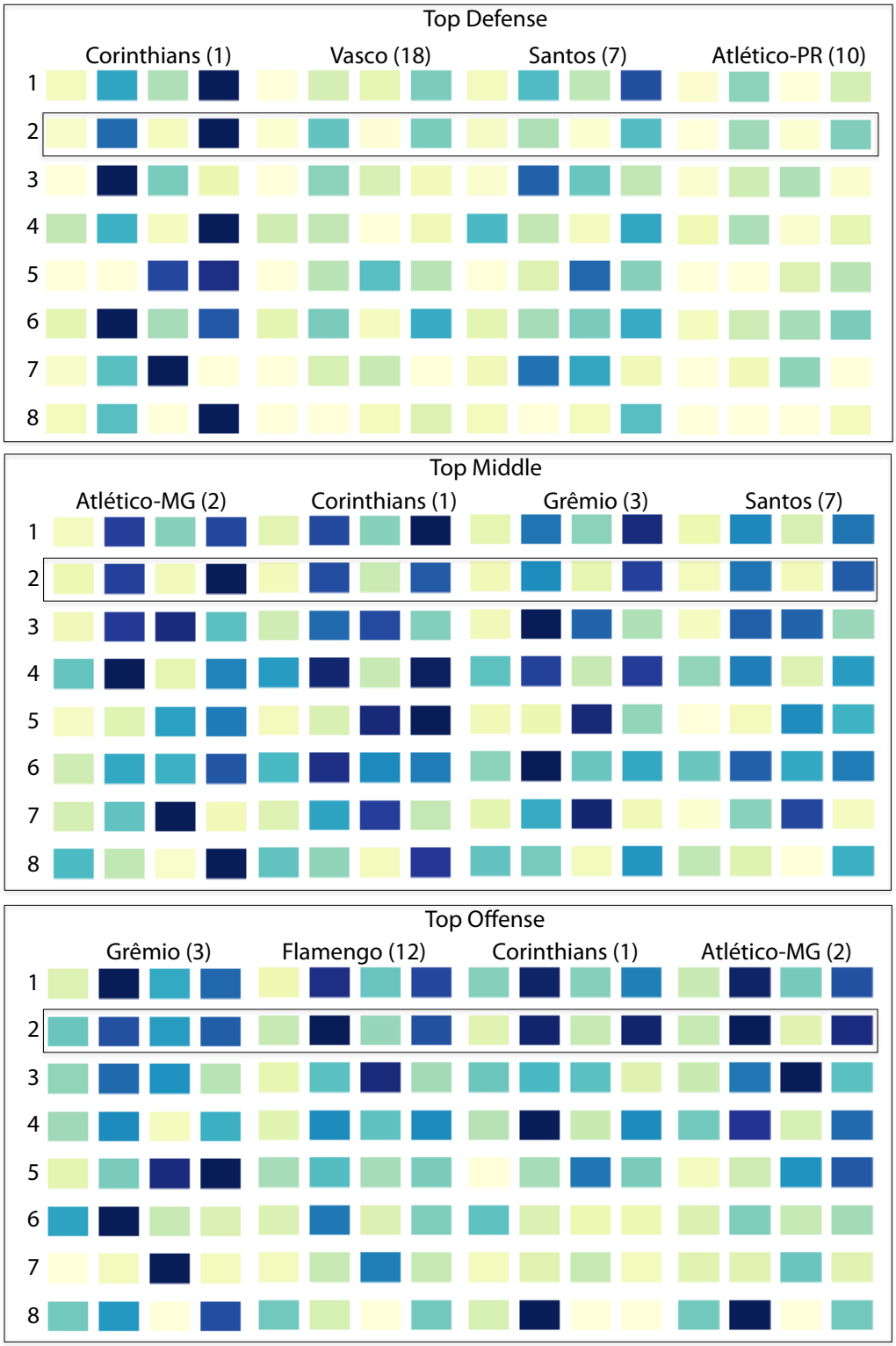
tion for cluster 5 and 6. On the following subsections, we use the most two frequent clusters 1 and 2 to guide our research about teams passing behaviour. Figure 5.1 shows the histogram of clusters per team. The histograms were ordered following the standings ranking of the tournament. Top tier teams presented a larger number of consecutive passing subsequences than lower table teams.

5.1.2 Teams behaviour analysis

We use the frequency stripes visualization to obtain an overview of the motif distribution during the tournament. We placed a visualization per team side by side and order them by the total number of passing sequences on cluster 2. The frequency stripes visualization was filtered to show a single region pitch at the same time. In Figure 5.2, we compared four teams with most cluster 2 passing sequences in defense, middle and offense On defense, Corinthians far exceeds the second place Vasco on defensive motifs, mainly on clusters 1,2,4 and 8 with ABCB structure. The ABCB and ABAC structures both consist on a simple pass followed by a wall pass. In soccer, a wall pass is defined as a movement in which one player passes the ball to another and sprints forward to receive the ball. They differ in whether the individual pass was made at the beginning or the end of the sequence. ABAC and ABCB were the two most frequent motif structures used by the leader of the tournament. In the other hand, offensive motifs are equally distributed among teams; however, Corinthians maintained the high frequency in clusters 1,2,4 and 8 but in this case in the structure ABAC with the long pass in the end. From the ordering, we notice that the top 3 teams from the final standings appear in top positions in the offensive region. However, of those three teams, Corinthians is the only one which appears in the defensive ranking.

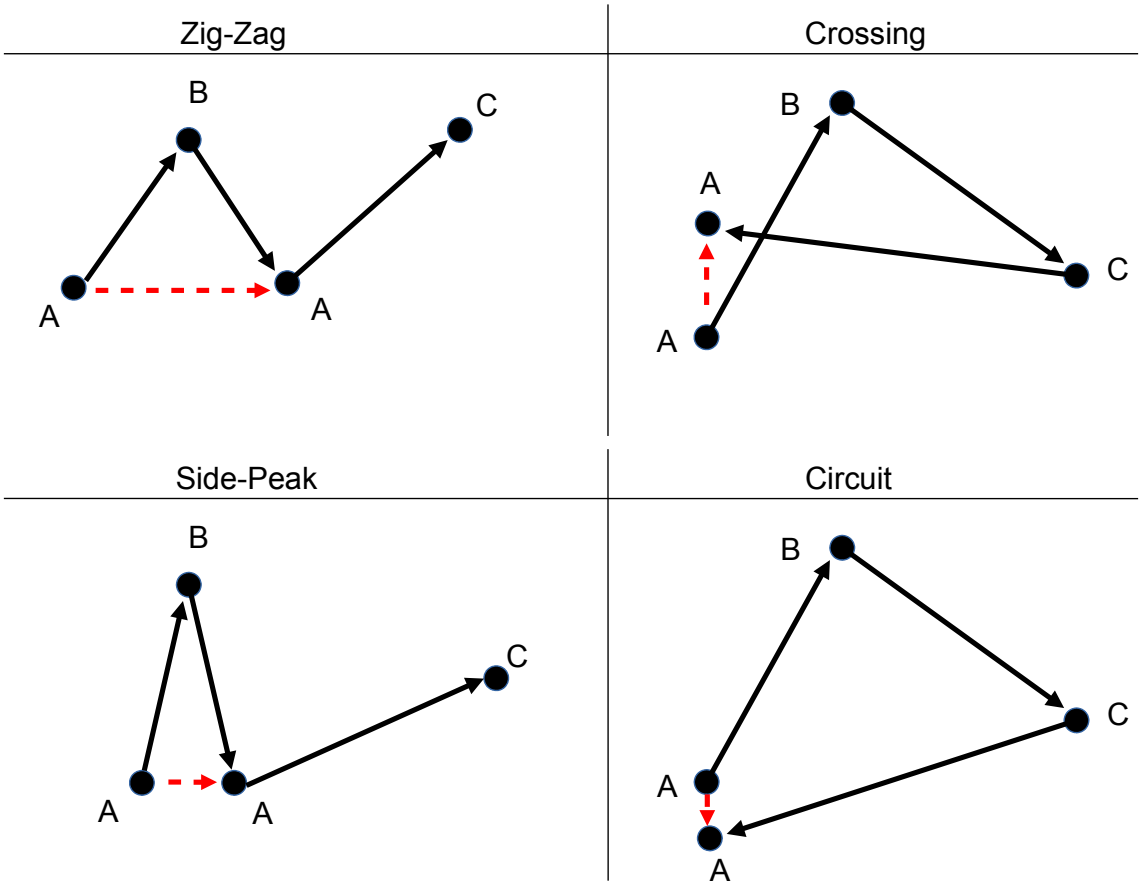
The structure of the passing sequence ABAC and its shape present useful insights for our analysis. Being Corinthians awarded with the best defense of the tournament, the team had the need for rapid ball movement with short passes to reach the rivals goal. The shape of the passing sequences of cluster 4 (zig-zag) support the need for counter-attacking that had Corinthians. From all the eight shapes, the zigzag shape is the one which allows to cover a larger area, move the ball from right to left and avoid the rivals on the way. In figure 5.3, we illustrate the shape structure of cluster 4 along with the player movement needed to perform the structure ABAC. This passing subsequence allows rapid movement of the ball from the starting point of A to the receiving player C. The intervention of the pivot (B) allows to make room for the player A to run a considerable distance, probably to avoid marking rivals. The advantage of this passing sequence is the wide movement of the ball without involving much player physical strain. In comparison with the “Side-Peak” trajectory, we observe the shorter distance run by player A. We have two reasonable posibilities, the first is that player A stayed its place without moving for almost five seconds which is the maximum time B would take to perform the pass. The other one

Figure 5.2: FSV divided by pitch region for “Side-Peak” motifs



Top four teams on defense, middle, and offense FSV matrices ordered by cluster 2 (Side-peak). Team matrices with higher cluster 2 frequency are shown in decreasing order from left to right. Along the name of the team, we show the final fixture position in the league. Observe that the first three teams in the league appear consistently in this visualization, in particular the first place Corinthians, that appeared in the three views as first on defense, second on middle and third on offense.

Figure 5.3: Players and ball movement comparison in different clusters



Cluster 2 vs 4. Red arrows represent player movements. The right parts of image show the ball movement with black arrows and the player movement with red arrows.

is that player A passed the ball and run; however, B return the ball almost instantly which gave no time to A for running longer distances. We call this a rapid wall pass. Similarly, the differences between the "Crossing" and "Circuit" shape regarding player movement is the distance that the first player A can run after executing the pass and recieving the ball.

Figure 5.3 compares shapes between clusters 4 (Zig-zag) and 6 (Crossing). For cluster 6, the structure ABCA was used because it was one of the most used passing sequences in the middle field by Corinthians. A significant difference is observed between both clusters related to the final position of the receiving player. In cluster 4, the three pass sequence is used to move the ball forward (or backwards). We observe a significant distance between first player (A) and final receiver (C) on the left of Figure 5.3. On the other hand, in cluster 6 the position of the fourth player is close to the starting point (the same is true for shapes from cluster of type "Circuit"). The shape from both passing sequences from clusters 4 and 6 support a style of playing, whether the team wants to move the ball from one point to another or to exchange passes while keeping the ball

possession. Since the middle region of the pitch is where most of the passes occur on a soccer game, and due to the high marking in this area of the field, we notice that Cluster 6 and 7 passing sequences may be the more beneficial for middle field if the objective is to keep possession. The high frequency of ABCA structure is explained as we observe the player A convenience of moving only a small distance to sucessfully complete the pass sequence in Cluster 6.

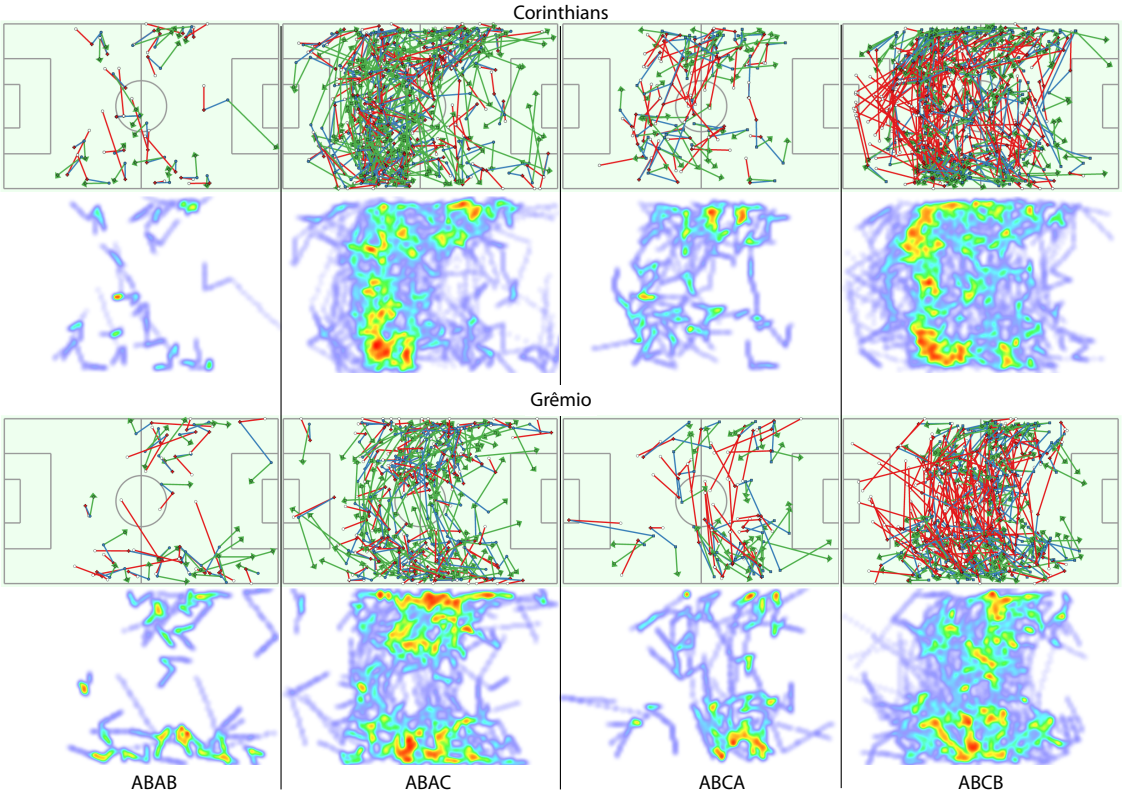
In the middle region, we observe that teams with more structured passing finished in the top positions of the final fixture. Although the pattern of motifs occurrence in the top three teams is similar, the biggest difference lies in the first column (ABAB). Corinthians clearly used more wall passes of type 6 which correspond to the crossing shape. The crossing and the circuit shapes both begin and end in a close position, which made them perfect to avoid rivals and maintain control of the ball possession without compromising their ability to transport the ball forward.

5.1.3 Motifs on pitch analysis

To answer the question about where in the pitch the passing sequences occur, we used the sub-possesions on Pitch visualization together with the trajectory heatmap to highlight the pitch regions where the ball was moved. To do this, we picked two teams Corinthians and Grêmio and separated the motifs by their structure (Figure 5.4). As we were interested on structured motifs with different players, we chose to analyze the most frequent cluster without the ABCD motif.

Figure 5.4 shows a predominance of vertical green (third) passes on structures ABAC and vertical red (first) passes on the structure ABCB. We attribute this behaviour to the fact that the long pass from cluster 2 is dependent on the position of the third different player on the sequence. Results show that the the long pass after or before the wall pass, is mainly between left and right side, probably to change the direction of play. From the trajectory heatmaps we confirm the defensive passing sequences of Corinthians, compared to Grêmio preferences over the left and right sides.

Figure 5.4: Sub-possession on pitch visualization plus trajectory heatmaps.



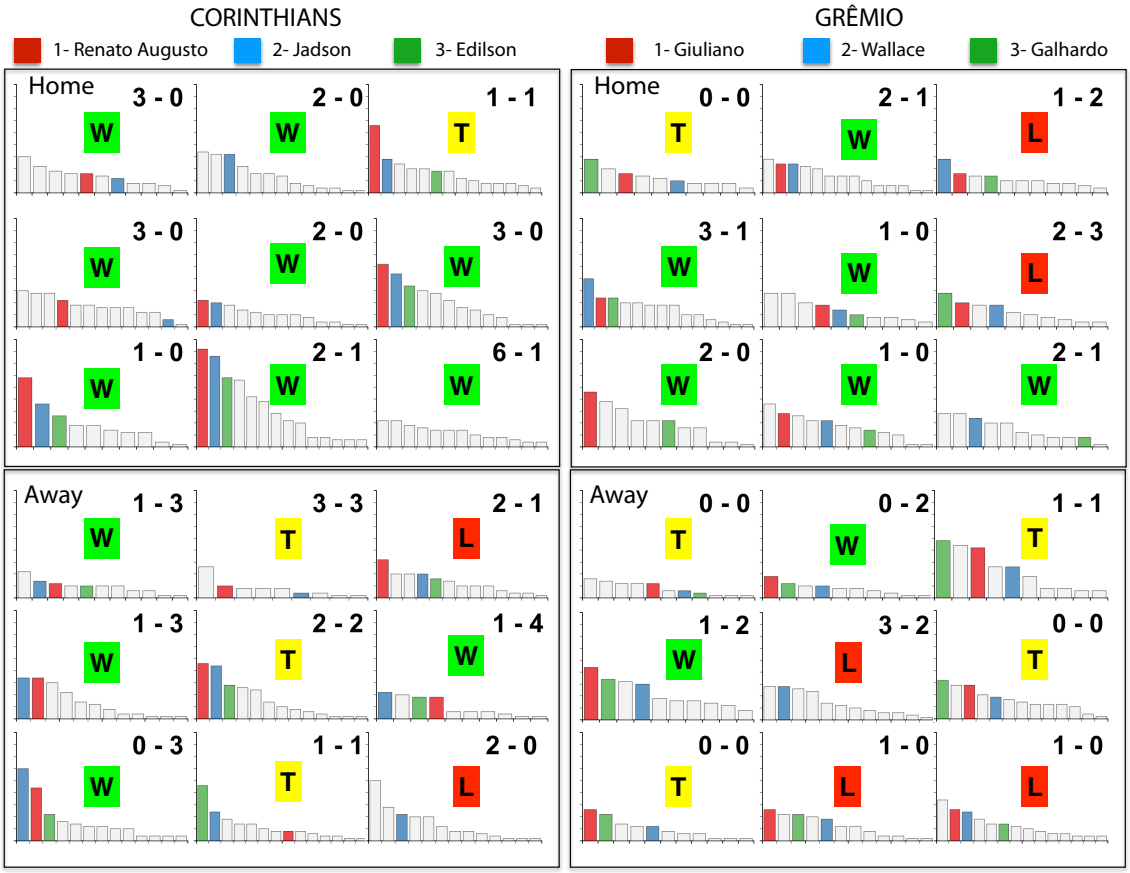
Top: Corinthians. Bottom: Grêmio. Columns represent the type of the soccer motif structure. The trajectory heatmaps represent places where players moved the ball with more frequency in *Side peak* passing

5.2 Individual Player Analysis

To find which players practice similar patterns on passing strategies, we use two visualizations: a histogram grid per match and the sub-possessions on Pitch plus glyphs mapped to players positions. First, we present a player histogram with all games from Corinthians and Grêmio to understand which players participated on the passing motifs and how they variated along the tournament. Figure 5.5 shows 18 matches per team. The histograms per match are organized as follows: the left half belongs to Corinthians and the right to Grêmio. The first three rows represent the matches where they play at home and the last three rows represent away games. Each histogram is annotated with the final score of the game in the format: HomeTeam - AwayTeam. Additionally, a colored letter represents whether the game was won, lose or draw by the team in the current analysis.

In Corinthians games, we see 5 of 9 home matches were Renato Augusto appeared as the player with more participation in structured passing. This pattern does not occur in away games where Renato Augusto appeared only in two games in the top of structured passing. The histogram colored by player, reveals that frequently Jadson (blue) appeared

Figure 5.5: Comparison between Corinthians and Grêmio: Appearance of top three players in the "Swoosh Cluster".

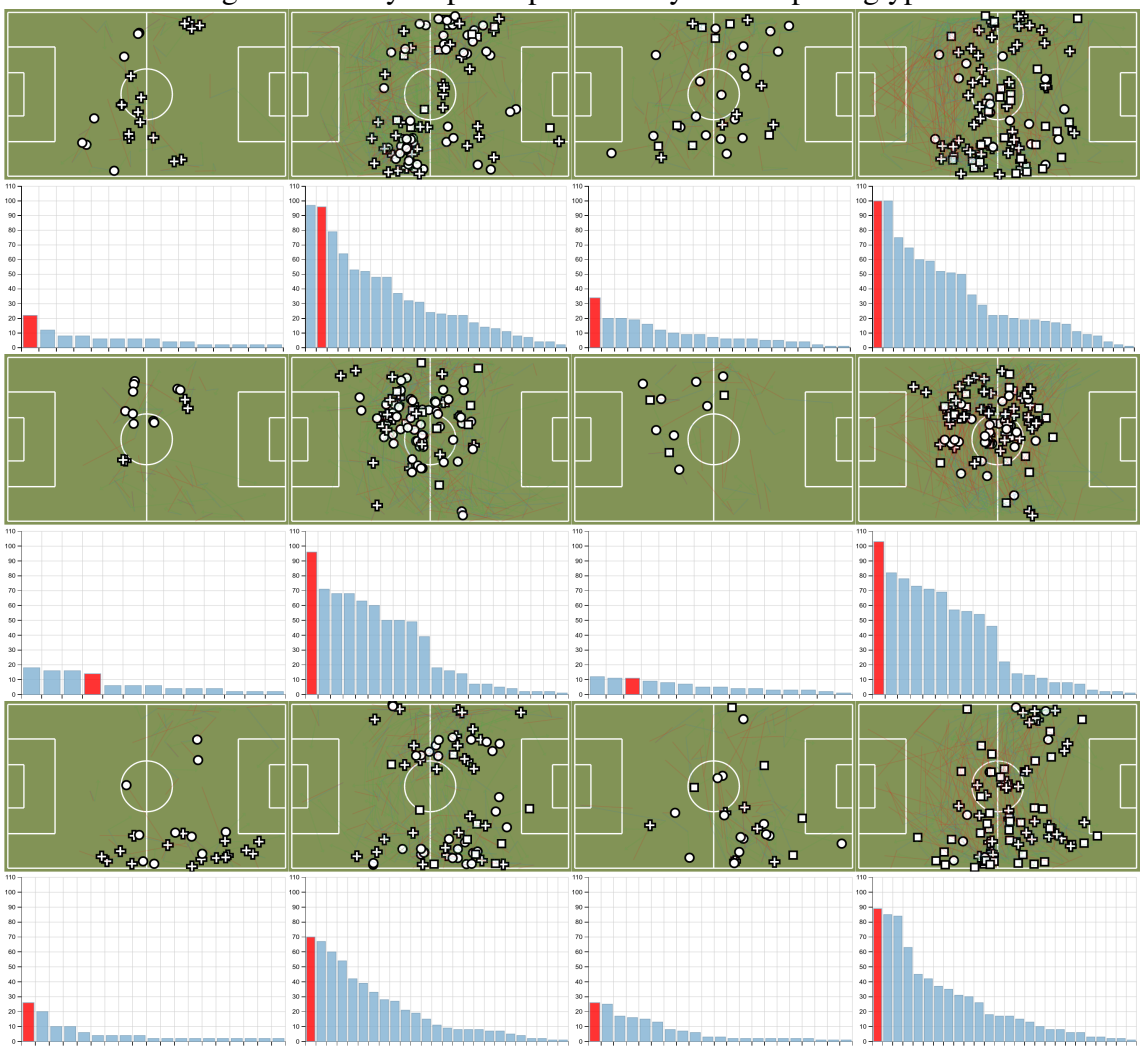


Players histogram for Cluster 1 with all motifs except ABCD. Corinthians and Grêmio. The upper 9 histograms correspond to home matches and the bottom 9 to away games. Colors were used to map the player names which appear more frequent in the top places. Renato Augusto was the top Corinthians player at home, but away Jadson was more important. Giuliano was more important for Grêmio in away matches than at home, where it divided the task with Wallace and Galhardo. The letters W, L, T represent won, lost and tied matches respectively.

along with Renato Augusto (red) as the most participative players. Moreover, in Gremio matches the pattern is less visible due to the reduced number of Cluster 2 motifs in comparison with Corinthians. However, we see that Giuliano appeared in 4 of 9 away games as the most participative player but in less quantity than Renato Augusto. Although the expected pattern was to find most of the colored players in all matches we noticed two outliers on Corinthians games against São Paulo (6-1) and visiting Sport (2-0), matches that were played by reserve players. From the visualization, we can also see that Corinthians did not lose a match while playing at home for the last half of the tournament.

We visualize the positions of players who appeared most in the passing motifs for three teams: Corinthians (top) and Atlético Mineiro (bottom). For this task, we chose two visualizations: a player histogram along with the player positions glyphs. As can be seen

Figure 5.6: Players participation analysis with pitch glyphs



(a) ABAB (b) ABAC (c) ABCA (d) ABCB

Soccer motifs player visualization for Cluster 2 with players participation histogram. Corinthians (top), Atlético-MG (mid) Grêmio (bottom) most participative players are displayed in red (Jadson, Rafael Carioca and Giuliano). Players positions are mapped to glyphs which correspond to the position of the player in the structure (A: circle, B: cross, C: square)

in Figure 5.6, Jadson from Corinthians presents a centralized positioning closer to the defense for ABAC and more offensive for the ABCB. Representing the player positions as glyphs helps us identify that more circles appear on the ABCA pitch for Jadson and on the ABAB for Rafael Carioca (Atl. Mineiro). We can see that Jadson frequently begin a passing sequence and recieve the final pass in the structure (ABCA). Furthermore, we observe that Rafael Carioca participates more as the first player in the wall passing ABAB (circles). We notice from Giuliano visualization that he appeared more on the right side for wall passes, but in general, there exists a pattern of presence on the both sides in the ABAC and ABCB structure.

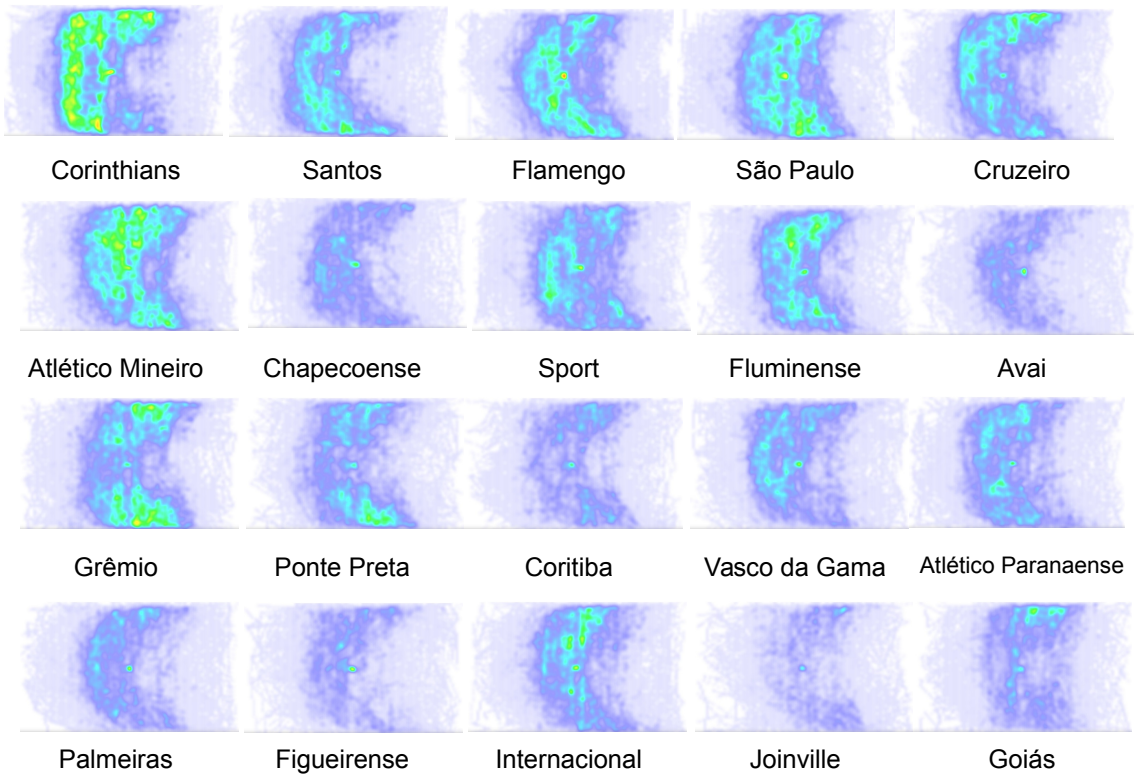


Figure 5.7: All heatmaps ordered by Defensive Cluster 1 Frequency

Based on requirement R3, we used the trajectory heatmaps to summarize the passing sequences, regardless of the cluster they belong. Figure 5.7 shows the 20 teams from Brazilian tournament ordered from top to bottom by the quantity of defensive passing sequences for Cluster 1. Different patterns arise from teams on the top, middle and bottom positions. The first five teams (top row) present high number of structured passing sequences on their own field with ball movement through the sides. However, the result clearly shows Corinthians preference over the left side in the last half of the pitch. In the bottom row we observe the teams with the lowest frequency of consecutive passing in the defense. The position of Internacional show us that the team strategy was entirely different from Corinthians or Santos regarding the defensive zone. An alternative strategy for teams with a low frequency of passing sequences could be the long ball or individual talent, which is out of the scope of the current analysis.

5.3 User Feedback

We gave a demo of our system to a soccer analyst, who is not a computer scientist. The analyst initially identified that the passing patterns and clusters were not intuitive to him, but this could be improved upon familiarity with the system. Also, the preferred analysis in the system is the one that focuses on players. For example, the analyst demonstrated special interest in identifying the player involved in a specific pattern, such as the ABAB scheme of a particular cluster. They were mostly interested in the pass sequences that take the ball to the ofensive zone. For instance, they suggested to show what kind of structures were important for taking the ball to the offensive zone or which sequences result in a key pass, cross or shot. The analyst agreed that our prototype would improve the information useful for scouting purposes. Finally, he wondered about the difference between soccer motifs of the team in the presence and abscence of a specific player.

As we can see in the accompanying video (<https://vimeo.com/206172940>), the interface allows different selections and the visualization of global and individual patterns. In particular, the dynamic update of the histograms allows this analysis to be performed. The analyst questioned the usefulness of the specific heatmaps in comparison to the standard global heatmap, but we explained that situations with many passing sequences could be complemented with specific heatmap visualizations to narrow the analysis to specific regions, patterns, or time intervals. Finally, the analyst demonstrated special interest in the zig-zag pattern, and believes this can lead to interesting insights.

6 CONCLUSIONS AND FUTURE WORK

In this work, we proposed a methodology for geometry feature extraction from passing sequences and a visualization system to study the trajectories created by ball sub-possessions. Geometrical analysis of passing strategies allows experts to explore the correlation between ball, player position and contrast them with the intended tactics planned by managers. In our exploratory study using a two-phase clustering technique, we found eight representative types of passing shapes. We introduced the Frequency Stripes Visualization (FSV) a visualization technique to analyze the frequency of the passing sequences which considers multiple attributes.

Additionally, we introduced extended visualization designs to support the exploration of the resulting clusters using the second part of the Brazilian Serie A soccer league. In such scenario, the FSV technique allowed us to compare the usage of sub-possessions from 20 teams in a complete tournament dividing frequencies by shape, passing structure and pitch region. Using our approach we recognized the defensive style of the league champion Corinthians for *Side Peak* (single + wall) passes. Additionally, we provide methods for space and time exploration of passing sequences using the sub-possessions on Pitch visualization with the support of trajectory heatmaps to reduce cluttered passes while analyzing multiple matches. Our final contribution is to demonstrate the use of the visualization designs in individual players style detection. We used a histogram grid and players glyphs to analyze their participation. Results show a player partnership between Renato Augusto and Jadson, the two players with most occurrences on passing strategies with wall pass. Both our visualization designs helped us to identify players who participated more in determined shapes, serving as a quantitative support for managers and scouts while signing new players.

Regarding future work, we plan to test our system with data from other leagues, adapt our prototype to load other soccer formats and work with positional player data. We think that in addition to ball movement, the position of players when they do not have the ball plays an important role in defensive and ofensive tactics. Additionally, we will explore the correlation of ball subpossessions with other soccer events, such as ball losses, fouls, corner kicks and goals.

REFERENCES

ANDERSON, C.; SALLY, D. **The Numbers Game: Why Everything You Know about Football is Wrong**. [S.l.]: Penguin Books, Limited, 2013.

BIALKOWSKI, A. et al. Large-Scale Analysis of Soccer Matches using Spatiotemporal Tracking Data. **Data Mining (ICDM), 2014 IEEE International Conference**.

BIALKOWSKI, A. et al. Identifying Team Style in Soccer Using Formations Learned from Spatiotemporal Tracking Data. **2014 IEEE International Conference on Data Mining Workshop**, p. 9–14, 2014.

FOOTOSCOPE: FIFA World Cup South Africa. 2010. <http://www.footoscope.com/worldcup2010/>. Acesso em: 24 mar. 2017.

GIULIANOTTI, R. Scotland's tartan army in italy: the case for the carnivalesque*. **The Sociological Review**, Blackwell Publishing Ltd, v. 39, n. 3, p. 503–527, 1991.

GOLDSBERRY, K. Courtvision: New visual and spatial analytics for the nba. **MIT Sloan Sports Analytics Conference**, 2012.

GUDMUNDSSON, J.; WOLLE, T. Football analysis using spatio-temporal tools. **Computers, Environment and Urban Systems**, v. 47, p. 16 – 27, 2014.

GYARMATI, L.; ANGUERA, X. Automatic extraction of the passing strategies of soccer teams. **2015 KDD Workshop on Large-Scale Sports Analytics**, p. 0–3, 2015.

GYARMATI, L.; KWAK, H.; RODRIGUEZ, P. Searching for a Unique Style in Soccer. **Social and Information Networks**, p. 5–8, 2014.

HOLTEN, D.; WIJK, J. J. van. Force-directed edge bundling for graph visualization. In: **Proceedings of the 11th Eurographics / IEEE - VGTC Conference on Visualization**. Chichester, UK: [s.n.], 2009. (EuroVis'09), p. 983–998.

HUGHES, M.; FRANKS, I. Analysis of passing sequences, shots and goals in soccer. **Journal of Sports Sciences**, v. 23, n. 5, p. 509–514, 2005. PMID: 16194998.

JANETZKO, H. et al. Feature-driven visual analytics of soccer data. In: **Visual Analytics Science and Technology (VAST), 2014 IEEE Conference**. [S.l.: s.n.], 2014. p. 13–22.

KODINARIYA, T. M.; MAKWANA, P. R. Review on determining number of Cluster in K-Means Clustering. **International Journal of Advance Research in Computer Science and Management Studies**, v. 1, n. 6, p. 2321–7782, 2013.

LEGG, P. A. et al. Matchpad: Interactive glyph-based visualization for real-time sports performance analysis. **Computer Graphics Forum**, Blackwell Publishing Ltd, v. 31, n. 3pt4, p. 1255–1264, 2012.

LEGG, P. A. et al. Transformation of an uncertain video search pipeline to a sketch-based visual analytics loop. **IEEE Transactions on Visualization and Computer Graphics**, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 19, n. 12, p. 2109–2118, dec. 2013.

LI, Y. Protractor: a fast and accurate gesture recognizer. **Proceedings of the 28th international conference on Human factors in computing systems**, p. 2169–2172, 2010.

LINK, D.; WEBER, H. Using Individual Ball Possession as a Performance Indicator in Soccer. **International Journal of Performance Analysis in Sport**, p. 0–2, 2015.

LUCEY, P. et al. Representing and discovering adversarial team behaviors using player roles. In: **Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on**. [S.l.: s.n.], 2013. p. 2706–2713.

LUCEY, P. et al. Quality vs Quantity: Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data.

LUCEY, P. et al. Assessing Team Strategy Using Spatiotemporal Data. **Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, p. 1366—-1374, 2013.

MILO, R. et al. Network Motifs : Simple Building Blocks of Complex Networks. **Science**, v. 298, n. 5594, p. 824–827, 2002.

OPTA. 2017. <http://optasports.com/>. Acesso em: 24 mar. 2017.

PERIN, C.; VUILLEMOT, R.; FEKETE, J. D. SoccerStories: A kick-off for visual soccer analysis. **IEEE Transactions on Visualization and Computer Graphics**, v. 19, n. 12, p. 2506–2515, 2013.

PEñA, J.; TOUCHETTE, H. A network theory analysis of football strategies. **arXiv preprint arXiv:1206.6904**, p. 1–6, 2012.

PILEGGI, H. et al. SnapShot: Visualization to propel ice hockey analytics. **IEEE Transactions on Visualization and Computer Graphics**, v. 18, n. 12, p. 2819–2828, 2012.

PLANCK, M.; LUXBURG, U. V. A Tutorial on Spectral Clustering A Tutorial on Spectral Clustering. **Statistics and Computing**, v. 17, p. 395–416, 2006.

ROSENTHAL, S. **Football Drawings**. <http://www.susken-rosenthal.de/fussballbilder/>. Acesso em: 24 mar. 2017.

RUSU, A. et al. Dynamic visualizations for soccer statistical analysis. In: **Information Visualisation (IV), 2010 14th International Conference**. [S.l.: s.n.], 2010. p. 207–212.

SHAO, L. et al. Visual-Interactive Search for Soccer Trajectories to Identify Interesting Game Situations. p. 1–10, 2016.

SPORTS, S. **Man. United vs Swansea City heatmap.** 2016. <http://www.scisports.com/news/2016/heat-map>. Acesso em: 24 mar. 2017.

STEIN, M.; SACHA, D. Enhancing Parallel Coordinates : Statistical Visualizations for Analyzing Soccer Data. **Electronic Imaging**, p. 1–8, 2016.

UW, S. et al. On spectral clustering: Analysis and an algorithm. **Advances in Neural Information Processing Systems 14**, p. 849–856, 2001.

WANG, Q. et al. Discerning Tactical Patterns for Professional Soccer Teams: an Enhanced Topic Model with Applications. **KDD Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, p. 2197–2206, 2014.

WEI, X. et al. Large-Scale Analysis of Formations in Soccer. **2013 International Conference on Digital Image Computing: Techniques and Applications (DICTA)**, p. 1–8, 2013.

WILSON, J. **Inverting The Pyramid: The History of Soccer Tactics**. [S.l.]: Nation Books, 2013.

WOBBROCK, J. O.; WILSON, A. D.; LI, Y. Gestures without libraries, toolkits or training: a 1 recognizer for user interface prototypes. **Proceedings of the 20th annual ACM symposium on User interface software and technology UIST 07**, v. 85, p. 159, 2007.

7 APÊNDICE - RESUMO DA DISSERTAÇÃO

7.1 Introdução

Futebol sempre tem sido um esporte muito popular no mundo todo e entre todos os esportes é um dos que posui mais rendimentos. Além dos grandes investimentos, é bem sabido que futebol é dos primeiros esportes que lidera a popularidade entre os torcedores. O desempenho dos times não é só acompanhado pelos torcedores, mas também pelos técnicos, patrocinadores e proprietários dos clubes. Em contraste com outros esportes, as baixas probabilidades de fazer gol e as estratégias dos jogadores adicionam a complexidade da análise de partidas. A análise de jogos de futebol é útil porque permite aprender os erros das equipes e permite estudar aos riváis para tirar proveito de suas fraquezas. Desde o começo do esporte, a análise estratégica também tem sido sempre estudada. Um exemplo disso, são as estratégias populares "Catenaccio" ou "Total Football" que dominaram no futebol por muitos anos.

Recentemente, incrementaram as pesquisas em dados analíticos esportivos com a ajuda dos avanços tecnológicos com os rastreadores GPS e processamento de imágens nos vídeos. Estes dados estatísticos dos jogos beneficiam aos técnicos e jogadores, adicionando informações de desempenho aos seus conhecimentos. Porém, alem de eventos chave num jogo como chutes, gols, faltas, número de passes, existe um interese da comunidade de pesquisa em entender os aspectos dinâmicos do jogo.

O objetivo de nosso trabalho é investigar as trajetórias de bola mais comúns criadas por sequencias de passes e como é que elas se relacionam com as estratégias em periódos específicos de tempo ou regiões no campo. A nossa abordagem consiste em definir uma distancia de similaridade entre trajetórias para assim encontrar as oito formas mais comúns das trajetórias formadas por sequencias de passes de tamanho três.

A nossa abordagem presenta as seguintes contribuições:

- Providenciar um método não supervisado para descobrir padrões de forma nas sequencias de passes baseados em um agrupamento por similaridade geométrica.
- Uma análise mutlifaceta de sequências de passes utilizando heatmaps de trajetórias coordinadas com a visualização *Frequency Stripe* para ter uma noção da frequencia das trajetórias para jogadores e regiões do campo especificas.
- Um caso de estudo de análise de estratégia de passes utilizando um dataset do Campeonato Brasileiro Serie A junto com feedback de um experto em futebol.

7.2 Abstração geométrica de estratégias de pases

Trabalhos anteriores introduzem o conceito de "flow motifs" para caracterizar estatísticamente os padrões nas sequências de passes. Para isso, eles usam o z-score para medir a significância dos subgrafos que geralmente possuem três ou quatro nós. Como uma generalização, neste trabalho foi desenvolvido um algoritmo de clustering para agrupar as seqüências de passes de tamanho tres por similaridade de forma.

A abordagem proposta presenta três partes: preprocessamento, agrupamento e visualização de trajetórias. No preprocessamento, calculamos todas as possessões de bola realizadas em um intervalo de tempo de 5 segundos. A trajetória da bola sempre é composta por quatro pontos que representam a posição dos jogadores no campo quando realizaram os passes. O preprocessamento de trajetorias consiste em quatro operações: reamostragem, rotação, escala e traslação. Em seguida, dado que temos o interesse em similaridade pela forma, definimos uma métrica para encontrar o melhor alinhamento entre multiples trajetórias. Para achar o menor ângulo em que duas trajetorias deven ser rotacionadas é utilizada a distancia ótima angular, que permite encontrar o ângulo entre dois vetores em um espaço de alta dimensão.

Na etapa do agrupamento, foi utilizada uma abordagem com dois algoritmos de clustering: Kmeans e Spectral Clustering. O Kmeans utiliza a distância euclideana para realizar o agrupamento ponto a ponto da trajetória. Esse primeiro procedimento serve para diminuir o tamanho do dataset de trajetórias para em seguida aplicar um clustering baseado em grafos com os centroides do Kmeans como entrada. A matriz de distancias que serve como entrada para o Spectral Clustering é calculada com a distância ótima angular. Evidência empírica sobre nosso dataset demostra que o “elbow method” selecciona corretamente o número de clusters. Kmeans consegue reduzir as trajetorias a 50 formas principais e depois foram reorganizadas em 8 clusters finais depois de aplicar o agrupamento Spectral Clustering.

Finalmente, na última etapa de visualização utilizando um algoritmo de edge bundling desenhamos as trajetórias centralizadas ao ponto 0,0. Cada segmento de reta foi mapeada para uma cor diferente. Vermelho, azul e verde para o primeiro, segundo e último passe respetivamente.-

7.3 Visualização e codificações

Para estudar a ocorrência dos oito clusters de passes, foi projetado um sistema de visualização interativa para suportar a análise de jogos individuais ou multiples. Além disso, é possível filtrar as partidas por resultado (vitórias, derrotas ou empates) e por jogos em casa ou fora de casa. As operações de filtragem e seleção ajudam a análise de estratégias e o estudo de comportamento de equipes.

Propusemos um novo esquema de visualização para explorar o número de sequências de passes estruturados. A visualização Frequency Stripes apresenta uma visão geral do uso de passes divididos por cluster e regiões do campo em que ocorreram. Junto com uma linha de tempo modificada e heatmaps de trajetórias, foi construída uma ferramenta que permite a visão geral e detalhes específicos com consultas espaciais e temporais por demanda.

7.3.1 Frequency Stripes

A visualização foi denominada “Frequency Stripes” (FSV), pois codifica informações sobre o número de subposseções possíveis nos jogos selecionados e as agrupa por equipe. Utilizamos uma representação de matriz, que em quanto ao nosso conhecimento, tira melhor proveito do tamanho da tela. O FSV consiste em múltiples matrices que relacionam padrões de estrutura das sequencias com os clusters geométricos. Cada matriz apresenta oito linhas e quatro colunas, as quais correspondem aos oito agrupamentos encontrados e quatro dos cinco possíveis subsequências dependendo da intervenção do jogador: ABAB, ABAC, ABCA, ABCB. A estrutura ABCD foi excluída de nossa análise para nos permitir focar em padrões estruturados envolvendo um ou dois jogadores repetidos. As células da matriz representam o número de vezes que uma sequência de passe foi usada por uma equipe numa estrutura ou cluster. As cores são mapeadas para representar as freqüências da sequências de passes.

7.3.2 Subposseções no campo

Os três passes que formam uma subsequência são desenhados no campo utilizando a convenção de cores vermelho, azul e verde. Uma característica adicional da nossa pro-

posta é a capacidade de desenhar linhas em diferentes paletas em tempo real, dependendo das características específicas de cada sequência de passe. Dependendo da análise conduzida por um especialista, isso pode revelar padrões interessantes se eles querem detectar grupos de passes horizontais ou verticais para separar as seqüências de cruzamento de zona, detectar quais das seqüências foram passes de paredem para identificar associados de jogadores ou examinar o comprimento dos três segmentos e compará-los entre clusters. Adicionalmente, a visualização representa cada posição de um jogador como un glifo com a forma mapeada para o papel do jogador na estrutura de passe sendo possível: A, B, C ou D.

7.3.3 Heatmap de subposseções

Utilizamos um heatmap de trajetórias como uma visão alternativa para a visualização no campo. Em vez de usar o posicionamento tradicional do jogador, diferimos em utilizar o heatmap para destacar as regiões do campo que tinham mais movimento de bola. Para criar o mapa de calor de sub-possessões, dividimos o campo em caixas finitas e para cada trajetória aumentamos a contagem das divisões que se cruzam com a trajetória. O heatmap da trajetória é usado para representar a trajetória do movimento da bola durante a sub-posseção em uma janela de tempo dada. Além disso, nossa ferramenta integra a visualização no campo e o heatmap de trajetória com um widget de seleção de tempo.

7.4 Conclusões

Neste trabalho, propusemos uma metodologia para a extração de características geométricas a partir de sequências de passes junto com um sistema de visualização de dados para estudar as trajetórias criadas por sub-possessões de bola. A análise geométrica das estratégias de passagem permite aos especialistas explorar a correlação entre a posição da bola, do jogador e contrastá-la com as táticas planejadas pelos técnicos. Em uma análise exploratória, utilizando uma técnica de clusterização de duas fases, achamos oito tipos representativos de formas de passes. A Frequency Stripes Visualization (FSV), é uma técnica nova de visualização proposta para analisar a freqüência das seqüências de passes que considera a forma dos passes e a estrutura dependendo da intervenção de um jugador.

Adicionalmente, apresentamos formas de visualização complementarias para apoiar a exploração dos clusters resultantes usando a segunda parte do campeonato brasileiro de futebol. Nesse cenário, a técnica FSV permitiu comparar o uso de sub-possessões de 20 equipes em um torneio completo separando as freqüências por forma, estrutura de passe e região do passe. Usando a nossa abordagem, reconhecemos o estilo defensivo do campeão da liga Corinthians utilizando o tipo de passe Side Peak (passe simples + parede). Além disso, fornecemos métodos para a exploração espacial e temporal de sequências de passes usando a Visualização de sub-possessões no campo, com o apoio de heatmaps de trajetórias para reduzir a sobreposição enquanto se analisam múltiples jogos.

Nossa contribuição final é demonstrar o uso da visualização proposta na detecção de estilo de jogadores individuais. Utilizamos uma grade de histogramas e glifos de jogadores para analisar a sua participação nas sequências de passes. Os resultados mostram uma parceria de jogadores entre Renato Augusto e Jadson, os dois jogadores com mais ocorrências em estratégias de passe de tipo parede. Ambos designs de visualização nos ajudaram a identificar jogadores que mais participaram em formas determinadas, servindo como suporte quantitativo para técnicos e olheiros enquanto analizam novos jogadores potenciáis para um clube.